

**Studying Relevance Judging Behavior
of Secondary Assessors**

by

Aiman Al Harbi

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirements for the degree of

Doctor of Philosophy

in

Computer Science

Waterloo, Ontario, Canada, 2016

© Aiman Al Harbi 2016

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

Secondary assessors, individuals who do not originate search topics and are employed solely to judge the relevancy of documents, have been found to differ in their relevance judgments. Their relevance judgments are used in constructing test collections, which play a significant role in evaluating search systems. These judgments are also used in e-discovery to assist with locating relevant material. To a large extent, our existing understanding of secondary assessors' judging behavior is limited to quantitative measurements. The goal of this thesis is to better understand the relevance judging behavior of secondary assessors. Therefore, we conducted two user studies to achieve this objective. The first study, which forms the main part of this thesis, was a think-aloud study, and provides what may be the first of such qualitative studies of secondary assessors' judging behavior. The second study of the research was to capture the uncertainty in secondary assessors' relevance judgments. Further examination of the behavior of secondary assessors when judging multiple types of documents was also carried out based on the data from the think-aloud study. Data obtained through the think-aloud method, permitted us to achieve more in-depth insight into secondary assessors' relevance judging behavior. We were able to directly listen to and note their thoughts during the assigned search tasks. Based on this data, we found that relevance judgments are made with differing levels of certainty. These levels of certainty vary from low to high. We also found that the varying factors of a search topic, the document, and the assessor can each impact differing judgments. The think-aloud study also reveals preliminary evidence regarding how the amount of detail stated in a search topic's description influences the relevance judging behavior of secondary assessors.

To capture the uncertainty in secondary assessors' relevance judgments, we designed four user interfaces in our second user study. The objective was to study the uncertainty in secondary assessors' relevance judgments when the level of uncertainty is self-reported. We found that they tend to make high certain relevance judgments despite the consensus level of a document. In judging high consensus documents, assessors' accuracy was lower when making low certainty relevance judgments, and the judgments were more accurate and tended to agree with NIST assessors when making high certainty relevance judgments.

For low consensus documents, we found assessors' accuracy to be low regardless of their certainty level. Finally, we found that assessors tend to spend less time when making high certainty relevance judgments, regardless of the consensus level of the document.

Further study of the behavior of secondary assessors when judging multiple types of documents, identified that relevance judgments are occasionally based on incorrect perception. We show how factors such as lack of familiarity, lack of understanding the search topic, absence of keywords and other reasons could be a source of not only incorrect relevance judgments, but also of those which are correct. We also illustrate how the length of search topics and documents, and their level of difficulty may further contribute to the issue of variations in the judgments.

Our research overall contributes to a more extensive, meaningful understanding of the behavior of secondary assessors. It establishes a foundation for more pertinent work in the future on the impact of uncertainty in secondary assessor's relevance judgments. Our findings also show that assessor training and background, search topics, and document length should be all considered and given additional attention in order to obtain more reliable results.

Acknowledgements

This thesis would not have been possible foremost without the assistance and support of Allah (God) and then the help and support of many individuals in so many ways.

At the top of the list of whom I wish to credit is my supervisor, Professor Mark Smucker. I would like to express my deepest appreciation and thanks to him for all of the things I have learned from him during the years I spent under his supervision in the PhD program. He has been always available and open to discuss and provide advice on different issues either on my research or outside the research area. I will never forget not only his positive attitude during the PhD years, but also his understanding and support when my father passed away in January 2013. It was a genuine honor to work with him. I also would like to thank my co-supervisor, Professor Charles Clarke, for the help and advice I received from him throughout the PhD program.

I wish to express my sincere gratitude to the examining committee, Mark Smucker, Charles Clarke, Gordon Cormack, Ellen Voorhees, Parmit Chilana, and Olga Vechtomova for participating on my committee and their valuable and helpful feedback.

My appreciation and thanks extend also to my parents: Lafi Al-Harbi and Fatimah Al-Tamimi. My father, who passed away two years ago, was the person who most encouraged and supported me to pursue a higher education. He was consistently available to offer a variety of support and advice. He will remain the greatest teacher in my life, from whom I learned a lot. I also cannot find enough words to express gratitude to my mother for her presence in my life. I thank her for all of the support and patience she has shown since my first year in Canada almost 8 years ago to the present day.

I also thank my wife, Taghreed, for being part of my daily life during my PhD years. Her presence has added valuable things to my life. Without her support and encouragement, completion of this thesis would not have been possible. She shared in all of the experiences I had in my PhD program day and night. I thank Taghreed for all of her support and encouragement. I am really blessed to have a wife like her. I wish also to express gratitude to my brothers and sisters for their usual supportive contacts and encouragement during my years in Waterloo. Their kind and uplifting words were a meaningful source of empowerment.

I would like also to thank my officemate and friend, Gaurav Baruah, for the great friendship we established during the PhD studies. I have been very fortunate to be his officemate since the beginning of the PhD years. We spent countless hours discussing different issues in research and academia in general, from which I benefited greatly. My thanks extend to the colleagues in the IR lab, Adam Roegiest, Bahareh Sarrafzadeh, Luchen Tan, Haotian Zhang, Mustafa Abualsaud, and Alexandra Vtyurina, for all of the enjoyable times we spent together.

I thank also the Saudi Bureau in Ottawa for their role in the completion of this thesis by providing me with all which was necessary to help me to complete my degree successfully. I also cannot overlook to thank the Saudi Students Association for being my “second home” in Waterloo. We, along with our family members, eagerly looked forward to Friday evenings when the Saudi Students Association organized socials. Their presence had also made our time in Waterloo seem to pass more quickly.

Finally, I would like to thank King Saud bin Abdulaziz University for Health Sciences for sponsoring me in the PhD program and acknowledge the the support received from Natural Sciences and Engineering Research Council of Canada (NSERC), and the University of Waterloo during my PhD program.

Dedication

To my parents, my wife, and my daughters.

Table of Contents

List of Tables	xvi
List of Figures	xix
1 Introduction	1
1.1 Background Overview	2
1.1.1 IR Evaluation with Test Collections	2
1.1.2 Primary Assessors vs Secondary assessors?	5
1.2 Problem Statement	6
1.2.1 Disagreement on Relevance judgments	6
1.2.2 Uncertainty in Relevance Judgments	9
1.3 Thesis Overview	10
1.4 Contributions	13

2	Literature Review	16
2.1	The concept of Relevance	16
2.2	Relevance Judgments	19
2.2.1	Disagreement about Relevance Judgments	19
2.2.2	Impact of Disagreement on the Performance of IR Systems	24
2.3	Factors that Influence Relevance judgments	28
2.4	Crowdsourced Relevance Assessments	33
2.5	Think-Aloud Method	35
2.5.1	Why To Use it?	36
3	Data Set	40
3.1	Topic Selection	41
3.2	Topic 336: Black Bear Attacks	41
3.3	Topic 310: Radio Waves and Brain Cancer	42
3.4	Topic 383: Mental Illness Drugs	43
3.5	Topic 436: Railway Accidents	44
3.6	Document Selection	45
4	Think-Aloud Study	49
4.1	Think-Aloud Method	49

4.2	Study Design	50
4.2.1	Study Protocol	50
4.2.2	User Interface	52
4.2.3	Measuring Judging Behavior	53
4.2.4	Participants	55
4.3	Cleaning of Data	57
4.4	Certainty in Relevance judgments	58
4.4.1	Low Certainty Relevance judgments	59
4.4.2	Medium Certainty Relevance judgments	60
4.4.3	Certain Relevance judgments	61
4.4.4	Frequency of Certainty Levels	62
4.5	Making Incorrect Relevance judgments	63
4.5.1	Completion Time for Incorrect Relevance judgments	63
4.5.2	Categories of Incorrect Relevance judgments	64
4.6	Judging Behavior	72
4.7	Short vs. Long Search Topics	74
5	Low Consensus Documents	78
5.1	Summary	78

5.2	Definitions	82
5.2.1	Assessor Causes of Differences	82
5.2.2	Assessor Decision Making	84
5.2.3	Findings	86
5.3	Topic 336: Black Bear Attacks	87
5.4	Topic 310: Radio Waves and Brain Cancer	94
5.5	Topic 383: Mental Illness Drugs	103
5.6	Topic 436: Railway Accidents	111
6	High Consensus Documents	118
6.1	Relevant Documents	119
6.1.1	Summary	119
6.1.2	Key Findings	121
6.1.3	Topic 336	121
6.1.4	Topic 310	130
6.1.5	Topic 383	138
6.1.6	Topic 436	145
6.2	Non-Relevant Documents	150
6.2.1	Summary	150

6.2.2	Key Findings	151
6.2.3	Topic 336	151
6.2.4	Topic 310	155
6.2.5	Topic 383	160
6.2.6	Topic 436	164
7	Certainty Interfaces	170
7.1	Methods and Materials	171
7.1.1	Study Protocol	171
7.1.2	User Interfaces	172
7.1.3	Participants	175
7.1.4	Latin Square	176
7.1.5	Measuring Accuracy Rate	176
7.2	Assessors' Judging Behavior	177
7.2.1	Low consensus Documents	177
7.2.2	High Consensus Documents	180
7.2.3	Low vs. High Consensus Documents	182
7.3	Analysis of Time	186
7.3.1	Binary and Ternary Certainty Interfaces	187
7.4	Conclusion	189

8 Conclusion and Future Directions	190
8.1 Conclusion	190
8.1.1 Certainty in Relevance Judgments	192
8.1.2 Categories of Incorrect Relevance Judgments	192
8.1.3 Low Consensus Documents	195
8.1.4 High Consensus Documents	196
8.1.5 Certainty Interfaces	197
8.2 Contributions	199
8.2.1 Limitations	201
8.3 Future Work	201
References	205
APPENDICES	218
A Forms Used in the Studies	219
A.1 Think-Aloud Forms	219
A.1.1 Information and Consent Forms	219
A.1.2 Questionnaire	222
A.2 Certainty Interfaces Study Forms	225
A.2.1 Information and Consent Forms	225

A.2.2 Questionnaire	228
-------------------------------	-----

List of Tables

3.1	TREC topics used in the studies	41
3.2	Distribution of Documents in Low and High Consensus Level	47
3.3	Probability of Relevance and number of assessors for each document used in our studies	48
4.1	Confusion Matrix	54
4.2	Certainty levels and their frequency	63
4.3	Relevance judgments and Average Completion Time	65
4.4	Categories of Incorrect Relevance judgments	69
4.5	Judging Behavior	74
4.6	Assessors' Behavior with short and long topics' description	77
5.1	Causes of Differences and Assessors Decision Making For Low Consensus Documents	86
7.1	Confusion Matrix	177

7.2	Results. This table reports the percent of judgments made with binary certainty, as well as the accuracy of the relevance judgments. The standard error for each percentage is reported as well. Results are shown for low consensus documents.	178
7.3	Results. This table reports the percent of judgments made with binary certainty, as well as the accuracy of the relevance judgments. The standard error for each percentage is reported as well. Results are shown for low consensus documents.	179
7.4	Results. This table reports the percent of judgments made with binary certainty, as well as the accuracy of the relevance judgments. The standard error for each percentage is reported as well. Results are shown for high consensus documents.	181
7.5	Results. This table reports the percent of judgments made with binary certainty, as well as the accuracy of the relevance judgments. The standard error for each percentage is reported as well. Results are shown for high consensus documents.	182
7.6	Results. This table reports the percent of judgments made with binary certainty interfaces, as well as the standard error at each level of certainty. Results are shown for low and high consensus documents.	184
7.7	Results. This table reports the percent of judgments made with ternary certainty interfaces, as well as the standard error at each level of certainty. Results are shown for low and high consensus documents.	185

7.8	Results. This table reports the average judging time in seconds with different levels of certainty, as well as the standard error. Results are shown for the binary certainty interfaces.	188
7.9	Results. This table reports the average judging time in seconds with different levels of certainty, as well as the standard error. Results are shown for the ternary certainty interfaces.	188

List of Figures

1.1	Cranfield Paradigm as Followed in TREC	4
3.1	Topic 336	42
3.2	Topic 310	43
3.3	Topic 383	44
3.4	Topic 436	44
4.1	The user interface for judging a single document	53
4.2	Expressions used in conjunction with making a relevance judgment	60
4.3	Relevance judgment Process	71
5.1	Possible Relevant Sentences for DocID: APW20000323.0200	87
5.2	Attractive Sentences for DocID: APW20000703.0186	93
5.3	Possible Relevant Sentences for DocID: NYT20000224.0139	95
5.4	Possible Relevant Sentences for DocID: XIE19970506.0203	98

5.5	Possible Relevant Sentences for DocID: XIE20000628.0163	101
5.6	Attractive Sentences for DocID: NYT19990121.0380	104
5.7	Relevant Sentences for DocID: NYT19991214.0159	107
5.8	Relevant Sentences for DocID: NYT19991206.0109	109
5.9	Attractive Sentences for DocID: APW19990914.0022	112
5.10	Attractive Sentences for DocID: XIE19980303.0229	114
5.11	Possible Relevant Sentences for DocID: NYT19991206.0299	116
6.1	Relevant Sentences For DocID: APW19990809.0179	122
6.2	Possible Relevant Sentences For DocID: NYT20000706.0242	125
6.3	Relevant Sentences For DocID: NYT20000602.0371	129
6.4	Possible Relevant Sentences For DocID: APW20000608.0153	131
6.5	Possible Relevant Sentences For DocID: NYT19991003.0452	134
6.6	Possible Relevant Sentences For DocID: NYT19991025.0333	136
6.7	Relevant Sentences For DocID: NYT20000925.0105	139
6.8	Relevant Sentences For DocID: XIE19991207.0246	145
6.9	Relevant Sentences For DocID: XIE19990802.0027	147
6.10	Relevant Sentences For DocID: XIE19981020.0034	148
7.1	First Certainty Relevance Judgments Interface	172

7.2	Third Certainty Relevance Judgments Interface	173
7.3	Second Certainty Relevance Judgments Interface	173
7.4	Fourth Certainty Relevance Judgments Interface	174

Chapter 1

Introduction

Immediately upon opening any textbook, website, or reading any article, or any other source of information in Information Retrieval (IR), you will find that terms such as “relevance”, “relevant”, and “relevancy” are among the most frequently used terms. This is in fact entirely predictable. The reason is that relevance has in the past and will be linked with all work in the area of information retrieval. To IR researchers, the ultimate goal is to provide searchers with what they are looking for, and as a result, satisfy their information need. However, the unstructured nature of information, the continuous growth of the data volume (particularly electronic data), and the complexity of new technologies make the process of finding relevant information a complicated task.

Research in IR is divided into two main approaches. First is the system-driven approach in which greater emphasis is placed on developing algorithms and techniques that are able to find relevant pieces of information which satisfy users’ information need. Second is the

user-oriented approach, where users are the main targets of research. This approach pays greater attention to the behavior of users and the cognitive processes they go through when performing search tasks (Borlund, 2003b).

The integration of both of these approaches is important in order to foster development in IR. Researchers who work more frequently with the second approach, study users' behaviors, cognitive processes, and their information need when performing searches. On the other hand, researchers who work with the first approach take advantage of the outcomes of research in the second approach in order to develop or alter search systems, and subsequently to better suit users' information needs. The work in this thesis falls mainly under the second approach in IR, which is user-oriented.

1.1 Background Overview

1.1.1 IR Evaluation with Test Collections

How can we evaluate IR systems and know that System A for example is better than System B? Researchers and developers might invent excellent algorithms that are capable of performing very well in retrieving relevant documents, but how can we know if these algorithms are working better than others?. In the 1960s and after working on a 10-year project called the "Cranfield Project", Cyril Cleverdon initiated the Cranfield Paradigm, which has become the main paradigm to be used in evaluating IR systems. In the Cranfield paradigm, researchers use what are called "test collections" and evaluation metrics such

as recall and precision, where the recall is the fraction of relevant documents that are retrieved, while precision is the fraction of retrieved documents that are relevant.

A typical test collection in the Cranfield paradigm consists of three main components: a set of documents, a set of search topics (also referred to as information need in IR), and a set of relevance assessments (Baeza-Yates et al., 1999). Each search topic is run against the set of documents, and as a result a list of documents is retrieved. Subsequently, the retrieved documents are pooled (top-k documents) and then judged by assessors. These assessors generally are the same individuals who create these topics. After making relevance judgments, the metrics we mentioned above and others, are used in comparing and distinguishing between IR systems (Voorhees et al., 2005; Sanderson, 2010). Figure 1.1 illustrates the Cranfield Paradigm that is followed in TREC¹ when evaluating IR systems using test collections.

When test collections were first created, they did not contain more than several thousand documents and a few hundred topics. For instance, the Cranfield collection, which was created in the late 1960s, consisted of 1400 documents and 225 search topics (Harman, 1993). Over the years, and to satisfy the need to evaluate IR systems in more contemporary situations, where large full-text searches are required, the number of documents in test collections has increased dramatically. A typical test collection in the present day would range from millions to billions of documents such as the ClueWeb12 Dataset, which its compressed size is 5.5 terabytes and the uncompressed size is 27.3 terabytes.

¹TREC refers to Text REtrieval Conference. It is an annual workshop started in 1992 and sponsored by National Institute of Standards and Technology (NIST) and the U.S. Department of Defense. The aim of designing TREC is to speed up the process of transferring the necessary technology and resources for evaluating large-scale systems into the commercial sector (Voorhees et al., 2005).

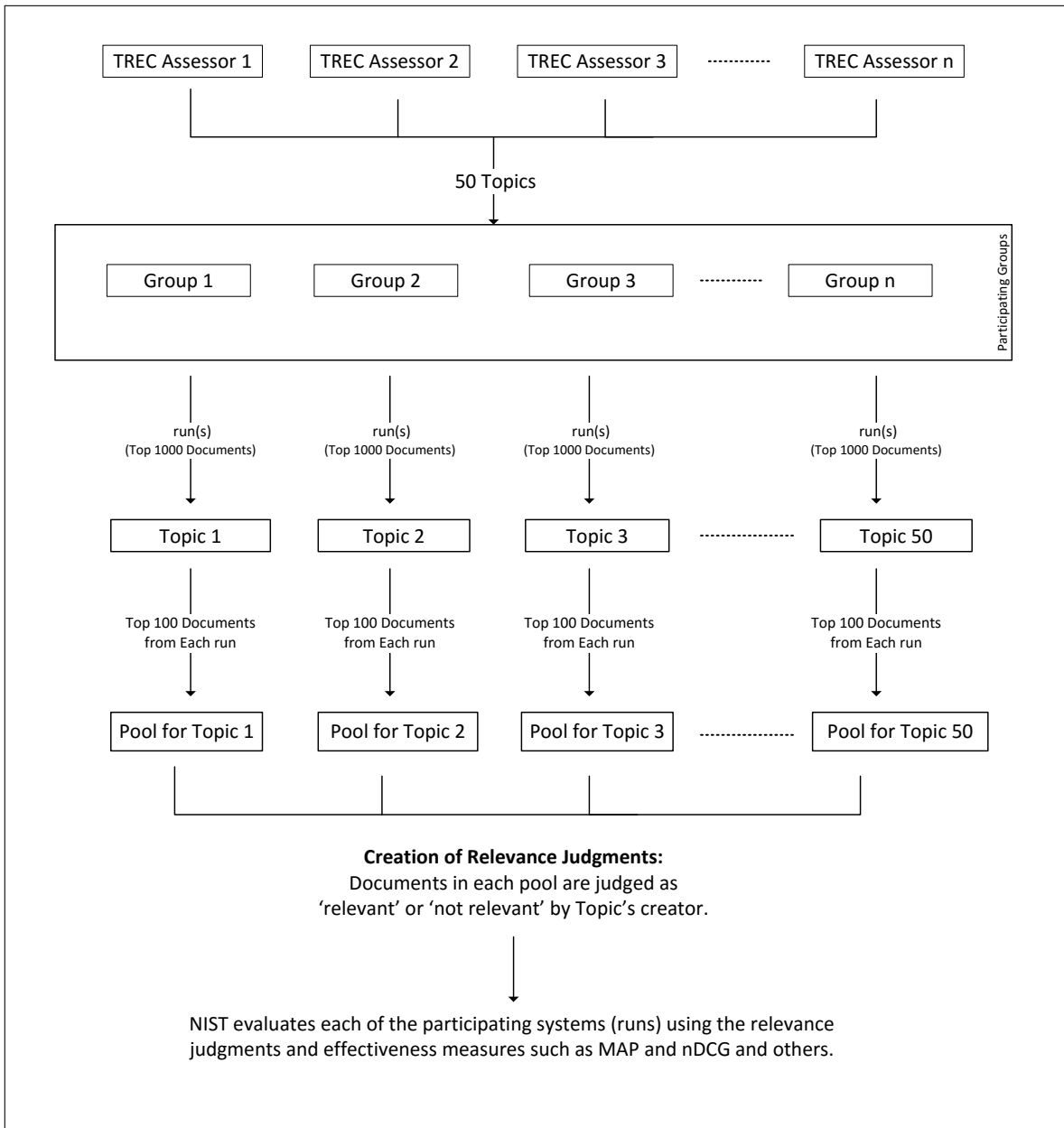


Figure 1.1: Cranfield Paradigm as Followed in TREC

1.1.2 Primary Assessors vs Secondary assessors?

Primary assessors are those who originate the search topic. They may be researchers, lawyers, or experts in other fields. They generate search topics for different reasons that serve their needs, such as testing search systems or finding important pieces of information for a lawsuit, as in legal e-discovery. Even though primary assessors have a good level of knowledge of the search topic, they might make mistakes when assessing the relevance of documents (Grossman and Cormack, 2011).

On the other hand, secondary assessors are individuals who are hired to fulfill a specific job that is assigned to them. Their job is to judge relevance of documents to a given search topic. These search topics are given to them and they are provided with description of the search topic and guidelines that they have to adhere to when making relevance judgments. That means their role is limited only to what is written in the description of the search topic or the instructions that are given to them. They might or might not have background on the search topics that are assigned to them. This may cause them not to understand fully the intent of the search topic given to them (Kinney et al., 2008). They also may not have training in judging relevance of documents except what is given to them on the tutorial and qualification test before working on the assigned tasks.

In this thesis, our interest is to gain a better understanding of secondary assessors' relevance judging behavior, and the causes of judging disagreements among them. Secondary assessors are a great asset to the IR evaluation process since they can assist in producing relevance assessments that will help in testing IR systems.

1.2 Problem Statement

1.2.1 Disagreement on Relevance judgments

Research has shown that secondary assessors produce different relevance judgments (Voorhees, 2000; Harter, 1996). In fact, the same assessor has been also found to produce different relevance judgments at different times during the relevance judgments session (Schamber, 1994). There are a number of studies that address the factors that influence assessors' relevance judgments (Barry, 1994; Schamber, 1994; Park, 1993; Saracevic, 2007). However, within the factors themselves, some are more influential than others (Chu, 2011). In addition, researchers have studied the potential causes behind the disagreement on relevance judgments. For instance, Webber et al. (2012a) suggested that causes such "assessor inattention", "differing relevance conceptions", and "variable threshold for detecting relevance" have an impact on assessors' relevance agreement. Also, assessors' levels of expertise have been found to impact the level of assessors' relevance judgments disagreement (Wang and Soergel, 2010; Wang, 2011). Furthermore, researchers have gone one step further and built models to predict the differences in assessors' judgments (Chandar et al., 2013).

The level of agreement between assessors and the impact of variations in relevance judgments on IR systems' performance has been studied (Lesk and Salton, 1968; Cleverdon, 1970; Burgin, 1992; Cormack et al., 1998; Voorhees, 2000; Sormunen, 2002; Trotman and Jenkinson, 2007; Bailey et al., 2008; Kinney et al., 2008; Carterette and Soboroff, 2010; Li and Smucker, 2014). In all of these works, the authors discuss how the variations could or could not impact the performance of IR systems. For example, (Voorhees, 2000) discusses

how the variations in relevance judgments do not affect the relative performance even though the absolute performance is affected. However, the assessors in Voorhees' study were all experts in judging documents since they were TREC assessors and UWaterloo assessors who are experts in IR and judging documents as well. However, this does not reflect a real world user that might be recruited to make relevance judgments (Harter, 1996).

With the advent of technology and the dramatic increase in test collections' sizes, the process of making relevance judgments have become more time consuming and more expensive. Therefore, researchers have become more interested in recruiting secondary assessors to help construct test collections more quickly and in a more cost-effective manner. However, these assessors, as mentioned earlier, might not have the domain expertise and are not experts in judging documents. How do these types of assessors would impact the performance of IR systems when their relevance judgments are used? Bailey et al. (2008) conducted a study where three groups of assessors were used to make relevance judgments. They found that relevance judgments made by the group that neither created the search topics nor had domain expertise (they refer to them as bronze standard judges) do impact the ranking of IR systems. The results found in (Bailey et al., 2008) revealed to us how the sets of judgments made by assessors who lack the domain expertise could impact the performance of IR systems.

Also, in similar studies where errors in sets of judgments were simulated, both Carterette and Soboroff (2010) and Li and Smucker (2014) found that conservative judging behavior and pessimistic model increase the level of rank correlation; however, Li and Smucker

(2014) used additional effectiveness measures and found that some of these measures were affected more by the errors in relevance judgments.

Understanding the sources of disagreements allows us to obtain fundamental knowledge about the relevance judgment process, which can have implications for the design of systems that collect relevance judgments. Furthermore, studying secondary assessors judging behavior is important for legal e-discovery where secondary assessors can be recruited to find what is called “responsive documents” to the production request. Responsive documents and productions requests are legal jargons which refer to relevant documents and search topics respectively in IR (Oard et al., 2010).

From the above paragraphs, we can realize the importance of studying the judging behavior of secondary assessors. However, most of the published work focuses on studying this issue exclusively from the quantitative approach, and only sometimes did researchers use certain qualitative research tools, such as interviews or questionnaires, to gain a greater understanding of assessors’ relevance judgments. To the best of our knowledge, none of the published work conducted mainly a qualitative study to deeply understand the judging behavior of secondary assessors. What we mean by this is that the researchers observe and record what assessors articulate while they are engaged in the relevance judgment process. We believe that there is a need to know what kind of thoughts go through assessors’ minds while they perform search tasks. Not capturing this results in losing a more profound understanding and an instant access to what stimulates relevance judgments and what influences those judgments as well. In order to develop more appropriate search systems that are capable of providing more levels of satisfaction to users, a deeper access to assessor’s thoughts must take place.

Therefore, we conducted a qualitative study (Think-aloud study), which forms our first user study in this thesis, to deeply understand the judging behavior of secondary assessors. Many interesting findings were revealed by the help of the Think-aloud data as we will discuss in the coming chapters of this thesis. For example, we were able to know not only if assessors make guesses during relevance judgments but also why. From previous studies that have been published (Jethani, 2011; Smucker and Jethani, 2011b, 2010), and based on the measurements used in a university lab, Intentional random decisions are found to happen very few times or not at all. By listening to assessors' stream of thoughts while they work on the assigned search tasks, we found useful information that can help the IR community to better understand the relevance judging behavior of secondary assessors.

1.2.2 Uncertainty in Relevance Judgments

In many of experiments in IR, it is common to have a binary scale when asking assessors to make relevance judgments (Voorhees et al., 2005). Even though the binary scale has been widely used in IR experiments since the last century, it might not be the best scale to satisfy assessors' judgments (Kekäläinen, 2005). From the results of our first user study (Al-Harbi and Smucker, 2014), we found that assessors expressed different levels of uncertainty when judging documents. The analysis of the verbal protocols and the type of expressions assessors used in the study conveyed to us that the common binary relevance scale is not capable of capturing the uncertainty in assessors' relevance judgments and as a consequence does not truly reflect assessors judgments.

Moreover, by providing just two options from which to choose (relevant or not-relevant), we force assessors to choose an option that might not represent accurately their decisions and satisfy their needs. Therefore, we are not getting accurate relevant judgments. [Sormunen \(2002\)](#) found that 50% of relevant documents were judged as partially irrelevant when a four-level scale was used. However, there are times when researchers used a multi-level relevance grading scale, such as a three-level or four-level, up to eleven-level scale ([Kekäläinen and Järvelin, 2002](#)).

Our goal in the second user study, which forms the second part of the work presented in this thesis, is not to update the current binary relevance scale; however, it is to add the uncertainty factor to the binary relevance scale in order to capture the uncertainty in assessors' relevance judgments. The design of the second user study was based on the results we have discovered in our previous study ([Al-Harbi and Smucker, 2014](#)). The results in the second user study allowed us to find the tendencies in secondary assessors relevance judgments when self-reporting their certainty levels. Chapter 7 discusses these results and more in detail.

1.3 Thesis Overview

While the study of relevance judgments and the variations between them has existed for many years, little research existed regarding the thoughts that go in secondary assessors' minds when they work on assigned search tasks. There is a need to understand more how secondary assessors perform the assigned search tasks and what they think of when making relevance judgments. Our work in this thesis adds to the current knowledge and existing

research and further fills the gap of our understanding of secondary assessors' relevance judging behavior. The entirety of the work in this thesis is divided into three main parts.

Two parts of this thesis are represented by two user studies. The first study, which forms the main part of this thesis, was a think-aloud study, and provides the first of such qualitative studies of secondary assessors judging behavior. The second study of the research was to capture the uncertainty in secondary assessors relevance judgments. The third part is represented by a further examination of the behavior of secondary assessors when judging multiple types of documents. This part was carried out based on the data from the think-aloud study.

We discuss in Chapter 3 the criteria that we followed and applied when choosing the documents and search topics for our two user studies. We explain in detail the two groups of consensus we used and what conditions that qualify the documents to be either at the low consensus group or the high consensus group.

In Chapter 4, based on the data collected from the think-aloud study, we look at the different levels of certainty in assessors' relevance judgments. In this study, we used a binary relevance scale and assessors were instructed to think aloud while they work on their assigned tasks. They worked without any interruptions from the researcher side and they only were encouraged to continue thinking aloud when there was a lengthy period of silence. Based on the transcribed data and the recording videos, we categorized assessors' relevance judgments into three levels of certainty. We show also how the binary relevance scale that we used is not capable of capturing those certainty levels. We also discuss

assessors performance and their error rates and give interpretation for the results we found supported by examples from the transcribed data.

In Chapter 5 and 6, low and high consensus documents are further examined. We study each document we used in our first user study separately and analyze assessors' relevance judgments in regard to it. This examination was based not only on the transcribed data but also on the recorded videos. Supported by examples from the transcribed data, we discuss the causes and reasons that we found assessors to state or express when making relevance judgments. At the beginning of Chapters 5 and 6, we provide a summary and key findings table for the results we found in the whole chapter.

Chapter 7 is devoted to our second user study which forms the second part of this thesis. Four new interfaces were designed in order to investigate how secondary assessors judge documents with certainty. We study in greater depth the behavior of secondary assessors when incorporating the uncertainty factor with the typical binary relevance scale. The interfaces we used in the study fall into two groups, with two interfaces in each group: binary certainty interface group, and the ternary certainty interface group. All of the interfaces have the same answers for the first question about relevance judgments, either relevant or not relevant. However, the answers for the second question, which is about the level of certainty, differ in each interface. Each interface in the binary group has a different set of words at each level of certainty. Likewise, each interface in the ternary group has a different set of words at each level of certainty. In this chapter, we computed the percentage of relevance judgments and the accuracy rate at each level of certainty.

1.4 Contributions

In this thesis, we make the following contributions:

1. We found secondary assessors to judge relevance of documents at different levels of certainty. These levels of certainty vary from low certainty to the other end of the spectrum of high certainty. (Chapter 4)
2. We show the differences between primary and secondary assessors. These can be divided into four categories, according to: difficulty in applying the search topic, difficulty in processing the document, secondary assessor factors, and true error in primary assessor judgment. (Chapter 4)
3. We found preliminary evidence regarding the impact of the number of details in a search topic's description on a secondary assessors judging behavior. In examining the differences in relevance judgments at a per-topic level, we noticed a difference between topics with long descriptions and topics with short description in regard to the rates at which participants misidentified relevant documents as non-relevant and vice versa.(Chapter 4)
4. We show that assessors differ in their ability to articulate their thoughts; however, this does not impact their performance in judging relevance of documents. Assessors (participants) are divided into three groups: slow, medium, and fast deciders. However, being in one group does not indicate superiority in performance. (Chapter 4)

5. We found assessors' age to affect the understanding of certain search topics. Young assessors, who are mostly undergraduate students, were not able to understand some old technologies (like car phones) since they have not heard about them. (Chapter 4)
6. We found that the length of a document and location of the relevant material in the document to play an essential role in making it low consensus. This is not only an issue for a secondary assessor but also for a primary assessor as well. (Chapter 5)
7. We show how factors such as lack of familiarity/knowledge, lack of understanding the search topic, absence of keywords and other reasons could be a source of the variations in relevance judgments. (Chapter 5 & 6)
8. We show how the length of search topics may further contribute to the issue of variations in the judgments. We provide a number of examples about how assessors only consider one part of the description and forget about the others. (Chapter 5 & 6)
9. We show that relevance judgments are occasionally based on incorrect perception. Our data shows that when an assessor judges a document correctly, that does not mean his/her decision is based on correct perception. He/she might misunderstand the search topic or the content of the document. (Chapter 5 & 6)
10. We found that secondary assessors tend to be certain in their relevance judgments regardless whether a document is high or low consensus. That is found in all the certainty interfaces we used in our second user study. Assessors produced more certain relevance judgments than uncertain ones. (Chapter 7)

11. We found when judging high consensus documents, assessors' accuracy to be lower when making low certainty relevance judgments, and the judgments to be more accurate and tended to agree with NIST assessors when making high certainty relevance judgments. (Chapter 7)
12. We found when judging low consensus documents, assessors' accuracy to be low regardless of their certainty level. (Chapter 7)
13. We found certainty to be lower when judging low consensus documents using the ternary certainty interface. (Chapter 7)

Chapter 2

Literature Review

2.1 The concept of Relevance

The concept of relevance is at the heart of research in the area of information retrieval (IR) (Croft et al., 2010). Researchers and other experts in this field have proposed several interpretations of relevance based on their understanding of this concept. A number of published works on the concept of relevance have been published during the last decade of the last century (Froehlich, 1994; Green, 1995; Harter, 1992; Mizzaro, 1998; Park, 1994; Saracevic, 1996; Schamber et al., 1990; Cosijn and Ingwersen, 2000a). In all of these works, the concepts of relevance and notions of system-based relevance, and user-based relevance have been discussed at greater length. What most researchers have been trying to illustrate is that the concept of relevance is not limited only to the topical relevance (system-based relevance) (Borlund, 2003a).

In fact, it has much wider meaning. Situational relevance for example, which is a type of user-based relevance, is a representation of the relevance that might be more realistic to users than topical relevance, and as a result, it is thought to be more useful for interactive information retrieval (Borlund, 2000; Borlund and Ingwersen, 1997). In this type of relevance, matching words between the user's query and the document might not satisfy the user's information need. A document may be a perfect match to the query; however, a user may find it irrelevant for several reasons, such he/she has previously read the material in the document, the document is outdated (old), or the source of the document is not reliable.

Despite all of the advancement in the understanding of the concept of relevance, there is still no complete agreement on it, as pointed out by Mizzaro (1998). A paper that focuses only on the concept of relevance, Borlund (2003a) discusses this issue in detail. She talks about the different grades, levels, and classes of relevance. She states that there are two main classes of relevance, which are associated with the two main approaches followed in IR. The two main classes are: (1) objective, or system-based relevance; and (2) subjective, or human (user)-based relevance. Additionally, she discusses the nature of relevance in this paper, examining how relevance is dynamic during the information-seeking process session. Moreover, Kuhlthau (1991) addresses the concept of dynamic relevance in greater detail, when she states that the relevance preferences of users may vary at different points in time during the relevance judgment process. However, the notion of dynamic relevance, though it might be true when we discuss information seeking and primary assessments, it would be difficult to consider when dealing with relevance assessments made by secondary

assessors. This is stated, since they are hired only for the purpose of judging documents, and merely required to follow what is given to them in the description of the search topic.

Büttcher et al. (2010) defines the notion of a relevant document in the following statement: “a document is considered relevant to a given query if its contents (completely or partially) satisfy the information need presented by the query”. Therefore, satisfying a user’s information need is important in order to consider whether or not a document is relevant. This definition does not limit the meaning of relevance to the topical relevance (system-based relevance); however, it centers on the satisfaction of the information need of the user. Therefore, retrieving documents that are on topic, but do not satisfy the user’s information need, may not be considered relevant.

Though the concept of relevance is important to the work presented here, it is beyond the scope of this thesis to go into any further detail about this concept. Rather, the remainder of this chapter will focus on the relevance judgments, the disagreement assessors encounter when evaluating documents, the factors that influence assessors’ relevance judgments, crowdsourcing and why it is important to the IR community and toward the end of the chapter we will give an overview about some of the methods used in generating verbal reports and in particular the think-aloud method and its importance.

2.2 Relevance Judgments

2.2.1 Disagreement about Relevance Judgments

As mentioned earlier, relevance judgments are one of the key components of a test collection. The evaluating process cannot be carried out without the existence of relevance judgments. Therefore, its importance to evaluating IR systems is obvious. However, there are variations in relevance judgments (Voorhees, 2002). Assessors were found to produce different relevance judgments, as we will illustrate in the following paragraphs. This observation has made researchers in IR worried about the correctness of their effectiveness measures. Therefore, there has been effort to study the impact of these variations on the performance of IR systems, as well as the causes of these variations. In the following paragraphs, we will first discuss the impact of the variations of relevance judgments on IR, and secondly we will consider the reasons and causes researchers think stand behind these variations.

Variations in relevance judgments has been observed since the 1950s (Gull, 1956). Barhydt (1964, 1967); Rees and Schultz (1967); O'Connor (1969) have also observed the variations in relevance judgments and how this disagreement could raise question about the validity of results obtained when evaluating the performance of IR systems.

Barhydt (1967) studied the performance of two groups of assessors: subject experts and system specialists. The two groups were given abstracts and asked to judge them based on given search topics. These assessors are considered as secondary assessors since they are not the originators of the search topics. They are hired only to judge the relevance of the

abstracts. Based on his analysis of the collected data, no relationship was found between the type of group and performance, though their ranges of effectiveness were wider for the subject experts group. The used effectiveness measure was based on the sensitivity and specificity scores.

Janes and McKinney (1992) conducted a qualitative study on the difference between the judgments made by different groups of assessors (secondary assessors, those who did not originate the search topic). They divided their participants into three groups with 4 participants in each group. The study found that secondary relevance judgments were comparable to those made by the users who created the search topics (primary assessors). Janes (1994) conducted another study where he recruited 48 participants (assessors) divided into three groups as well: incoming students to a school of information/library (SIL) science, continuing students in SIL, and academic librarians. The judges were instructed to assess documents based on either relevance to the query, their similarity to the search topic, or their perceived usefulness to the users. His observation was that secondary assessors made comparable judgments to the ones produced by users (primary assessors). Also, Schamber (1994) discusses the issue of producing different relevance judgments. She points out that the same assessors produce different relevance assessments at different times during the search seeking process.

When comparing between relevance judgments made by TREC assessors and a user study participants, Al-Maskari et al. (2008) found a high level of agreement between the TREC assessors and the assessors in their user study (63%) when identifying relevant documents. There were 56 participants, who were asked to judge documents for 56 search topics. They provide explanation for the variations in relevance judgments made by their

user study's assessors and TREC assessors. They ascribe these variations to: (1) the users tendency in being more liberal due to the few number of relevant documents they found, (2) the difficulty of some search topics.

Surprisingly, [Efthimiadis and Hotchkiss \(2008\)](#) found that assessors who did not have legal expertise were much better at assessing the relevancy of documents than those with legal expertise. They asked assessors to judge the relevancy of documents for search topics that were designed for the Legal Track Interactive Task Challenge.

[Wang and Soergel \(2010\)](#) and [Wang \(2011\)](#) examined if background has an impact on assessors relevance judgments agreements. In their studies, 8 students in total participated: 4 with a law background and 4 with a library and information studies (LIS) background. They found that although the two groups were from different disciplines (law and LIS), both produced almost identical relevance judgments. However, the LIS group judged the non-relevant documents slightly less accurately than the law group.

In a qualitative study, [Grossman and Cormack \(2011\)](#) investigated two points: (1) if the vagueness of production requests (search topics) and its application is responsible for assessors disagreement, or (2) if human error is the cause of the disagreement in assessor relevance judgments. They determined that most of the disagreements in relevance assessments are due to human error. The results of their work led them to assert that the criteria of relevance that assessors have to apply is sometimes well defined (they pointed out the production request and assessment guidelines used at TREC 2009). They argue that most of the incorrect judgments are the result of human error. They also reported

that the TA (Topic Authority who is a senior lawyer familiar with the subject matter) as a human assessor might produce contradicting relevance assessments.

[Smucker and Jethani \(2011a\)](#) study assessors accuracy by comparing user study participants and NIST assessors. NIST assessors, who are considered usually to be primary assessors, were able to distinguish between relevant and non-relevant documents better than the user-study participants. However, the fraction of relevant documents judged as relevant were found to be comparable between the two groups of assessors; they also indicated that user study participants showed more liberal behavior when assessing documents than NIST assessors who were more conservative.

In a study conducted to test the quality of relevance judgments based on the query ownership when evaluating search engines, [Chouldechova and Mease \(2013\)](#) found that the ownership of query does impact the quality of relevance judgments positively. Since identifying the query owners in the context of search engines is impossible for some practical reasons, assessors were chosen and instructed to recall their recent queries and in this way the authors could link the queries with their owners. Their results show that the query owners made better relevance judgments than the ones made by non-query owners. In this experiment, query owners are basically primary assessors, who created the search topics (queries) while non-query owners are secondary assessors, who are only recruited to make relevance judgments.

The impact of differences between secondary and primary assessors on the quality of a text classifier was also studied ([Webber and Pickens, 2013](#)). Though the quality of

classification was negatively impacted when using secondary assessments, the effects on the results ranking was minimal.

Wakeling et al. (2016) conducted a study to examine differences between primary and secondary assessors. They emphasize the use of what they refer to as “real-life” search topics. Based on their review of a number of previous works, they think that the search topics used in information retrieval experiments are outdated. The aim of their work is to answer five research questions:

1. How does relevance assessment behaviour differ between primary and secondary assessors?
2. To what extent do secondary assessments agree with primary judgments?
3. To what extent do interest in and knowledge of the topic affect relevance judgments?
4. Does the length of the topic description affect secondary relevance judgments?
5. How does the level of confidence in judgments differ between primary and secondary assessors?

They determined that assessors viewed the description of secondary search topics with greater frequency than their own search topics. The scope and the details of the search topic instructions occasionally were found to be unclear to secondary assessors, and this explains the reasons that secondary assessors sometimes have required additional time to judge documents for secondary search topics. Moreover, unfamiliar vocabulary and terminology were found to be a significant source of confusion for assessors when they work

on secondary search topics. They found the percentage of agreement between primary and secondary assessors on all topics to be 79%, with variation between the open and closed topics. Moreover, it has been determined that closed topics were less challenging to judge compared to open topics. Open topics require more interpretation and discussion of the material in order to identify the relevant information.

When it comes to the levels of interest and knowledge of a search topic, primary assessors were at higher levels than secondary assessors. This result in fact, was anticipated since they are more knowledgeable about the topics which they themselves create, and are also more interested in relevant literature. The authors include some extracts from post-session interviews in which participants explain and justify their responses and behavior. In addressing their fourth question, the length of a topic's description was found to have no impact on the number of "relevant" decisions made by assessors. Finally, the level of confidence was found to be higher for primary assessors than secondary assessors. The authors' work includes several extracts from assessors' post-sessions interview responses. These interviews uncover interesting information about both primary and secondary assessors.

2.2.2 Impact of Disagreement on the Performance of IR Systems

The findings that were talking about the variations in relevance judgments (as seen the previous paragraphs) raised the concern and posed questions about the validity of the measurements used to evaluate the effectiveness of IR systems. The first researchers who led the work in studying the effect of the disagreement of relevance judgments on IR systems were Lesk, Salton, and Cleverdon. Their work was in 60s and 70s of the last century where

the sizes of test collections were small (thousands of documents). [Lesk and Salton \(1968\)](#) studied the impact of different relevance judgments sets on the performance of retrieval systems. They used four sets of relevance judgments. The first two sets were produced by experts and non-expert assessors while the third and the fourth sets were either an intersection of the first two sets or combination of them. What they ultimately found was that even though there was just 30% agreement on relevance judgments between the four sets, the stability of retrieval systems' performance was not impacted. Lesk and Salton provided three points that explain why the variations in relevance judgments do not impact system ranking. First, the reported scores of evaluation are averaged over many topics rather than individual topics. Second, the variations in relevance judgments are mostly caused by the borderlines documents that have lower rank level. Lastly, the variation in the relevance judgments sets may be of a little impact on the recall and precision since these measures rely on the relative position of the relevant and non-relevant documents in the ranking list.

In another study, [Cleverdon \(1970\)](#) used four relevance judgments sets as well to test if they had an impact on the ranking of nineteen index methods. Even though the sets were different, they did not have an impact on the ranking of the index methods that were used in the experiment. In a study similar to that conducted by Lesk and Salton's, [Burgin \(1992\)](#) used four sets of relevance judgments and six different document representations. These four set of relevance judgments were made by different groups. Therefore, aim was to test how the varied relevance judgments made by these different groups could affect the evaluation of retrieval systems. His findings do comply with the findings of ([Lesk and](#)

Salton, 1968) that the variations in relevance judgments have no impact on the relative performance of retrieval systems.

One of the extensively cited works in IR is the work of Ellen Voorhees on the variations in relevance judgments and the measurement of retrieval effectiveness Voorhees (2000). She conducted a study on the impact of different relevance judgment sets on the performance on retrieval systems, but this time with larger test collections. According to Voorhees, previous studies that touched on the same issue (Lesk and Salton, 1968; Cleverdon, 1970) were based on relatively small test collections (less than 1300 documents). However, TREC test collections are many times larger than this as in the ClueWeb Dataset, which contains 733,019,372 documents. Her goal was to test if what previously published work has found about the stability of the comparative evaluation of retrieval performance would still hold true, despite the size of the test collection. The study is composed of two parts. One with TREC-4 relevance assessments and the other one with TREC-6 relevance assessments. In TREC-4 relevance assessments, three NIST assessors were used to construct the relevance judgments sets. Each topic was judged by three assessors, one primary assessor who represents the topic's creator, and two secondary assessors, who are not the topic creators. The agreement between these three assessors were in fact higher than what was reported in the previous studies by Lesk, Salton or Cleverdon. Voorhees ascribed this finding to: (1) similarity in background between NIST assessor, (2) equal level of training in judging documents and (3) identical judging conditions. When the 33 systems were evaluated using these relevance assessments sets, the relative performance on retrieval systems was stable. In TREC-6 relevance assessments, the aim was to determine whether the similarity in background was behind the stability of the evaluation results. Therefore, Voorhees

obtained relevance assessments made outside of NIST. These assessments were made by a participating group in TREC-6 from the University of Waterloo. The rationale behind this choice is that the group from Waterloo had different backgrounds from NIST. Likewise with TREC-4 relevance assessments, the same was found in regard to the relative performance on retrieval systems. Therefore, the results of the Voorhees study reaffirm to the IR community that the comparative evaluation of retrieval performance will remain stable even when variant relevance assessments sets are used; consequently, the TREC test collections will remain as a valid tool for comparative retrieval experiments.

In an effort to investigate if the domain expertise would play a role in the quality of relevance judgments, [Bailey et al. \(2008\)](#) conclude that assessors who are experts, have a tendency to make more accurate relevance judgments than non-expert assessors. They used three groups of assessors in the study: gold, silver and bronze. Gold assessors are those who created the search topics and are experts in search tasks. Silver assessors are merely experts in search tasks. The bronze group consists simply of assessors who were neither creators of the search topics, nor experts in search tasks. They observed that there is a low level of agreement between the three groups, especially the level of agreement between gold and bronze assessors.

In another work, [Kinney et al. \(2008\)](#) studied the performance of domain experts and what they called “generalists” evaluators, who essentially are non-experts on a given search topic. In their work, they targeted the computer programming and biomedical domains only, in the experiments they conducted. In comparison of the quality of ratings of the search results, they found domain experts to be more accurate. They also found that assessors who lack domain expertise, tend to judge documents based on the occurrence

of keywords, while domain experts judge the quality of the document and its relationship to the search topic. They determined that the ratings of the search results reveal that the domain experts understand the intention of the query to a greater extent than the generalists. Indeed, generalists demonstrate a shallow level of understanding the meaning of a query, as well as its intention. They recommend that generalists to be provided with written intent statements if we need to improve the quality of the ratings made by them. These intent statements will give clear and concise instructions about the query intent.

[Carterette and Soboroff \(2010\)](#) studied the impact of assessors errors on the methods used in Million Query track test collections. In such collections, the number of queries are high, while the number of relevance judgments for each query is low. Their results are based on eight models of possible assessor error, and they argue that these errors might have a great impact on systems ranking, though the averages tend to give more stable results.

2.3 Factors that Influence Relevance judgments

There are plenty of works in the literature about this issue ([Barry, 1994](#); [Schamber, 1994](#); [Park, 1993](#); [Saracevic, 2007](#)). According to [Chu \(2011\)](#), those factors can be grouped into several categories. These categories are: (1) studies that cover all factors, nothing in specific; (2) studies that cover certain factors, such as the order of presentation, knowledge, interest and experience in the process of relevance judgment; (3) studies that cover the impact of associating different relevance factors with different phases of a task; and (4) studies that focus on the specific subject of research, and/or specific document type.

Cool et al. (1993) conducted two users studies to understand what types of factors impact assessors relevance judgments rather than the topicality factor. Their two user studies targeted two different groups; in the first study, they recruited university students where they gave them search tasks and asked them to fill out questionnaire when selecting each document that they think it is relevant. The aim of this questionnaire was to obtain all the factors that users considered when assessing the relevance of documents. The design of the second user study was different from the first one since they targeted scholars in humanities where these scholars were required to search for relevant documents for their information need. Intensive interviews took place with those scholars in order to understand what impacted their relevance judgments. What the authors have found in these user studies was that the concept of topical relevance is not the only factor that have an influence on users relevance judgments. There are several number of influencing factors that could contribute to the decisions when assessing documents. They summarize those under what they related to facets of the judgment process. There facets are six and they are: (1) Topic, (2) Content/Information, (3) Format, (4) Presentation, (5) Values, (6) Oneself.

In another study, Schamber (1994) gathered 80 relevance factors and grouped them in a table under six categories. Those categories were: judges, requests, documents, information system, judgments condition, and choice of scale. Similarly, and in another attempt to identify influential factors, Barry (1994) created seven classes where there were a total of 23 factors.

One example of a study that focuses on only one factor is the study that was conducted by Eisenberg and Barry (1988). In their study, they focused on the impact of the order of

presentation on the assessors' relevance judgments. Assessors were assigned two different groups of documents based on their relevance degree. The former group was sorted from high to low, whereas the latter group was sorted from low to high. They argue that the order of presentation impacts assessors' relevance judgments. In a high to low approach, they noticed that the highly relevant documents at the top of the rank were consistently underestimated. However, in a low to high approach, documents at the low and medium range were overrated. [Ruthven et al. \(2007\)](#) conducted a study to investigate the impact of three factors: (1) the assessor's level of knowledge of a search topic, (2) his/her interest in the search topic, and (3) his/her confidence level when assessing the relevancy of documents. Based on their results, all of these factors were found to have an impact on assessors' relevance judgments. The part of Ruthven's study that is most relevant to our work in this thesis is the level of assessors' confidence when assessing the relevancy of documents. In the study, they asked assessors direct questions about their level of confidence when assessing documents and they allowed them to choose just one answer from: (1) Confident, (2) Depends on the documents retrieved, and (3) not confident. Their goal was to determine if the confidence level impacts the accuracy of assessors' relevance judgments. What they found is that the higher the level of confidence one has when assessing documents, the more relevant decisions one makes.

In a similar study, [Taylor et al. \(2007\)](#) studied the process of making relevance judgments that an assessor goes through when making relevance judgments. The study's results confirmed what others studies ([Spink et al., 1998](#); [Tang and Solomon, 1998](#); [Vakkari, 2000](#); [Vakkari and Hakala, 2000](#); [Wang and White, 1999](#)) found in regard to the variety in assessors' relevance judgments during the information seeking process. They also emphasized

that the topical relevance is not the only type of relevance that assessors apply during the information seeking process.

Xu and Chen (2006) followed a cognitive approach and designed a study to investigate the importance of their five-factor model of relevance. Their aim was to concentrate on other factors that assessors use in assessing documents rather than the topicality factor. Their five-factor relevance model consisted of: (1) topicality, (2) novelty, (3) reliability, (4) understandability and (5) scope. Their results found that topicality and novelty are the most critical factors, whereas understandability and reliability has some importance as well. However, the scope factor was found to have no importance on assessors' relevance judgments.

Finally, a comprehensive study conducted by Bales and Wang (2005), identified and grouped relevance factors that influenced assessors' relevance judgments. They covered 16 empirical studies which reported in 19 journal papers. In their study, they identified 230 factors that affect assessors' relevance judgments and grouped them into four dimensions: (1) situations and context, (2) user criteria, (3) document information elements and (4) value judgments. They concluded their work by stating that all of the empirical studies they covered in the study provide proof that relevance is a multidimensional concept.

In Scholer et al. (2011), they investigate deeply the inconsistency in relevance judgments made by assessors in TREC. They used several test collections and compared assessors relevance judgments when judging duplicates documents. They found that inconsistency in relevance judgments were affected by the proximity of duplicates from each other; in other words, if the distance (the number of documents between duplicates) is shorter,

assessors tend to be more consistent in the judgments they make. They also discuss what they call “judgment inertia”, where they claim that it could have an impact on the consistency of relevance judgments; since assessors tend to assign to the next document the same relevance value that he/she assigned to the preceding document.

Villa and Halvey (2013) conducted a study to investigate if the length of documents and the degree of relevance of documents have an impact on the correctness of assessors’ relevance judgments and the effort that they need to make those judgments. They found that assessors tend to pay much greater attention as the document increases in size. Moreover, documents that are either highly relevant or not relevant needed less effort to be assessed. The level of accuracy, while not impacted by the length of judged documents, was impacted by the degree of relevance of a document.

In the same context, Scholer et al. (2013) emphasize the importance of varying the level of relevance of documents for better assessors’ relevance threshold. They claim that the first documents that an assessor encounters during the relevance judgments process will affect his or her relevance threshold. Scholer et al. (2011) points out that the same assessor might produce different relevance assessments for the same document at different points of time during the relevance judgments process (15% to 19% when a binary-relevance scale is used and 19% to 24% when a three-level scale is used).

In an advanced step, Chandar et al. (2013) designed a study to predict assessor disagreement. In this study, two of their hypotheses were found to be true: (1) “longer documents will provoke more disagreement”, and (2) “less coherent documents will provoke more disagreement”; however, and surprisingly, a third hypothesis, that “documents

that are more difficult to read will provoke higher levels of assessors disagreement” was found to be untrue. In fact, they found the opposite to be true: “documents that are easier to read are the ones that provoke more assessors disagreement”.

2.4 Crowdsourced Relevance Assessments

With rapid development in technology and the dramatic increase of the web contents, the size of test collections that were built during the 1960s and 1970s was not reflective of the modern demand of IR systems. Therefore, there has been effort to create large test collections such as ClueWeb12 Dataset that contain hundreds of millions of documents, and the hope is that these types of test collections will help in evaluating IR systems that are equal to or similar to current market standards. Therefore, researchers have worked on different techniques to deal with the creation of these test collections in order to use them in evaluate IR systems correctly.

In a typical NIST setup (or scenario), the assessor (primary assessor) who creates the search topic, usually judges the documents in the pool and is required to decide whether or not they are relevant to the search topic. Therefore, it is the role of the assessor to check whether each document in the pool is relevant. The concept of crowdsourcing (Howe, 2008), has opened the door widely for a new technique to be used in IR to create relevance judgments. Now, with the help of the crowd, relevance assessments can be created not only faster but also less expensively than recruiting costly primary assessors (Alonso and Mizzaro, 2009; Snow et al., 2008; Alonso and Mizzaro, 2012; Alonso, 2011; Alonso et al., 2008).

For example, [Clough et al. \(2013\)](#) studied whether relevance judgments made by crowdsourced workers would be reliable and comparable to those made by domain experts. The authors found that crowdsourced workers were able to produce relevance judgments that were helpful in ranking two search engines which they used in the experiment. However, they reported that domain experts were able to observe subtle differences between various levels of relevant results returned by the system.

Alonso and Mizzaro ([Alonso and Mizzaro, 2012](#)) argue in their paper that relevance judgments made by crowdsourced workers (i.e. secondary assessors) are comparable in reliability to those made by primary assessors. In order to obtain reliable relevance judgments, they emphasize on the careful design of the experiment, its execution, and finally quality control. They claim that using crowdsourced workers is less expensive than recruiting primary assessors, who are not only limited in their numbers but are financially more costly than crowdsourced workers. The quality of relevance judgments made by crowdsourced workers is the main issue that researchers in IR have worked diligently to address and enhance ([Alonso and Baeza-Yates, 2011](#)). As pointed out in ([Alonso and Baeza-Yates, 2011](#)), the issue of the quality of the produced relevance judgments is challenging, and such a challenge could be encountered and overcome by the cooperation between groups from different areas such as HCI aspect or game design, incentive engineering and of course IR.

[Eickhoff et al. \(2012\)](#) claim that the quality of relevance judgments could be improved by applying concept of gaming in crowdsourcing tasks. Their suggested technique is used to attract more reliable assessors and distract possible cheaters and others who produce low quality work. They described the type of secondary assessors in crowdsourcing tasks as: money-driven assessors and entertainment-driven assessors. The entire goal of their work

is to attract the entertainment-driven assessors since they are sources of more reliable work in crowdsourcing tasks. They compare their suggested design to typical design in crowdsourcing to collect relevance judgments, where they found that their proposed design improves the quality of relevance judgments.

Our work in this thesis adds to the existing research knowledge about the behavior of secondary assessors and the goal is to deepen our understanding of their behavior and how they work on the assigned search tasks and how they make relevance judgments as we will see in the remaining chapters of this thesis.

2.5 Think-Aloud Method

As we stated earlier in Chapter 1, the aim of this work is to better understand secondary assessors' judging behavior and what causes them to make different relevance judgments. This would be better achieved if assessors' thoughts were verbalized as they make the judgments. Raya Fidel describes the importance of the qualitative methods in (Fidel, 1993). She states

The qualitative approach offers the best methods for exploring human behavior. It is exploratory because it is the best for investigating complex phenomena when very little is known about them, and it is not usually employed for studying retrieval systems from purely mechanical or computational perspectives.

Kelly (2009) also points out the benefits of qualitative methods, and these types of methods are recommended to be used when we do not have sufficient information on a

particular phenomenon. Therefore, we designed the first part of the PhD research work on this thesis to be exploratory, and our aim is to gain a better understanding of assessors' relevance disagreements.

From consideration of a number of qualitative methods, we chose the think-aloud method to accomplish our goal. In a typical think-aloud study, users' (in our case we call them assessors) voices are recorded as they perform the required search tasks. The following paragraphs discuss the reasons behind choosing the think-aloud approach in greater details.

2.5.1 Why To Use it?

In a typical IR experiment, assessors are engaged in a relevance judgments process. This process (relevance judging) can be characterized as a cognitive process that has a goal and requires effort and attention. The think aloud is one of the techniques that is used to help in studying cognitive processes. This method is widely used in many disciplines such psychology, education, sociology, information science, information retrieval etc.

Think aloud is a useful data collecting method that helps in gaining more insight into how users think, behave, and conduct their tasks (Ericsson and Simon, 1980, 1993). In think-aloud studies, users are asked to speak aloud while they engage in assigned tasks. The objective is to listen to what goes through their minds during the assigned search tasks, and produce what is called "Verbal Protocols". However, some would argue that this type of method would potentially intervene with the assessors ability to perform the search tasks, and as a result the quality of the results would be jeopardized. Ericsson and

Simon (1993) pointed out this concern, and indicated that there is no interference between thinking aloud and the users' performance. Having said this however, the type of the task that we want users to perform and its level of difficulty will play a major role in determining whether we should pick think-aloud as a tool for generating verbal protocols, since verbal protocols might not be effective with cognitively overwhelming tasks (Charters, 2003).

There are several methods to gather verbal protocols: Retrospection, Introspection, Questioning and Prompting, Dialogue Observation, and Think-Aloud (Someren et al., 1994). Each technique has its own advantages and disadvantages. For example, retrospection is the technique that is used after a user completes the required task. He/she will be asked to consider the process which he/she followed. Though this technique is effective for giving interpretation and explaining what a user did, it might not reflect what occurred when the user was performing the task, since there is a distance in time between the task he/she performed and the retrospection process. The Introspection is another method for gathering verbal protocols. The user is asked to work on the assigned task and then at intermediate points chosen by him/her, he/she stops and reports what he/she is doing. However, this method also suffers from the same issue that the retrospection suffers from and that is they are subject to memory errors and misinterpretations.

The questioning and prompting method is subject to experiments' interruption. The experimenter interrupts the user at different time points and asks him/her about what goes in his/her mind and what he/she is doing and other questions that can give the experimenter more information about the behavior of the user and the cognitive process. The last method is the dialogue observation. The experimenter in this method asks the user questions while he/she works on the assigned tasks and record the dialogue in order

to analyze it later. This method suits the best the tasks that requires dialogue in order to work on them. However, not all tasks suits this method.

The method that we chose for our first user study is the think-aloud method. As stated earlier, in this method, users are instructed to talk aloud while working on the assigned task. The experimenter or researcher cannot intervene while the user is talking. The only role for the researcher is to encourage the user to talk if there is a lengthy period of silence. In this way, we can ensure the thought process is not interrupted and the user is focused in generating more useful information. In think-aloud, the user does not interpret what he/she is doing. He/she just says what thought goes into his/her mind while performing the tasks. One of the most advantages of the think-aloud method that it minimizes the occurrences of memory errors since it requires concurrent verbal reports; unlike the other methods that wait until the end of the assigned task and then ask about what happened or has been done.

As stated by [Ransdell \(1995\)](#), the information that is obtained by the think-aloud method can be used to check the validity of a hypotheses and models of behavior. Although it is capable of accomplishing this, Ransdell believes that verbal protocols are a better source of generating hypotheses. She states, “Protocols are also important because they yield rich data and thus promote hypothesis generation, but are not as powerful for hypothesis testing purposes”.

In information-seeking and IR research, think-aloud has been recognized as a tool for data collection as well ([Kelly, 2009](#)). According to [Branch \(2000\)](#), none of the published work in information-seeking research ([Yang, 1997](#); [Hughes et al., 1998](#); [XIE and Cool, 1998](#);

Hirsh, 1999) reported any problems with use of the think-aloud method as a data collection tool.

Chapter 3

Data Set

The results reported in this thesis are based on a set of search topics and a set of documents taken from the AQUAINT collection. This collection consists of 1,033,461 documents taken from the New York Times, the Associated Press, and the Xinhua News Agency newswire. The AQUAINT collection was used in the Robust Track in TREC 2005. Robust Track was designed to enhance retrieval technology consistency through work on topics that produced poor results. Relevance judgments which are produced by NIST assessors are on a ternary scale: “not relevant”, “relevant”, and “highly relevant”. Throughout this thesis, 0 refers to “not relevant”, 1 to “relevant” and 2 to “highly relevant”. In our user studies, we consider both “relevant” and “highly relevant” to be relevant.

3.1 Topic Selection

In total, five search topics were used in the two studies. In both of them, we used four for the main search tasks in the Task Part, and one for the tutorial in the Tutorial Part. The topics that we used, and their titles, are shown in Table 3.1. As it is followed in TREC, every search topic has a title, a description, and a narrative (Voorhees et al., 2005). In our studies, we combined the description and the narrative and call this the “Search Topic’s Description” as will be shown in the coming sections.

Number	Phase	Topic Title
427	Tutorial	UV Damage, Eyes
310	Task	Radio Waves and Brain Cancer
336	Task	Black Bear Attacks
383	Task	Mental Illness Drugs
436	Task	Railway Accidents

Table 3.1: TREC topics used in the studies

3.2 Topic 336: Black Bear Attacks

This topic is about Black Bear Attacks. Figure 3.1 presents the description for Topic 336 that assessors are required to read in order to assess relevance of documents. As can be noticed in Figure 3.1, this description starts with giving only brief information about the information need. Then, more sentences are given to explain and give details about what an assessor should consider when looking for relevant material in the documents that he/she is going to judge.

A relevant document would discuss the frequency of vicious black bear attacks worldwide and the possible causes for this savage behavior. It has been reported that food or cosmetics sometimes attract hungry black bears, causing them to viciously attack humans. Relevant documents would include the aforementioned causes as well as speculation preferably from the scientific community as to other possible causes of vicious attacks by black bears. A relevant document would also detail steps taken or new methods devised by wildlife officials to control and/or modify the savageness of the black bear.

Figure 3.1: Topic 336

3.3 Topic 310: Radio Waves and Brain Cancer

This topic is about Radio Waves and Brain Cancer as illustrated in Figure 3.2. It requires a good level of understanding of some electrical waves terminologies. Also, a good level in reading comprehension is recommended since an assessor has to find the link between radio waves and brain cancer as required by the description of the search topic. When he/she judges documents for this search topic, he/she might go over a number of results from experiments, statistical studies, articles and so on. Therefore, his/her role is to decide if these results fit or satisfy the information need in the description or not.

Relevant documents will provide evidence that radio waves from radio towers or car phones affect brain cancer occurrence. Persons living near radio towers and more recently persons using car phones have been diagnosed with brain cancer. The argument rages regarding the direct association of one with the other. The incidence of cancer among the groups cited is considered, by some, to be higher than that found in the normal population. A relevant document includes any experiment with animals, statistical study, articles, news items which report on the incidence of brain cancer being higher/lower/same as those persons who live near a radio tower and those using car phones as compared to those in the general population.

Figure 3.2: Topic 310

3.4 Topic 383: Mental Illness Drugs

Topic 383 is about finding drugs used in treating mental illness. As shown in Figure 3.3, this topic might seem easy from the first glance since the description is short and concise; however, an assessor would come across different drug names and some of these names might be a source of confusion as we will discuss in Chapters 5 and 6 of this thesis. Also, some assessors would be in doubt about some illnesses if they are mental illnesses or not.

Relevant documents will identify drugs used in the treatment of mental illness. In particular, a relevant document will include the name of a specific or generic type of drug. Generalities are not relevant.

Figure 3.3: Topic 383

3.5 Topic 436: Railway Accidents

Topic 436, as shown in Figure 3.4, should be an easy topic since it is a general topic and does not require familiarity in a specific domain. Its description is concise and mentions clearly what should be considered when looking for relevant material in the documents. Relevant information for this search topic should be on the surface of the documents.

A relevant document will provide data on railway accidents of any sort (i.e., locomotive, trolley, streetcar) where either the railroad system or the vehicle or pedestrian involved caused the accident. Documents that discuss railroading in general, new rail lines, new technology for safety, and safety and accident prevention are not relevant, unless an actual accident is described.

Figure 3.4: Topic 436

3.6 Document Selection

When choosing documents for our work in this thesis, we relied on the data collected in previous studies (Jethani, 2011; Smucker and Jethani, 2011b, 2010). In these studies, documents are judged by many participants; therefore, we have a good level of knowledge about their consensus levels. We refer to a consensus level here as the general agreement upon the relevancy of a document to a given search topic. For instance, if a document was judged relevant or not relevant by 90% of assessors, then the majority of assessors agreed on its relevancy, either relevant or not relevant; therefore, it is a high consensus document. In contrast, if just 50% of assessors judged it as relevant or not relevant, then we do not have a general agreement on its relevancy, and, as a result, it is a low consensus document. The criteria we applied in selecting high and low consensus documents is as follows:

The probability of relevance

The probability of relevance is equal to the fraction of participants that judged a document to be relevant. To make this point clearer, let us take the following scenario. Suppose document C was judged by 10 assessors. Nine of them judged it as not relevant and just one of them said the document is relevant. Then, the probability of relevance for document C is not exactly 0 but it is 0.1. However, if all the 10 assessors judged it as not relevant then its probability of relevance would be 0. We divided the documents into two levels of consensus: high and low. We selected 9 documents for each search topic in the study: 6 documents at the high consensus level and 3 documents at the low level of consensus. Of the 6 documents at the high consensus level, 3 are relevant and 3 are not relevant per NIST.

Table 3.2 shows the number of documents at each consensus level and their corresponding NIST qrel scores.

The probability of relevance for each group is as follow:

- For a relevant document at the high consensus level, its probability of relevance has to be ≥ 0.8 .
- For a non-relevant document at the high consensus level, its probability of relevance has to be ≤ 0.2 .
- The probability of relevance of a low consensus level document has to be ≥ 0.4 and ≤ 0.6 .

Number of Assessors

When we selected the documents, we considered documents which were judged by at least six participants or more in (Jethani, 2011; Smucker and Jethani, 2011b, 2010). The number of participants who judge our documents vary as can be seen in Table 3.3 .

Reciprocal Rank Fusion (RRF)

To add more strength to our choice of documents, we also considered choosing the documents that are more likely to be returned by actual search systems. We are interested in user behavior with top ranked results. Therefore, we used the parameters given by Cormack et al. (2009) to calculate the reciprocal rank fusion (RRF) for each document that we chose. This was obtained by using all the runs submitted to the TREC 2005 Robust track. When a document is retrieved near the top of a ranked list, this means its RRF is high,

		NIST qrel		
		NR=0	R=1	HR=2
Topic 310: Radio Waves and Brain Cancer				
High Consensus	Non-Relevant	3		
High Consensus	Relevant		2	1
Low Consensus	Non-Rel/Rel		3	
Topic 336: Back Bear Attacks				
High Consensus	Non-Relevant	3		
High Consensus	Relevant		1	2
Low Consensus	Non-Rel/Rel	1	2	
Topic 383: Mental Illness Drugs				
High Consensus	Non-Relevant	3		
High Consensus	Relevant			3
Low Consensus	Non-Rel/Rel	2		1
Topic 436: Railway Accidents				
High Consensus	Non-Relevant	3		
High Consensus	Relevant		1	2
Low Consensus	Non-Rel/Rel	2	1	
All 4 Topics				
High Consensus	Non-Relevant	12		
High Consensus	Relevant		4	8
Low Consensus	Non-Rel/Rel	5	6	1
Total:		17	10	9

Table 3.2: Distribution of Documents in Low and High Consensus Levels and their NIST qrel (relevance judgments are called qrels (Voorhees et al., 2005)) Scores. NIST qrel Scores 0, 1, and 2 stand for “Not-Relevant”, “Relevant” and “Highly Relevant” respectively.

while it will be low if it is not among the top retrieved documents. From this perspective, we chose documents that have higher RRF scores. Again, this means the greater the RRF score, the better the choice.

	ProbRel	#Assessors
Topic 336:		
APW20000323.0200	0.5	14
APW20000622.0185	0.5	16
APW20000703.0186	0.533	30
APW19990809.0179	0.875	16
NYT20000706.0242	0.864	44
NYT20000602.0371	0.846	26
NYT19990927.0436	0.2	9
NYT20000718.0206	0.2	8
XIE19980113.0247	0.2	9
Topic 310:		
NYT20000224.0139	0.545	11
XIE19970506.0203	0.571	14
XIE20000628.0163	0.545	11
APW20000608.0153	0.853	34
NYT19991003.0452	0.909	22
NYT19991025.0333	0.93	43
NYT19990405.0532	0.05	20
NYT19990805.0436	0.071	28
NYT19990907.0397	0	27
Topic 383:		
NYT19990121.0380	0.538	13
NYT19991214.0159	0.519	27
NYT19991206.0109	0.6	35
NYT20000925.0105	0.808	26
NYT20000319.0216	0.882	17
APW20000307.0001	0.833	6
NYT19980727.0419	0.143	28
NYT19990711.0089	0.04	25
NYT20000717.0206	0.053	19
Topic 436:		
APW19990914.0022	0.474	19
XIE19980303.0229	0.5	28
NYT19991206.0299	0.6	10
XIE19991207.0246	0.967	30
XIE19990802.0027	1	36
XIE19981020.0034	0.943	35
XIE19960718.0212	0	29
XIE19970503.0122	0	24
XIE20000724.0250	0.042	24

Table 3.3: Probability of Relevance and number of assessors for each documents used in our studies.

Chapter 4

Think-Aloud Study

4.1 Think-Aloud Method

The use of qualitative research methods, such as think-aloud, is established in Information Retrieval (Kelly, 2009). It is considered a valuable source of information for researchers. It helps in, not only, supporting the results that researchers find with other quantitative research methods, but it also reveals important data about the relevance judgments process and users' behavior. The data obtained by think-aloud (or other qualitative research methods) will be used alongside other data obtained by quantitative research methods, in order to foster the research in the area of evaluation in IR.

Several types of data can be collected by using qualitative methods such as video, audio, text and other formats. These types of data can provide far more explanations about users' behavior that we are not able to obtain with quantitative research methods.

In think-aloud studies, users are asked to speak aloud their thoughts while they work on assigned tasks. This will give instant access to what is going on in their heads at that moment. The research facilitator does not intervene while the assessors work on the assigned tasks. The research facilitator's only task is to encourage them to continue speaking if there is a lengthy period of silence.

4.2 Study Design

We created an experiment that consisted of two parts: The Tutorial, which helped train the participants in how to judge the relevancy of documents; and The Task, in which we collected real data.

4.2.1 Study Protocol

As mentioned above, the study is divided into two parts. The first part of the study was devoted to train the assessors about the required tasks, and the second was to perform the assigned search tasks. Our study was a lab-experiment. It was held in a private room at the university. The room provides a good level of privacy and quietness for conducting the study.

Upon their arrival at the laboratory, participants were handed an information letter which explained to them everything that they needed to know about the study. Subsequently, and if they agreed to participate, we asked them to sign a consent form. Partic-

ipants then logged into the system and went through a number of pages which provided them with information about the study and instructions on how to perform it.

A demographic collection page followed the instructions and we collected basic experience data from participants. This included their age, academic discipline, search habits, and whether they had received any training or education in information retrieval. After that, participants completed the two phases of the study.

We recorded not only what was shown to participants on the screen and their interaction with it, but also their voices. The researcher and an outside vendor transcribed the collected data, the audio and video. All the reported results in this thesis are based on the transcribed data. The following is detailed information about each part of the study.

Tutorial Part

As it should happen in any study and to produce more reliable data for a think-aloud study (Branch, 2000), participants must have adequate training. Therefore, our main goal in the Tutorial Part was to give them the required training so that they felt confident performing the Task Part of the study. Moreover, the Tutorial Part helped them to practice how to verbalize their thoughts. Four documents were given to each participant in this part, and participants were required to judge the documents relevance to a given search topic. Two of the documents were relevant to the given search topic, and the other two were not relevant. The task of judging documents for a search topic is referred to as a search task, and that is different from a task for judging a single document. This distinction should be clear.

During the Tutorial Part, participants were told if their answers were correct or not and, in few cases, the researcher gave some explanation or gave more detail if there was a need. The participants were informed that if any elements were unclear during the tutorial, they should seek clarification, as during the Task Part, they would work solo, with no intervention from the researcher.

Task Part

In the main part of the study, the participants were left to work on their own. At this part, the only role of the researcher was to encourage the participant to continue speaking when there was a lengthy period of silence. The researcher did so since this is the primary characteristic of the think-aloud protocol, where participants speak while performing the required tasks, thereby allowing for data collection. Each participant in this part performed 4 search tasks. In each search task, he/she judged 9 documents as either relevant or not relevant. As a result, 36 relevance judgments were made by each participant.

4.2.2 User Interface

We designed our user interface to accept only binary judgments, where assessors must judge a document as either relevant or not relevant. We did so since the collection of binary relevance has been the most standard form of relevance judging.

The designed interface showed participants one full document at a time. To ensure that participants remembered the search topic and its description during the task, the title of the search topic and its description remained available on the interface. Figure [4.1](#) shows

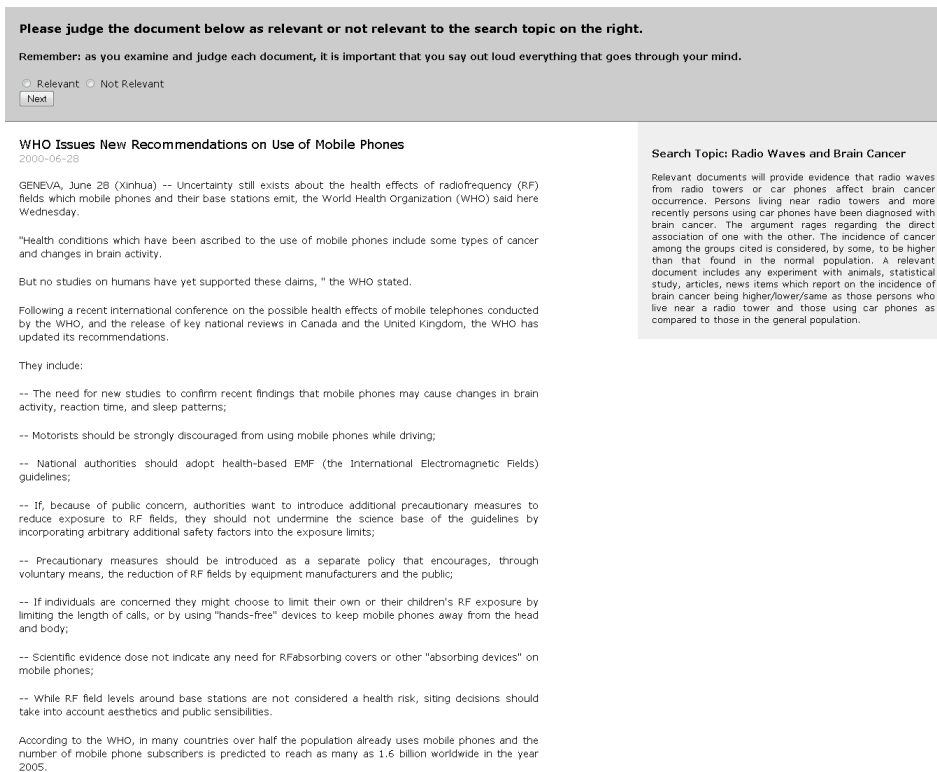


Figure 4.1: The user interface for judging a single document

the interface. The participant cannot rejudge the document once he/she goes to the next one. The judgment that he/she made is final.

4.2.3 Measuring Judging Behavior

Judging a document as relevant or not relevant borrows from signal detection theory, where Yes and No are the only choices (Abdi, 2007). In this theory, the number of hits, which represents the times when the assessor was right in saying Yes, and the number of false

alarms (FA), where the assessor was wrong in saying Yes, are what we need to infer the performance and the behavior of assessors.

In information retrieval, we call the hits True Positives and we call false alarms False Positives. Table 4.1 illustrates all types of response that we obtain while using a binary relevance judgment scale.

From the type of responses in Table 4.1, we can compute the true positive rate (TPR), which represents the fraction of relevant documents judged as relevant, and the false positive rate (FPR), which represents the fraction of non-relevant documents judged as relevant. The TPR and FPR are widely used in information retrieval as measurements of users’ performance and behavior; therefore, we used them for this purpose. The true positive rate is measured as:

$$TPR = \frac{|TP|}{|TP| + |FN|} \tag{4.1}$$

and the false positive rate is measured as:

$$FPR = \frac{|FP|}{|FP| + |TN|} \tag{4.2}$$

where TP, FP, TN, and FN are from Table 4.1.

Participant	Relevant (Pos.)	Non-Relevant (Neg.)
Relevant	TP = True Pos.	FP = False Pos.
Non-Relevant	FN = False Neg.	TN = True Neg.

Table 4.1: Confusion Matrix “Pos.” and “Neg.” stand for “Positive” and “Negative” respectively.

In our study, we compare the relevance judgments not only against the consensus levels, high and low which inferred from the data collected in previous studies (Jethani, 2011; Smucker and Jethani, 2011b, 2010), but also against NIST qrel scores as well.

We are also interested in calculating the d' and the criterion c due to their importance in revealing significant data about users' behavior in assessing the relevance of documents. The d' measures an assessor's ability to discriminate between relevant and non-relevant documents. By knowing the TPR and FPR, we can compute the d' as the following:

$$d' = z(TPR) - z(FPR) \quad (4.3)$$

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

From the TPR and the FPR, we can also calculate the criterion c , which tells us about the strategy that an assessor follows when making relevance judgments. A positive c indicates that an assessor has a conservative behavior, while a negative c indicates a liberal behavior. Criterion c can be computed as the following:

$$c = \frac{1}{2}(z(TPR) - z(FPR)) \quad (4.4)$$

4.2.4 Participants

We started the recruitment process immediately after getting the approval from the Office of Research Ethics (ORE) at the university. We used email to recruit participants. We sent an email to the Grad-Mailing List to which all graduate students in the university are

subscribed. The email was selected in the recruitment process since it helps to reach all students wherever they are. When recruiting via email, all students need to do is reply to the email and state their intention to participate.

We paid each participant \$20 dollars for the whole study. We informed them about the policy of the study were they could stop at any point and withdraw his/her consent. In this case, they would be paid on a pro-rated basis at a rate of \$5 per search topic completed.

In total, 14 graduate students participated in the study. We used the first two participants to test the settings of our study. Therefore, the actual number of participants is 12. However, from the remaining 12 participants, we excluded the data of 3 participants since there was unintended intervention from the researcher with these 3 participants. We provide more detail about this exclusion in the *Cleaning of Data* section. All reported results are only for the remaining 9 participants.

The participants were from different academic backgrounds. Six were science, technology, engineering, or mathematics students; two were arts students; three were environmental studies students; and one was a health studies student. The average age was 25 years, the minimum was 21, and the maximum was 35. Since we were dealing with graduate students who speak different languages, we required that all participants be native English speakers. Graduate students who are not native English speakers may be very good in English in their own fields but they may struggle to understand some English documents from outside their disciplines. Moreover, some students claim that they are fluent in English but when they come to the participate, we find that they are not. Rather than confound the

results with English comprehension issues, we excluded non-native English speaker. For this reason, we mentioned this condition in the recruitment email.

The majority of the participants stated that they use search engines several times a day and they are experts in finding information on the Internet. None of the them received training in information retrieval. However, only two of them indicated that they had received introductory lessons in accessing library resources and article databases.

4.3 Cleaning of Data

During the engagement of the participants in the relevance judgment process, there were three times where they were confused about one of the search topics and asked the researcher directly about it. The researcher answered them by accident since he was not expecting questions from participants during the relevance judgment process.

However, in order to obtain clean results that are not biased or impacted by the research facilitator, we identified all the places where the interventions occurred. Then, we excluded all the data that belonged to the participants who received intervention.

It is worth mentioning that all the interventions occurred with search topic 310, which is about Radio Waves and Brain Cancer. In these interventions, the participants were asking about car phones. It appeared that they knew nothing about them and never heard of them. We notice that these participants are under the age of 25 and to them this term is not known.

After cleaning all data in the study, we have 324 relevance judgments. All our analysis and results are based on these relevance judgments.

4.4 Certainty in Relevance judgments

This study is an effort to gain a better understanding of secondary assessors' relevance judgment behavior. It explores the causes of making different relevance judgments, and determines if assessors make guesses while judging documents. The think-aloud method is the tool that we used to address these issues. We were able to listen directly to what assessors said, while they were engaged in performing search tasks. The following sections discuss the results of think-aloud study in more detail.

When analyzing the transcribed data, which is produced by participants in our think-aloud study, participants express different levels of certainty in their relevance judgments. This is shown by the types of expressions and phrases that we noted in the transcripts. Figure 4.2 illustrates all the expressions that we found. At the two ends of the spectrum in this figure, there are two entirely different types of certainty. The most uncertain relevance judgments start from the far left of Figure 4.2, and, toward the right of the same figure, the level of certainty increases dramatically. Users sometimes make highly certain judgments while, in other cases, the certainty in their relevance judgments is completely absent. Between these two extremes, there are other levels of certainty that we are not able to identify clearly by using common quantitative research methods; however, the think-aloud method helped us to clearly achieve this.

The forthcoming subsections discuss the levels of certainty that we have found in the collected data.

4.4.1 Low Certainty Relevance judgments

Low certainty relevance judgments represent the fraction of judgments produced when participants are confused and do not know what to do. Judgments here are nothing more than guesses. Participants at this level of certainty believe that the binary relevance grading scale, which the interface offers, does not represent their judgments correctly. However, they have to make a decision; therefore, they just make a random choice.

In the following example, Participant 2 did not know what Attention Deficit Disorder (ADD) was. This lack of knowledge became clear as the participant worked on topic 383 (Mental Illness Drugs). Topic 383 requires the relevance assessor to find the names of drugs used to treat mental illness. From the transcribed data, which we obtained via the think-aloud method, we can note that the lack of knowledge about this term (ADD) led the participant to be confused and, as a result, produce a low level certain judgment by saying *I'm not sure*. Here is the complete think-aloud transcript for this example with comments in parentheses:

Are we overmedicating our kids? (Participant reads the title of the document.)
How is this one related to mental illness? Was impulsive, wouldn't sit down.
(Part of document read out loud.) Attention deficit, ADD, Ritalin. (Has found
the name of a drug.) Is ADD a mental illness? Hmm, I don't think so. (Answer
own question.) I'm not sure. *Participant 2*.

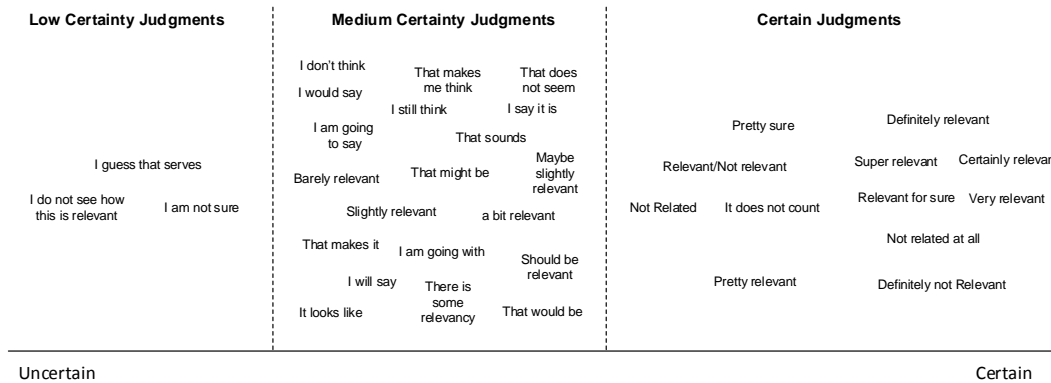


Figure 4.2: Expressions used in conjunction with making a relevance judgment. The expressions vary from those indicating little to no certainty on the left to complete certainty on the right.

4.4.2 Medium Certainty Relevance judgments

There is more certainty in participants' relevance judgments at the medium level than at the low level. However, the main characteristic of this level is a lack of total certainty. In the following example, participant 11 is trying to decide if a document is relevant to topic 436 (Railway Accidents). Topic 436 requires the relevance assessor to find documents that provide data on railway accidents of any sort. The description of topic 436 also provides more details about what one should consider or not when judging documents. The following is an excerpt from the transcript:

Germany to replace rail chief. It doesn't sound relevant. So far. But it is talking about the person. More talking about the person. No mention of an accident. Oh, what ... they do talk about one crash in 1998. Meant that killed 101 people. And that, talking about how he handled it. It provides some

information about the accident but not that much. Ah, it doesn't really focus on the, the accident. And it doesn't provide any really damage data to see there is an accident. That killed 101 people. But I don't know if that makes it relevant enough. Uhm, I think ... says, it's a ... it says how many people were killed that counts as data and I think that makes it relevant. *Participant 11.*

It is obvious in the above example that it was difficult for the participant to fit what he/she was reading in the document into the criteria mentioned in the description of the search topic. We can note that the participant was going back and forth between the two mandatory options, relevant or not-relevant. This is caused by a lack of ability to find a good criteria fit. Therefore, producing a medium level relevance judgment was the choice of the participant; however, like the judgments at the low level certainty, judgments at this level were transformed by force into a binary relevance judgments due to the two options offered by the interface.

4.4.3 Certain Relevance judgments

At the other end of the spectrum in Figure 4.2, participants had many ways to express that they were completely, or near completely, certain in their judgments.

For example, participant 8 is definite about his/her judgment. In this example, mentioning the names of the drugs in the document's title made the participant sure about the decision that he/she made.

Okay, So the White House seeks to curb. Alright, So this mentions drugs that are used for mental illness right in the title. Definitely relevant. *Participant 8.*

In the following example, Participant 11 is also certain about the relevance judgment that he/she made. It appears from the expressions that he/she was using *it does not sound that it could be relevant* that he/she is not definitely certain (highly certain) about his/her decision, but is at least certain.

It doesn't sound that it could be relevant because it doesn't talk about an attack in the title. It just talks about bigfoot and no attacks. It is not relevant.

Participant 11

4.4.4 Frequency of Certainty Levels

Table 4.2 shows each certainty levels we have found in the participants' answers and their frequency in our data. As noted in the table, most of the relevance judgments are made with certain to medium certainty. As the low certainty judgments represent a small fraction — 25 of the 324 total relevance judgments (just 7.71%) — this leads us to infer that assessors will occasionally be forced to guess the relevance judgments if the interface limits decisions to either relevant or not-relevant.

Certainty Level	Total
Certain	196
Medium Certain	100
Low Certain	25
Unknown	3
Total	324

Table 4.2: Certainty levels and their frequency. Unknown in this table refers to assessors’ relevance judgments that we were not able to categorize under one of our three main certainty levels.

4.5 Making Incorrect Relevance judgments

We analyzed participants’ incorrect relevance judgments based on high and low consensus levels and NIST qrel scores as well. Table 3.2 in Chapter 3 three shows the number of documents at each level of consensus and their corresponding NIST qrel scores. Please refer to that chapter if there is a need for more detail about the document selection process. There was a feeling before analyzing the data that the reasons for making incorrect relevance judgments might differ based on the consensus of the documents (high or low). However, based on our analysis of the transcripts and the recorded videos for the incorrect relevance judgments, we found the reasons that lead participants to make incorrect relevance judgments are the same despite the consensus levels (high or low).

4.5.1 Completion Time for Incorrect Relevance judgments

In Table 4.3, at the high level of consensus, we note that only 6 judgments of 108 non-relevant judgments were incorrect, and that in the relevant group 19 of 108 judgments were

incorrect. On the other hand, the relevance judgments at the low consensus level were split almost evenly, with 55 as non-relevant and 53 as relevant out of 108.

Table 4.3 illustrates that participants took more time when they were incorrectly judging the relevant (NIST qrel is 1) and highly relevant (NIST qrel is 2) documents in low and high levels of consensus (except for NIST qrel = 2 at the low level of consensus since there is only one relevance judgment). For example, the average completion time for incorrectly judging relevant documents in the high consensus level is 96.56 seconds while it is just 56.35 seconds when the judgments were correct. Also, the same thing occurred when incorrectly judging the relevant documents in the low consensus level and that is shown by the completion time of 65.05 seconds for incorrectly judging the relevant documents and just 57.35 seconds when judging the same documents correctly.

4.5.2 Categories of Incorrect Relevance judgments

Even though quantitative research methods are able to provide us with a good level of understanding of assessors' (users') behavior, qualitative research methods (in this thesis represented by the think-aloud method) are an excellent way to understand why secondary assessors make incorrect relevance judgments. Based on the analysis of transcribed data, we found that incorrect relevance judgments can be divided into four categories:

Search Topic: under this category, the participant is not able to apply the given search topic to the document. Sometimes, it was hard to know or unclear how to apply it. We call this *Difficulty in Applying Search Topic*. Difficulty in applying search topic occurs when the assessor finds what he/she thinks is relevant information. He/she rechecks the search

	$D^1=0$		$D^1=1$		TPD ²	Time ³
	#PD ⁴	Time ³	#PD ⁴	Time ³		
H.Cons: Not-Relevant	102	59.89	6	58.70	108	59.82
0	102	59.89	6	58.70	108	59.82
H.Cons:Relevant	19	96.56	89	56.35	108	63.42
1	5	115.63	31	66.65	36	73.46
2	14	89.75	58	50.84	72	58.41
L.Cons: Non-Rel/Rel	55	63.12	53	67.56	108	65.30
0	21	61.53	24	53.74	45	57.38
1	33	65.05	21	57.35	54	62.06
2	1	32.69	8	135.82	9	124.36
Total	176	64.86	148	60.46	324	62.85

Table 4.3: Relevance judgments and Average Completion Time

¹ Decision.

² Total Number of Participants Decisions.

³ Average Completion Time.

⁴ Number of Participant Decision.

topic to make sure that this information fits into the criteria mentioned in the description of the search topic. In the following example, Participant 2 is trying to decide if a document is relevant to topic 436 (Railway Accidents). Topic 436 requires the relevance assessor to find documents that provide data on railway accidents of any sort; however, documents that discuss railroading in general, (new rail lines, new technology for safety, and safety and accident prevention) are not relevant, unless an actual accident is described. In the example below, the assessor found some data about human deaths, but this was mentioned without describing an actual accident. He/she continued reading the document and found some information about safety. Therefore, he/she believes that since the document has some safety information in it, and because it mentioned data about human deaths then it is relevant. However, this document is judged by NIST assessors as not relevant and is among the high consensus documents (Not Relevant group) in our experiment.

China sets railway safety. OK, so does it have information on accidents? Yes, Human deaths. OK, And then, does it talk about prevention or safety? So documents that discuss railroading in general, new lines, new technology for safety. Safety and accident prevention. OK, but does it have prevention? So I would say it is still OK. *Participant 2.*

Document: even though the participant understands the search topic and knows what he/she is looking for, it is hard to process the document and find the relevant content. We call this *Difficulty in Processing Document*. The following example shows how the assessor's inability to process the document caused him/her to judge it incorrectly. From the transcribed data, we note that the assessor in this example knows what he/she should

look for in the document. However, he/she failed to find relevant content in the document. This document was judged as a highly relevant document by the NIST assessor, which means it does have relevant material in it but the participant missed it and was not able to find it.

Well, this article actually does not seem to be relevant because it does not discuss the frequency (Participant knew what he/she should look for.) It is just talking about one isolated incident and does not really talk about any causes of the attacks (Does not think this incident fit into the criteria). And there is no scientific, speculation or anything like that (Said as participant marks document non-relevant. The last sentence shows how the participant demonstrated understanding of the description of the search topic. However, he/she could not find the relevant content.) *Participant 12.*

Assessor: under this category, the participant lacks the required knowledge, which prevents him/her from judging the documents correctly or he/she lacks concentration. The following example illustrates how the assessor's lack of knowledge caused the incorrect judgment. In this example, the assessor was working on topic 310 (Radio Waves and Brain Cancer), which requires the relevance assessor to find documents that provide evidence that radio waves from radio towers or car phones affect brain cancer occurrences. As it is clear in the example, the assessor does not know that radio waves are a special kind of radiation and that caused him/her to stop reading and judged the document as not relevant.

So, I am scanning the article for cancer. Uhm, That says microwave. Radiation not radio towers. So, I would say this one is not relevant. *Participant 10.*

NIST Assessors' Mistakes: this category is different from the above three. It occurred with just one document and for a specific search topic. As we mentioned in this subsection, we compared our participants' relevance judgments against the judgments made by NIST assessors. Even though NIST assessors are well-trained, primary assessors, they still might judge documents incorrectly. We found that a non-relevant document to the NIST assessor was judged several times as a relevant document by our participants. The types of answers and proofs that participants provided led us to recheck that document carefully. After analyzing the whole document carefully, we believe that the NIST assessor was wrong in judging this document as a non-relevant one for the Mental Illness Drugs topic (Topic 383). The document in question has a docno of NYT19991214.0159.

Topic 383 requires the relevance assessor to find the names of drugs used to treat mental illness. The following two examples illustrate the judgments that were made by two different participants in the study in regard to the above mentioned document:

But then it mention specific, Prozac and Zoloft and depression (Participant has found a specific drug name). So it's relevant. *Participant 10.*

Prozac and Zoloft are laugh lines, even though such new drugs are effective in treating serious depression in many individuals (Has found specific drug names). So, that sentence makes it relevant. *Participant 11.*

	<i>UserDecision</i>		Total
	non-rel	rel	
H.Cons: Not Relevant	102	6	108
NIST qrel = 0:	102	6	108
Source of Error:			
Search Topic		5	5
Assessor		1	1
H.Cons: Relevant	19	89	108
NIST qrel = 1:	5	31	36
Source of Error:			
Search Topic	2		2
Document	2		2
Assessor	1		1
NIST qrel = 2:	14	58	72
Source of Error:			
Document	12		12
Assessor	2		2
L.Cons: Non-Rel/Rel	55	53	108
NIST qrel = 0:	21	24	45
Source of Error:			
Search Topic		19	19
NIST Assessor Error		5	5
NIST qrel = 1:	33	21	54
Source of Error:			
Search Topic	1		1
Document	29		29
Assessor	3		3
NIST qrel = 2:	1	8	9
Source of Error:			
Document	1		1
Total	176	148	324

Table 4.4: Categories of Incorrect Relevance judgments

Table 4.4 illustrates the categories of the incorrect relevance judgments. If we look carefully at the numbers in this table, we find that false positives (FP), which are non-relevant documents incorrectly judged as relevant, are caused by difficulty in applying the search topic.

On the other hand, false negatives (FN) or misses, which are relevant documents incorrectly judged as non-relevant, are caused by assessors having difficulty in processing documents. In general, this type of error means that the assessor was unable to find relevant material in the document.

The error types and their occurrences can be explained with a simple model of the relevance judging process. Figure 4.3 is a flow chart of the judging process as we understand it. In this figure, an assessor starts the relevance judgment process by searching for relevant material in the document on hand. If the assessor is not able to find relevant material, then the assessor judges it as not relevant. If potentially relevant material is found, the assessor will check if this material fits into the topic's description. If the assessor is not able to find a fit with the search topic's criteria, the assessor will judge the document as not relevant. Otherwise, the document is judged as relevant. Area A in Figure 4.3 represents what we earlier called "difficulty in processing the document." where the whole figure represents what we call "Difficulty in applying search topic".

Given this relevance judging process, document processing issues only result in false negatives, i.e. secondary assessors judging a document to be non-relevant that a primary assessor has judged to be relevant. If we want to prevent false negatives, we need to build relevance assessing systems that assists assessors in finding relevant material in documents.

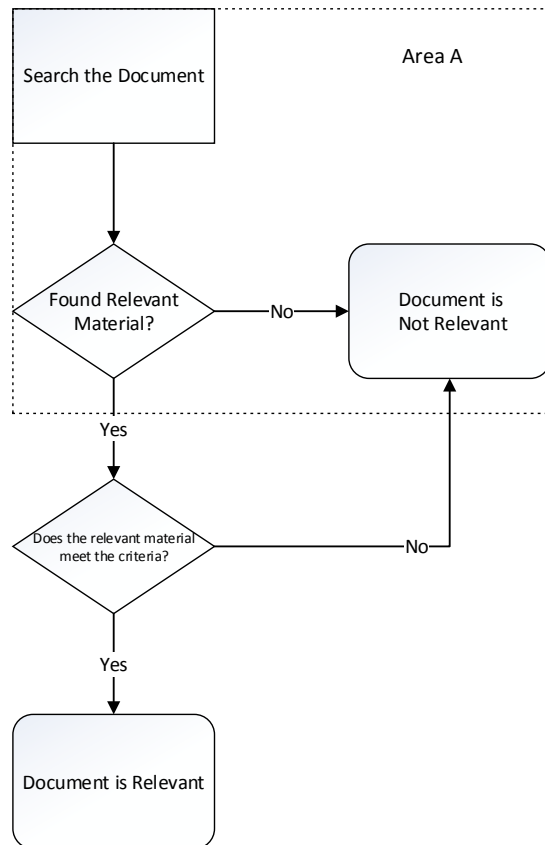


Figure 4.3: Relevance judgment Process. Area A represents “Difficulty in Processing Document”, while the whole figure represents “Difficulty in Applying Search Topic”

Difficulty in applying the search topic can result in judging issues during the search of a document or during the decision concerning topic fit. For example, if an assessor does not understand a topic, this can result in either missing relevant material or in the misidentification of material. Once the assessor finds potentially relevant material, difficulty with the search topic can again produce differences as the assessor makes his/her final relevance decision. In the case of documents that had high consensus, the few errors caused by the

search topic seem divided between false positives and negatives. In the case of low consensus documents, search topic errors were much more frequent and tended to result in false positives. In general, secondary assessors seem to understand the broad search topic and can identify on-topic material, but it is the application of detailed topic criteria that leads to false positive errors.

4.6 Judging Behavior

We know from previously published research that assessors' behavior differs. They are like any human group: some tend to share or almost share the same characteristics while others are totally different. In our study, we show that assessors (participants) are divided into three groups based on their speed of judging. However, being in one group does not mean that their true positive rate (trp), false positive rate (fpr), etc. are the same. In the following paragraphs, we explain and discuss everything we have found out about these groups. The three groups are:

Slow Deciders: Participants in this group needed more time to judge the relevancy of a document. Based on our results, taking more time does not imply more accurate judgments; it simply represents the way these participants conduct themselves. For instance, participant 5 is considered a slow decider based on the classification that we used. However, if we look at Table 4.5, we notice that his/her true positive rate is only 47.37%. On the other hand, the true positive rate for participant 11, who is also considered a slow decider, is 63.16%. Therefore, and based on the data in Table 4.5, it seems there is no relationship between a good true positive rate and being a slow decider.

Medium Deciders: These type of participants come in between slow and fast deciders. They do not take the same amount of time that fast deciders takes to judge, but they spend some time reading and thinking about the relevance of documents. Again, their trp and fpr are different.

Fast Deciders: A fast decider does not spend a great deal of time coming to a decision. He/she usually goes over the document quickly, and he/she sometimes provides a judgment based solely on the title or the first couple of lines of the document.

Also, we have found that slow and medium deciding assessors tend to talk or provide more explanations about their relevance judgments than fast deciding assessors. For instance, the average number of words for the slow and medium deciders groups are, respectively, 676.42, and 797 words, while it is just 318.75 words for the fast deciders group. The reason that the average number of words in the medium deciders group is greater than the slow decider group is the behavior of participant 9. Participant 9 tends to talk more than any other participant in this study, and he/she provides some unrelated comments about the search task in hand. His/her way of explaining is very quick, and it covers a lot of points, with most of these points being unrelated to the given search task.

Furthermore, being a fast decider does not indicate that the assessor is careless in his/her judgments. Some participants who are classified as fast deciders were accurate in most of their judgments. For instance, Participant 3, as illustrated in Table 4.5, is a fast decider. When we look at his/her true positive rate, which is 89.47%, we find that he/she was correct most of the time, regardless of his/her decision behavior.

Also, for the criterion c column in Table 4.5, we find that most of the assessors in the study have a conservative behavior and (The value of the criterion c is positive). We just notice a liberal behavior in the case of participants 3 and 8. Again, this does not imply that they are not providing correct relevance judgments. For these two liberal behaving assessors, their tpr and fpr are, respectively, 89.47%, 11.76% and 63.16%, 17.65%.

Group	Participant	Time	TPR	FPR	d'	Criterion c	Avg. # Words
Slow	P5	105.93	47.37	17.65	0.86	0.497	630.5
	P11	85.08	63.16	17.65	1.26	0.29	860.5
	P12	79.56	63.16	23.53	1.05	0.19	538
Average		97.60	65.79	23.53	1.19	0.13	676.42
Medium	P4	62.16	63.16	11.76	1.5	0.43	750.25
	P9	62.97	84.21	11.77	2.189	0.09	1219
	P10	54.69	68.42	29.41	1.02	0.03	421.75
Average		58.88	69.47	18.82	1.45	0.20	797
Fast	P2	42.98	53	29	0.60	0.24	439.5
	P3	45.26	89.47	11.76	2.40	-0.03	243.25
	P8	27.03	63.16	17.65	1.26	-0.27	273.5
Average		38.42	68.54	19.47	1.42	-0.02	318.75

Table 4.5: Judging Behavior

4.7 Short vs. Long Search Topics

In examining the differences in relevance judgments at a per-topic level, we noticed a difference in the rates at which participants misidentified relevant documents as non-relevant and vice versa. As shown in Table 4.6, the false negative rate (FNR), which is defined as

the fraction of relevant documents incorrectly judged as non-relevant, is higher when the description of the search topic is long. Our interpretation for this kind of behavior is that the description provides more details and more explanation, which is sometimes hard to remember by assessors or might mislead them. Moreover, when participants encounter this type of description, they seem to focus on one part of the description during the relevance judgment process while they forget about the other details in the description. For example, the description of topic 336, which is about Black Bear Attacks, is shown in Figure 3.1.

If we look at the description of topic 336 in Figure 3.1, we note that the first, the third and the last sentences describe what should be considered when an assessor looks for relevant information. These three sentences are as shown below and we start with the first sentence:

A relevant document would discuss the frequency of vicious black bear attacks worldwide and the possible causes for this savage behavior.

the third sentence is as the following:

Relevant documents would include the aforementioned causes as well as speculation preferably from the scientific community as to other possible causes of vicious attacks by black bears

finally, the last sentence says the following:

A relevant document would also detail steps taken or new methods devised by wildlife officials to control and/or modify the savageness of the black bear.

We can see clearly that every sentence in the above three sentences adds more data to consider in the relevance judgment process. We found from the transcribed data that participants are not able to accumulate all of these points in the description. The following examples are taken from the transcribed data:

So, boy is attacked by bear at scout camp. So, lets look to see if there is anything about frequency of attacks or causes. Talks about the incident. Or what happened during the attack. Candy wrappers were found in the tent, but officials did not know if that was related to the attack. It does not talk about the causes or frequency. It is just one incident. So, not relevant. *Participant 11.*

In the above example, the document that the participant was judging was judged as highly relevant by NIST assessors (score 2). However, participant 11 was trying to find information related to the first and the third sentences of the description of the search topic. However, the participant did not focus at all on the last sentence of the description, where the document on hand was discussing steps taken or new methods devised by wildlife officials to control or modify the savageness of the black bear. Giving participants more details in the description caused them to concentrate on only parts of the descriptions and ignore the others. As a result, a relevant document was missed due to this problem. The following example also illustrates this behavior:

So this just seems like a bear attacking a person but a bear being in a property. So, Uhm, and attacking the goats. So, I don't know. It is relevant yet. Ahh, Maybe an about [...]. The bear showing at a barbeque. So, I don't know if that

Topic	Topic Length	<i>ErrorRate</i>	
		False Positive Rate	False Negative Rate
310	Long (117 words)	0%	44%
336	Long (93 words)	5.6%	53%
436	Short (58 words)	37.3%	2.8%
383	Short (33 words)	24%	8%

Table 4.6: Assessors' Behavior with short and long topics' description

counts as an attack. Just scare some people. Yeah, This is more about [...].
 Ahh, Oops! So, I am going to say not relevant because it is just about scaring
 people. They are not actual attacks. *Participant 5.*

In the above example, Participant 5 was also looking just for the causes of the attacks and not considering the other parts of the description which might lead him/her to find the relevant information.

On the other hand, we found that the false positive rate (FPR), which is defined as the fraction of non-relevant documents incorrectly judged as relevant, for search topics that have short descriptions is higher than those which have longer descriptions as illustrated in Table 4.6. Since there is only a little information in the description about what to consider when looking for relevant documents, this lead participants to be more liberal and to consider non-relevant documents as relevant. This type of description encourages participants to be afraid of skipping or not seeing relevant information in the document; therefore, in order to be safe, they prefer to judge non-relevant documents as relevant. Figure 3.3 represents an example of a short topic's description.

Chapter 5

Low Consensus Documents

In this chapter, an in-depth analysis of the low consensus documents is given. We have used two types of documents in our Think-Aloud Study: Low and High consensus documents. The criteria that constitute the selection process of these documents are mentioned in *Chapter 3*.

5.1 Summary

In this section, we provide a summary of the findings for each of our search topics. We discuss what has been observed and found when judging the low consensus documents. Through the think-aloud data, we have obtained a deeper insight into secondary assessors relevance judging behavior. To the best of our knowledge, no such a study was conducted on the relevance judging behavior of secondary assessors.

Black Bear Attacks (Topic 336)

Topic 336 has a long description. It contains many points to consider when searching for relevant material in a document. An assessor should consider all of the points in a description. From what we have seen in the think-aloud data, assessors may focus on just one of the criteria and disregard the rest, causing them to overlook a relevant document. Also, documents that fall under this topic might not explicitly mention actual black bear attacks during which people were injured or killed. However, they might mention black bear invasion of peoples' homes and the factors that attracted bears to the properties. Some assessors may not judge these documents as dealing with black bear attacks since casualties or injuries are not mentioned. To improve assessors' accuracy, we think it is necessary that assessors complete a thorough tutorial in which all conditions of the search topic are reviewed. We believe this will help to reduce the amount of confusion and number of incorrect relevant judgments.

Radio Waves and Brain Cancer (Topic 310)

This search topic is not as straightforward to judge, since it requires greater consideration of the material in the document in order to locate relevant information. The description of this search topic includes many points that should be considered by an assessor while judging documents. One of the most common issues that we found assessors to encounter when judging documents for this search topic is the type of terminology. Documents for this search topic include many terms that might be less familiar to some secondary assessors. It is important to understand terms such as radio-frequency, radiation, ultra-high waves and waves in general in order to be able to judge the documents correctly. Moreover, not referring to the topic of brain cancer specifically is another source of confusion to

assessors. The documents occasionally discuss the harmful effects of cell phone waves and their impact on the growth of cancer; however, brain cancer is not explicitly mentioned. What makes the decision even more difficult is that it mentions both studies that either support or object this claim, since the description of the search topic instructs assessors to identify whether studies and experiments support or do not support these claims. In other cases, the documents talk about cell phone waves and their impact on several types of cancer and changes in brain activities. We found some secondary assessors to be wondering whether the term “brain activities” is equivalent to brain cancer or something different.

Mental Illness Drugs (Topic 383)

The nature of the search task on this topic is to locate names of drugs for mental illness. The most commonly cited issue with this type of search topic is that it requires a fairly high degree of familiarity with the subject of mental illness in general, the names of conditions, and related drugs. However, secondary assessors represent different backgrounds, skills, educational levels, and specialties; they also have no previous training on making relevance judgments. This makes the process of making relevance judgments for a topic such as this challenging due to the type of terminology that is encountered.

Also, some assessors, due to their inability to clearly comprehend what is stated in the description of the search topic, make incorrect relevance judgments. For example, in several instances when judging low consensus documents on this search topic, assessors were found to have chosen incorrect drug names. They believe, consider, or even guess that they are making correct decisions. However, these are in fact not correct. The Think-Aloud data revealed to us that though assessors occasionally make correct relevance judgments, this however, does not mean that the reasoning or rationale behind it is correct.

Moreover, the length of a document and location of the drug name in the document might also play an essential role in making it low consensus. In fact, we found this to be a reason for an incorrect relevance judgment made even by the NIST assessor. One of the documents included the drug name; however this occurs in a single line of the last third of the document. Therefore, we assume that the NIST assessor for an uncertain reason did not observe it and then decided to judge it as not relevant.

We believe most of issues that contribute to a document being considered low consensus might be solved by providing a clear and concise description of the search topic. We also suggest that assessors to be provided with a reference list of all names of mental illnesses and drugs. These steps will alleviate most dominating doubts assessors may experience while making relevance judgments. Therefore, a reference list should be available during the entire process of relevant judgments of a search task.

Railway Accidents (Topic 436)

This search topic is not very long and requires locating information on the subject of railway accidents; a second condition is that if the document is on the topic of safety or railway systems, it is not relevant unless it specifically describes a railway accident. However, what constitutes “sufficient description” is unclear to secondary assessors, unlike primary assessors who have better knowledge regarding whether a document contains adequate description about an accident. The documents for this topic contains a number of sentences that were confusing to some assessors. Confusion results from the type of information presented in the sentences. When considering such a search topic, we suggest that assessors to be tested on their ability to identify what is sufficient for a document to be judged as relevant. If an assessor is not performing well, he/she should be excluded from the main

search task of making relevance judgments. Furthermore, assessors should be provided with additional examples during the tutorial about what qualifies a document to be relevant. We believe that assessors should judge as many documents as possible in the tutorial, and they should be exposed to all criteria of a search topic in order to reduce the level of confusion they might experience in the main search task.

5.2 Definitions

We analyzed participants relevance judgments based on high and low consensus levels as well as NIST qrel scores. In this chapter, we discuss the ones that we found assessors to express when they judge low consensus documents. The the following two subsections lay the ground to more understanding of the coming sections in this chapter by providing definitions and examples for each of the reasons and causes we found our assessors to express or commit.

5.2.1 Assessor Causes of Differences

After analyzing the transcribed data and the video recordings, we found the following specific assessor causes of judgment differences:

- **Trouble Understanding the Search Topic:** this is caused by a participant who does not fully understand the search topic or part of it. Here are examples from the think-aloud transcripts where the participants misunderstand what a generically named drug is:

Specific or generic type of drug. Uhm. So, I guess that's a generic type of drug. Antipsychotic. Uhm. So I guess I would say it's relevant. *Participant 10*

It talks about antipsychotic drugs. I guess that's generic type of drug. I think this is relevant. *Participant 8*

- **Lack of Familiarity/Lack of Knowledge:** the participant is not able to figure out if the document is relevant or not due to a lack of familiarity with the given search topic or a lack of familiarity with some words or phrases in the document. Here is an example of this:

Not sure of that depression is a mental illness. So, I am going to think about that. Okay, so Ritalin is called stimulant because it belongs to a family of drugs that stimulate the central nervous system. Okay, Ahh, Parent's alertness and ability to pay attention. Oh, increases a person's alertness. So, I don't think that is about the mental illness. I am not quite sure what qualifies a mental illness. I am thinking. *Participant 5.*

- **Lack of concentration:** In very rare cases in our collected data, lack of concentration, which we describe as a kind of user behavior, was noticed. Lack of concentration was the cause of simple mistakes. Participants judged documents as relevant while those documents were entirely unrelated to the search topic.

In the below example, Participant 2 judged the document as relevant even though it discussed grizzly bears and not black bears. He/she did not notice this, and at-

tempted to fit the document into the criteria mentioned in the description of the search topic. This is clear when the participant said *See if there is any recommendations* and *See if there is any ways to control it*. The participant said that he/she knows generally about the search topic Black Bears Attacks in the pre-task questionnaire and Grizzly bears are totally different than Black Bears.

Behind conflict over new grizzly program, an endangered species war. Four weathered men stand around, dropped tailgate. Fruits leading cave. This is the kitchen that where the food is. Okay. Okay it was talking about food that's a cause. Threatened animals. See if there is any recommendations. See if there's any ways to control it. Oh. Okay. That's good. *Participant 2*.

5.2.2 Assessor Decision Making

In examining the transcripts in those cases where assessors made judgments that differed from the primary assessor, we found the assessors gave the following reasons for their decisions:

- **Insufficient Information:** Sometimes, participants might think that the document does not have enough information and this will cause incorrect judgments. The following are some think-aloud data examples from the transcripts:

So this is not relevant at all, we're not, it's just too short, we're not getting enough information about anything. *Participant 4*

I don't see any study type thing. So, I don't think it is relevant. It doesn't have information. *Participant 11*

- **Presence or Absence of Specific Evidence:** Participants affected by this factor cannot find evidence (specific information) or they think they find the evidence. Here is an example:

So it's not relevant because it didn't provide any evidence of that connection yet. Just about certain studies. *Participant 5.*

- **Lack of Topicality:** Participants incorrectly think that the document is off-topic. What follows is an example of this:

It's not talking about an incidents or anything. It's just talking about some crazy lady and why you should keep bears. *Participant 4*

- **Absence of Keywords:** Some participants were looking mainly for keywords in the documents. If key words were not there, that meant the document was not relevant to them.

Uhm. Black bear hunting. Black bear hunting, okay. But it is not black bear attacking. Okay. Why don't we search for attack. Search down. It is not there. Okay. Not relevant. *Participant 9.*

5.2.3 Findings

Table 5.1 summarizes all the causes and reasons we found assessors to express when judging low consensus documents.

Number of Documents	12
NIST Score	2 (1 Document) – 1 (6 Documents) – 0 (5 Documents)
Correct Relevance Judgments NIST Score (1 & 2)	Presence of Specific Evidence Presence of Keywords & Lack of Concentration Trouble Understanding the Search Topic
NIST Score (0)	Absence of Specific Evidence Lack of Topicality Insufficient Information
Incorrect Relevance Judgments NIST Score (1 & 2)	Absence of Specific Evidence Absence of Keywords Lack of Topicality Lack of Familiarity
NIST Score (0)	Trouble Understanding the Search Topic Presence of Specific Evidence Insufficient Information

Table 5.1: Causes of Differences and Assessors Decision Making For Low Consensus Documents

Title	Lawmaker introduces bill to ban bear hunt
NIST Score	1
Length (#Words)	559
ProbRelevance	0.5
Has Key Words	Yes
No. of Judgments	“black bears” , “bears” 8 Non-Relevant - 1 Relevant

5.3 Topic 336: Black Bear Attacks

Document ID: APW20000323.0200

Document’s Description

The document discusses a bill to ban bear hunts in New Jersey. It also considers the problem of increasing bear numbers, and how this current phenomenon is creating fear among people in the NJ area. Additionally, it mentions that complaints of a larger and more brazen black bear population in northern New Jersey which has invaded suburban back yards, killed pets, and scared small children, has promoted the state Fish and Game Council to vote authorize New Jersey’s first bear hunt since 1971.

“Bears come right on our front porch and get in the birdseed” said Slate, 62, of Wantage.
 One little bear, maybe about 150 pounds, was here night before last. He takes our bird feeders down, even broke our lattice under our porch to get feed that had spilled onto the ground.
 Complaints that the larger and more brazen black bear population in northern New Jersey have invaded suburban back yards, killed pets and scared small children prompted the state Fish and Game Council to vote this week to authorize New Jersey’s first bear hunt since 1971.
 State officials have tried "aversion conditioning" to scare off bears with pepper spray or rubber buckshot and urged residents not to feed the bears and to bring their garbage indoors at night.

Figure 5.1: Possible Relevant Sentences for DocID: APW20000323.0200

Causes of Not Relevant Judgments

- **Absence of keywords.** For example, Participant 2 judged the document as not relevant because he/she did not find any keywords in the document. He/she was looking for the keyword “attack” but did not find anything in the document regarding this. He/she stated, “What did they say about the attacks. It doesn’t even mentioned so I’d say irrelevant”. For this participant, not finding keywords was a sufficient reason to judge the document as not relevant.
- **Lack of Topicality.** Seven participants of 9 stated that the document is not relevant because it does not meet the criteria in the description of the search topic and considered the document to be on the topic of bear hunting, not about black bear attacks and this is what the document’s title imply. For example, Participant 10 said “So I’m scanning. It’s talking about bears but, I’m looking for attack data. And there’s something about, But it does mention control and modify the savageness. Uhm. So I’d say it’s not relevant. There isn’t talk about attack”. He/she judged the document as not relevant because it does not discuss the topic of black bears attacks, even though he/she located some data about controlling and modifying the aggressive behavior of black bears. For this participant, lack of reference to the topic of attacks was sufficient reason to judge the document as not relevant.

Causes of Relevant Judgments

- **Presence of Specific Evidence.** Participant 3 judged it as relevant because it meets one of the criteria that are stated in the description of the search topic, which is about controlling the savageness of the black bear. He/she stated, “So this is talking about

black bear damages and, but it seems more about the ... whether or not they should be hunted but it does mention tactics for keeping away bears. It does mention ways to control bears. So I guess that's serves relevant". He/she decided this, after reading the line which includes reference to "aversion conditioning ...". This participant shows different behavior than Participant 10 in the previous paragraph since he/she found the document to meet the criterion mention in the last sentence of the search topic's description.

Reflection

The document is graded as 1 by a NIST assessor. From initial glance and based on the title, the document seems to be about bear hunting in general. However, it also includes several points that might make relevant, and these include the following:

- It discusses what attracts bears, causing them to attack peoples' homes: "Bears come right on our front porch and get in the birdseed." said Slate, 62, of Wantage. It also mentions "One little bear, maybe about 150 pounds, was here night before last. He takes our bird feeders down, even broke our lattice under our porch to get feed that had spilled onto the ground."
- Complaints include that the larger and more brazen black bear population in northern New Jersey have invaded suburban back yards, killed pets and frightened small children promoting the state Fish and Game Council to vote this week to authorize New Jersey's first bear hunt since 1971.
- It also addresses that State officials have tried "aversion conditioning" to scare off bears with pepper spray or rubber buckshot, and urged residents not to feed the bears and to bring their garbage indoors at night.

This type of document does not directly discuss black bear attacks or actual incidents involving black bear attacks. It rather discusses what attracts the bears to peoples' homes. Another section of the document consider several methods of control that were used by wildlife officials to manage black bears' behavior. Therefore, quick scanning or skimming of the document is not very helpful in locating relevant material for a secondary assessor.

Document ID: APW20000622.0185

Title	Fish and Game Council approves black bear hunts in NJ
NIST Score	1
Length (#Words)	460
ProbRelevance	0.5
Has Key Words	Yes
	“black bear”
No. of Judgments	8 Non-Relevant - 1 Relevant

Document's Description

The document deals with the topic of approval of hunting black bears in NJ. It is similar to the previous document, except it focuses more on approval of black bear hunting in NJ, while the previous document focuses on the banning of black bear hunts in NJ. It places greater emphasis on the issue of the increase in the bear population and how this now is alarming people in NJ. It mentions, “Residents of many northern New Jersey communities complain that the bears invade their back yards, kill their pets and scare their children. It further describes how the complaints of damage caused by black bears have increased from 285 in 1995 to 1659 in 1999. The document also talks about the intention to reduce the density of black bears as a way to control the damage to NJ residents and their properties. It further considers how State officials have tried “aversion conditioning” to scare off bears

with pepper spray or rubber buckshot, and urged residents not to feed the bears and to bring their garbage indoors at night.

Causes of Not Relevant Judgments

- **Lack of Topicality.** Seven participants stated that the document is not related because it does not meet the criteria mentioned in the description of the search topic. They believed the document to be more about bear hunting not black bear attacks. For example, Participant 2 took just very quick look at the document and noticed that the title is about hunt, then the participant decided to judge it as not relevant. He/she said Hunt! We are not really talking about hunt. No. Participant 11, thought that the document was not relevant because it does not discuss black bears attacks. It only refers to bear hunting. He/she said “But again it doesnt say anything about, About an attack. It’s just talking about bears getting in to peoples garbage and using other methods. So, it’s just complaints. Not relevant. There’s no specific violent behavior or savage behavior.”. Participant

Causes of Relevant Judgments

- **Presence of Specific Evidence.** Participant 3 thought it was relevant because it meets one of the criteria that is stated in the description of the search topic, which is about controlling the savageness of the black bear. He/she stated, “So that seems like prevention mechanism and that seems related. That’s about ways of controlling bears.”. Even though the participant here did not find data on actual black bear attacks which represents the first criterion in the description of the search topic,

he/she found data on ways of controlling the savageness behavior of black bears which represent the last criterion in the description.

Reflection

The document is graded as 1 by a NIST assessor. That which has been stated about the previous document applies here as well. The relevant material is not explicitly presented. Careful reading and consideration of the content is required.

Document ID: APW20000703.0186

Title	Athlete Killed by Bear Attack
NIST Score	0
Length (#Words)	237
ProbRelevance	0.533
Has Key Words	Yes
	“bear attack”, “black bear”, “attacked”, “bear”
No. of Judgments	8 Non-Relevant - 1 Relevant

Document’s Description

The document discusses the killing of a 24-year athlete and unknown causes of death. It considers that the authorities are not certain whether the black bear attack was the main cause of death or if she suffered a heart attack or another factor occurred, permitting the animal to find her injured.

Causes of Not Relevant Judgments

- Absence of Specific Evidence. Participant 12 believed the document said nothing about the cause of the athlete’s death. He/she decided it is not relevant when he/she found no mention of the cause.

A 24-year-old athlete was killed in an apparent bear attack while running on a training course, police said Monday. Black bear tracks also were spotted near her body.
The autopsy will confirm whether the bear was the exact cause of death or whether she had a heart attack or something before and the animal found her injured," Coup-Fabiano said.

Figure 5.2: Attractive Sentences for DocID: APW20000703.0186

- Absence of Specific Evidence (Single case). Participant 4 was skimming the document looking for relevant material. However, he/she thought it was not relevant because it discusses only a single case.

Causes of Relevant Judgments

- Presence of Specific Evidence. However, Participant 5 here did not completely understand the search topic. Participant 5 was reading the first two lines of the document and then decided it to be relevant, stating, “this is quite relevant”. He/she did not continue to read the document further, in order determine whether this incident meets the criteria in the search topic description by stating the case of the attack or the other criteria in the description of the search topic.

Reflection

The document is graded as 0 by the NIST assessor. This document might be a source of confusion for assessors because of its title “Athlete Killed by Bear Attack”, may mislead one to consider that it might be relevant. The document discusses a single case where the cause of the death is unknown. This type of document requires not only a good understanding of the search topic but also careful processing of the information in the document since the

cause of the attack is not mentioned as well as the steps taken to control the aggressiveness behavior of black bears.

5.4 Topic 310: Radio Waves and Brain Cancer

Document ID: NYT20000224.0139

Title	How Business Gets What it Wants
NIST Score	1
Length (#Words)	773
ProbRelevance	0.545
Has Key Words	Yes “brain” , “cancers”
No. of Judgments	7 Non-Relevant - 2 Relevant

Document’s Description

The document discusses how microwave radiation generated by cell phones can increase a number of health issues, and among these is “brain barrier”. In another part of the document there is a mention that long-term, low-level exposure to the ultra-high frequencies in the microwave band have been implicated in increases in a variety of health problems, and among these “brain tumors”. The document focuses in general on how people presently are not able to defend their rights to living a healthy life, and that large companies ultimately obtain what they want, even at the expense of people’s health.

Causes of Not Relevant Judgments

- Absence of Keywords. Participant 2 found many instances of the word “brain” in the document. However, he/she was specifically searching for keywords like “brain

According to B. Blake Levitt, author of "Electromagnetic Fields: A Consumer's Guide to the Issues and How to Protect Ourselves," before 1996 protesters were able to cite numerous studies as well as data collected from foreign military bases that indicated various health effects from nonionizing microwave radiation generated by cell phones among them increased permeability of the blood-brain barrier; damage to the immune system; numerous cancers; and DNA damage".

Unlike the output of a 100-watt light bulb, 100 watts of ultrahigh frequencies in the microwave band are maximally absorbed by human tissue, with the result that long-term low-level exposure has been implicated in increases in breast cancer, leukemia, cataracts, immune suppression, and brain tumors.

Figure 5.3: Possible Relevant Sentences for DocID: NYT20000224.0139

cancer", and then subsequently he/she searched for "radio towers" and "car phones". When he/she found no connection which was clear to him/her, he/she judged the document to be not relevant. He/she stated, "Okay. Brain cancer. Ahh. Blood-brain barrier. High on microwave. Human tissues. Okay. But is there any relation to car phones? Okay. Radio towers? No. I don't think so".

- **Absence of Specific Evidence.** Participant 4 believed that the document is about political positioning. He/she said, "I am going have to say this is just a punch of political positioning or something about this issue. That makes it not relevant. It does not seem relevant to me". Participant 9 also shared the same opinion with Participant 4 and said, "This is an opinion article ... This is opinionation. Yeah. It's not really relevant".
- **Lack of Familiarity.** Confusion of electromagnetic and radio waves caused Participant 5 to judge the document as not relevant. He/she said "microwave radiation no it is probably not radio waves". Participant 10 was searching for the keyword "cancer". His/her lack of knowledge and familiarity with "microwave" and "radio

towers” caused him/her not to understand and observe the connection, and a result judged the document incorrectly. He/she said, “I am scanning the article for cancer. ... hmm, that says microwave radiation not radio towers”. Therefore, he/she judged it as not relevant.

- **Absence of Specific Evidence.** Participant 12 for example, did not find that the document discusses specific studies or experiments dealing with the issue of brain cancer and cell phones. He/she said, “Based on the criteria, it does not really talk about experiments with animals or .. well ... it does say that protesters cited numerous studies but ... does that mean it is relevant ... I am not a 100% sure... Does not really talk about these studies it mentions these studies were cited. I don’t know. I don’t think this is a relevant document”. Also, Participant 11 judged it as not relevant for the same reason. He/she said, “Talks about the effect of the ultra high frequencies on human tissues but it is not really a study providing an evidence just says that. I don’t see any study type thing”.

Causes of Relevant Judgments

- **Presence of Specific Evidence.** Participant 3 asked a self-question regarding whether mention is made of any specific evidence by saying, “Does it say anything specific?” and then responded by saying, “Yeah, it mentions a specific problem”. While he/she was answering, the movement of the computer’s mouse was on, “Unlike the output of a 100-watt light bulb, 100 watts of ultrahigh frequencies in the microwave band are maximally absorbed by human tissue, with the result that long-term low-level exposure has been implicated in increases in breast cancer, leukemia, cataracts, immune suppression, and brain tumors”. Also, Participant 8 said, “From the title and

the beginning, I do not see ...”, he/she meant that he/she does not see how this document is relevant. However, he/she continued to read and identified a paragraph where there is a mention of microwave radiation. He/she read this paragraph and said that the document is relevant. This is an excerpt from paragraph he/she read, “before 1996 protesters were able to cite numerous studies as well as data collected from foreign military bases that indicated various health effects from nonionizing microwave radiation generated by cell phones among them increased permeability of the blood brain barrier; damage to the immune system; numerous cancers and DNA damage”.

Reflection

The document is graded as 1 by NIST assessor. It does not talk directly about the problem, even though it mentions the harmful effects of radio towers in general. This type of document requires assessors to be familiar with the keywords and terms, as well as have a fair amount of knowledge on the topic since there is a mention of some electrical terms that might be unclear to some of them. As we pointed out in earlier chapters that secondary assessors come with different levels of knowledge and skills.

Document ID: XIE19970506.0203

Title	Israeli Company Recommends Using Headsets to Cut
NIST Score	1
Length (#Words)	328
ProbRelevance	0.571
Has Key Words	Yes “cancer”
No. of Judgments	3 Non-Relevant - 6 Relevant

Document's Description

The document discusses suspected health damage from cellular phones. It does not include any specific information about the brain. However, it mentions, “Bachar was commenting on media report about what some experts called *the first serious piece of lab research* to link electromagnetic radiation from cellular phones with cancer in mice”.

Bachar was commenting on media report about what some experts called “the first serious piece of lab research” to link electromagnetic radiation from cellular phones with cancer in mice.
The study, which was conducted by Dr. Michael Repacholi at the Royal Adelaide Hospital in South Australia and published in Radiation Research, found that exposure to functioning cellular phones doubled the risk of lymphatic cancer in mice.

Figure 5.4: Possible Relevant Sentences for DocID: XIE19970506.0203

Causes of Not Relevant Judgments

- Absence of Keywords/Absence of Specific Evidence. Participant 2 in this document was looking for specific keywords “brain” and “brain cancer” and did not find these. He/she stated, “Okay, it does not look like it is related to brain”. He/she scanned the document very quickly and found a sentence “fears of a link between cancer and cellular phone use”. He/she said at the end, “it does not have any experiment”. Participant 5 based his/her decision on just a single line of the document. He/she was focusing entirely on this one line, “until now, research studies have shown no danger to health”, when he made a judgment and stated, “Again it is about not showing any danger to your health. So, I am going to say it is not relevant”. Also, not mentioning “brain cancer” specifically, caused Participant 10 to judge the document as not relevant. He/she said, “So, I see cancer, but it doesn’t say brain cancer. Hmm,

lymphatic but not brain. So, I am just looking back at the criteria and it is brain cancer so it is not relevant. It is just other types of cancers”.

Causes of Relevant Judgments

- **Presence of Specific Evidence.** Though the document does not state anything specific about brain cancer, Participant 3 thought it is relevant because it discusses the topic of cancer in general. The participant was somewhat confused whether car phones and cell phones refer to the same item. He/she said, “I am kind of confused as to other cell phone and car phones are the same thing? I am going say that they are because I have never heard of a car phone . So, I have to look for brain cancer specifically”. He/she said, “It does not mention brain cancer specifically but it does seem relevant to cancer in general”. Also, Participant 12 was confused whether a single line of information is sufficient to judge a document as relevant. However, he/she was guessing and then subsequently decided to judge it as relevant. He/she said, “it mentions one line about linking electromagnetic radiation from mobile phones to cancer in mice in a lab. So, I guess that falls under the relevant category based on the description”. However, Participant 4 found the document to be very relevant. He/she said, “This is definitely relevant”. He/she said that while was reading the paragraph he/she located this point, “ found that exposure to functioning cellular phones doubled the risk of lymphatic cancer in mice”. Participant 11 thought the document does not report any evidence at the beginning. However, after reading several lines, he/she found a section of the paragraph in which he/she thought the material to be relevant. He/she was reading, “found that exposure to functioning

cellular phones doubled the risk of lymphatic cancer in mice”. He/she then decided that the document is relevant.

- Presence of Keywords/Lack of Concentration. Participant 9 found several instances of the word “cancer” in the document and then decided to judge it as relevant. However, he/she here did not focus while making the judgment and was speaking about non-related topics involving Israeli issues, such as the Palestinian and Israeli conflict.

Reflection

The document is graded as 1 by a NIST assessor. It is difficult to judge, since it has no specific information on brain cancer. However, it talks about how the radiation that is emitted from cell phones has a positive connection to lymphatic cancer in mice. Therefore, the mention of cancer an cell phone in the document was enough to convince some participants that the document is relevant.

Document ID: XIE20000628.0163

Title	WHO Issues New Recommendations on Use of Mobile Phones
NIST Score	1
Length (#Words)	345
ProbRelevance	0.545
Has Key Words	Yes “brain”, “cancer”
No. of Judgments	6 Non-Relevant - 3 Relevant

Document’s Description

The document discusses the recommendations made by World Health Organization (WHO)

in regard to the use of mobile phones. The document does not mention brain cancer in specific, however, it points out at changes in brain activities and other types of cancers.

Health conditions which have been ascribed to the use of mobile phones include some types of cancer and changes in brain activity.
The need for new studies to confirm recent findings that mobile phones may cause changes in brain activity, reaction time, and sleep patterns.

Figure 5.5: Possible Relevant Sentences for DocID: XIE20000628.0163

Causes of Not Relevant Judgments

- Absence of keywords/Lack of Topicality. Participant 2 was searching for keywords “animal”, “cancer” and then “brain cancer”. He/she read only the first several lines of the document and started to look for keywords after this. He/she said, “So it doesn’t look like they make it too much to brain cancer. So it’s not relevant”.
- Absence of Specific Evidence. Participant 4 believed the document is only about recommendations, with no reporting of incidents. He/she said, “These are recommendations. So, this is not relevant because they are not reporting on the incident at all. They are just basically giving recommendation”. He/she did not make an additional effort to read with greater attention, between the lines and find relevant material. He/she scanned the document very quickly and only focused on a few first lines. Participant 5 was also focusing on the beginning of the document to decide its relevance. After reading several lines and scanning the first half of the document, he/she said, “There is no data provided that shows a link between radio waves and cancer”. Participant 10 was interested in whether the document talks about brain

cancer and mobile phones. He/she said, “So, I am seeing that this one talks about mobile phones and cancer but it does not say brain cancer in specific”. He/she scanned the rest of the document very quickly and judged it as not relevant. The conclusion of that this participant made is correct since the document does not talk in specific about brain cancer. However, it talks about cancer in general. Participant 11 also said, “It does not actually talks about brain cancer”. He/she observed, “There is no evidence”. The participant was looking for information about brain cancer and its relationship to mobile phones.

Causes of Relevant Judgments

- Presence of Specific Evidence. Participant 3 said, “it seems relevant from the beginning”, He/she said that because of the following sentence, “*Health conditions which have been ascribed to the use of mobile phones include some types of cancer and changes in brain activity*”. Based on this sentence, he/she decided that the document is relevant. Also, Participant 8, after scanning the document said, “Okay, it is about the use of mobile phones”. He/she then said the document discusses, “health risks ... and types of brain cancer activity”. So, he/she judged the document as relevant.
- Presence of Keywords. Participant 9 here was first looking for the keyword “phone”. He/she located a number of occurrences of this word in the document. Then, he/she was looking for another keyword “cancer” and also found a number of occurrences. He/she read the first sentence that includes both keywords “phone” and “cancer” and was attempting to decide the documents relevance. He/she was not convinced by this sentence and scanned the remainder of the document; however, he/she said,

“We’ll say that’s relevant”. The presence of the keywords played an essential role on convincing the participant to judge this document as relevant even though he/she was not totally convinced about his/her decision.

Reflection

The document is graded as 1 by NIST assessor. Again, this document is difficult to assess by a secondary assessor. It does not talk directly about brain cancer. However, it does mention cancer in general, as well as changes in brain activities which makes it hard to judge and confusing for secondary assessors. Therefore, assessors who are searching specifically for the exact term “brain cancer” would not find it relevant. If the search topic therefore was not originated by the assessor, and the description of the search topic is not clear enough, he/she would not be able to decide easily whether the document is relevant.

5.5 Topic 383: Mental Illness Drugs

Document ID: NYT19990121.0380

Title	National Panel Reviewing Psychosis Studies
NIST Score	0
Length (#Words)	630
ProbRelevance	0.538
Has Key Words	Yes “Mental” , “Drugs” , “Illness”
No. of Judgments	5 Non-Relevant - 4 Relevant

Document’s Description

The document talks about the importance of stopping experiments that might harm people

or creating better regulations that mandate the stoppage of these types of experiments. There is no mention of a drug name that is used in the treatment of mental illness.

NIMH has conducted studies using such drugs as ketamine to induce psychotic symptoms in people with mental illness and healthy volunteers in an effort to better understand the biology of schizophrenia. The federal agency has also funded outside researchers for such work. Ketamine, a federally approved animal tranquilizer, is also a powerful hallucinogenic drug of abuse known as "Special K".

In New York, a current controversy centers around the death of a mentally ill patient given an experimental antipsychotic drug.

Figure 5.6: Attractive Sentences for DocID: NYT19990121.0380

Causes of Not Relevant Judgments

- **Absence of Specific Evidence.** Some participants here show good understanding of the search topic. For instance, Participant 3 was scanning the document to find drug names for treating mental illness. He/she found a drug name "Ketamine" but he/she said, "Mentions Ketamine but not necessarily a drug for mental illness. So, I don't think that is related. I don't see any specific drug names. Not relevant". Participant 5 also located the word "Ketamine", but did not consider it as a drug for treating mental illness. He/she said, "It does not talk about any drug".

Causes of Relevant Judgments

- **Trouble Understanding the Search Topic.** It seems that other participants here did not understand the search topic very well. There is a line in the document which mentions the drug name, "Ketamine". However, this drug is not used to treat mental illness. It is used to induce psychotic symptoms in people with mental illness. For example, Participant 4 was reading several lines in the document. He/she continued

to read random lines in the document and then observed a line which includes a drug name. The sentence is the following: “NIMH has conducted studies using such drugs as Ketamine to induce psychotic symptoms in people with mental illness”. He/she then stated, “I guess is it okay that Ketamine not to be used. Does that makes it relevant? Yeah, I guess it does. That makes it relevant. Giving Ketamin is really bad idea”. The Participant’s own interpretation of the texts in the document, without consulting the description of the search topic, caused him/her to judge the document incorrectly. The Participant was far off-base by assuming something that the description of the search topic did not ask him/her to do. He/she assumed that Ketamine was not to be used for treatment of mental illness.

- Trouble Understanding the Search Topic. Participant 8 thought that the mention of “antipsychotic drug” does make the document relevant. The topic’s description is, “Relevant documents will identify drugs used in the treatment of mental illness. In particular, a relevant document will include the name of a specific or generic type of drug. Generalities are not relevant”. The participant thought that inclusion of a phrase like “antipsychotic drug” would make the document relevant. However, this is not a drug name. He/she said, “It talks about antipsychotic drugs. I guess that’s generic type of drug. I think this is relevant”. The term “antipsychotic drug” is not a generic drug name. It is used here to describe any type of drug that is used to treat mental illness. Participant 10 judged the document as relevant for the same reason as well. He/she was reading random lines of the document. He/she thought at the beginning of the reading that this document is general. He/she said, So, it is talking

about drugs and psychosis, that seems too general though. He/she then said “so, it says mental illness and antipsychotic drugs. So, I am thinking that is relevant”.

Reflection

The document is graded as 0 by NIST assessor. It did not mention a specific drug for treating mental illness. The types of errors that assessors make here are easy to avoid if they concentrate while they judge, or if examples are given to them. This kind of document requires good understanding of the search topic. Moreover, assessors interpretation can lead to simple mistakes, as we have seen in the above examples.

Document ID: NYT19991214.0159

Title	Catch-22S FOR 22 Percent OF Americans
NIST Score	0 – This document was judged incorrectly by NIST Assessors
Length (#Words)	861
ProbRelevance	0.519
Has Key Words	Yes “Mental” , “Illness”
No. of Judgments	4 Non-Relevant - 5 Relevant

Document’s Description

The document talks about how it is important to discuss the issue of mental illness more seriously. There is mention of drugs used to treat mental illness, “Prozac” and “Zoloft”. This is the document that we discussed in our last full paper (Iiix 2014) where a NIST assessor was wrong in his/her judgment.

Causes of Not Relevant Judgments

- Absence of Specific Evidence. The name of drug is located within the document and mentioned in a single line in the middle of the document: *Prozac and Zoloft are laugh*

Prozac and Zoloft are laugh lines, even though such new drugs are effective in treating serious depression in many individuals. They're also knocked by many social commentators as a means to lull the non-ill into Stepford tranquillity.

Figure 5.7: Relevant Sentences for DocID: NYT19991214.0159

lines, even though such new drugs are effective in treating serious depression in many individuals. Participant 3 said, “This is talking about the problem of mental illness. But it does not seem to mention specific drugs. I don’t see any mention of drugs. It is not relevant”. Due to the length of the document, he/she omitted reading this line, and then judged the document as not relevant. Participant 8 also made the following observation: “I don’t think this is relevant. It does not have to do with drug. It has to do with stigma. Not relevant”. Another participant, Participant 9, said “I think this article is extremely generalities”.

Causes of Relevant Judgments

- Presence of Specific Evidence. The beginning the document was confusing to the participants. It talks about the topic of the liver in the first several lines and then begins to discuss mental illness. Some participants were carefully scanning the document looking for specific drug names which they found in the middle of the document. For example, Participant 2 said “Prozac and Zoloft are laugh lines even though, are effective in treating depression. Okay. It will, they do. At least they talk about them”. Participant 5 also said, “It talks again about depression. Yes. It talks about

Prozac and Zoloft. Uhm. Being effective in treating serious depression. So I'm gonna say this is somewhat relevant".

Reflection

The document is graded as 0 by NIST assessor. However, it is judged incorrectly by the NIST Assessor. The document did include drug names, however, these were stated in the middle. Assessors sometimes do not have the willingness and perseverance to read a very long document and check it line by line. They simply scan it very quickly for relevant material. They may just choose to consider the beginning, and if they feel it is not related to the search topic, they judge it as not relevant.

Document ID: NYT19991206.0109

Title	Researchers are Searching for and Treating Early Signs of Schizophrenia
NIST Score	2
Length (#Words)	1779
ProbRelevance	0.6
Has Key Words	Yes "Illness", "Mental"
No. of Judgments	1 Non-Relevant - 8 Relevant

Document's Description

The document talks about the early intervention for treating schizophrenia. It is a very long document, it is composed of 1779 words. There is mention of drugs used to treat mental illness, "Risperdal" , "Zyprexa" near the last third of the document.

Causes of Not Relevant Judgments

- Absence of Specific Evidence The drugs names were mentioned at the end of the document and it is very long. Not checking the document carefully line by line, will

Since the study began in 1996, he added, 4 of the 31 subjects who for six months received low doses of the anti-psychotic drug Risperdal and a specially tailored form of psychotherapy developed psychosis in the six months after they were taken off the drug.

In contrast to the participants in the Australian trial, neither subjects nor researchers in the Yale study know which participants are on medication in this case the anti-psychotic drug Zyprexa and which are getting dummy pills.

Treating people with drugs for a condition they do not yet have is a highly unusual approach in psychiatry, though it has precedent in other areas of medicine, the trial of tamoxifen as a prophylactic treatment for women at high risk for breast cancer is one example, and is being explored for Alzheimer's and some other diseases.

Many critics say they are concerned about the possible stigma attached to being labeled at high risk for psychosis, and about the potential side effects of even the newest anti-psychotic medications.

Figure 5.8: Relevant Sentences for DocID: NYT19991206.0109

result in locating no relevant material in these types of documents. Participant 3 for example, was only scanning it for drug names. He/she said, “It does not sound like it is related to drug treatments. Have a quick look. Hmm, mentions drugs in general but not any specific drugs. So, I am going to say not relevant”. The participant here did not pay a more careful look at the document.

Causes of Relevant Judgments

- **Trouble Understanding the Search Topic.** Even though Participant 2 judged the document as relevant, he/she entirely based his/her judgment on incorrect, inferential information. He/she was reading, “Treating people with drugs for a condition they do not yet have is highly unusual approach in psychiatry, though it has precedent in other areas of medicine- the trial of tamoxifen as prophylactic treatment for women at high risk for breast cancer is one example and is being explored for Alzheimer’s and some other diseases” and thought “Tamoxifen” is the drug that he/she was looking for. However, Tamoxifen is not used for treating mental illness. It is simply

mentioned in the text as an example of treating individuals with drugs for a condition they do not yet have. Therefore, this is an example of incorrect interpretation of the search topic's description.

- **Trouble Understanding the Search Topic.** Participant 8 here did not seem to understand the search topic correctly, though he/she judged the document accurately. He/she thought that mention of “antipsychotic drug” does make the document relevant. The topic's description is, “Relevant documents will identify drugs used in the treatment of mental illness. In particular, a relevant document will include the name of a specific or generic type of drug. Generalities are not relevant”. However, he/she thought that presence of terms such as “anti-psychotic drug” do make the document relevant. However, this is neither a specific nor a generic drug name. He/she said, “So, Ahh, Early signs of schizophrenia. To predict. Uhm, that. Newest anti-psychotic drugs. Because there's generic drug and their side effects. So relevant”.
- **Presence of Specific Evidence.** Participant 5 was not convinced that the document is relevant though he/she judged it as relevant. He/she thought that the document talks more about prevention, not about treatment of mental illness. However, he/she noted a drug name of an anti-psychotic drug “Zyprexa” at the end of the document. He/she therefore decided to judge it as relevant. He/she said “So the Zyprexa. It's you, What is this about. So I'm just tracking back because it's, Ah, This, Ah, A psychotic drug. Okay. So here at the end it does, Ah, Mention about a drug that is used in, Signed contracts that participants in the, Neither subjects. Hmm. Tough, To judge. It's mainly about, Uhm. A preventive but then it talks about this,

Ahh. Antipsychotic drugs at the end and the studies, So I'm gonna say relevant just because of that but not, Not much".

Reflection

The document is graded as 2 by NIST assessor. Drug names were not mentioned until the end of the document. This kind of documents requires careful processing of the content and a line by line check for the relevant material (drug names). Assessors sometimes do not have the willingness and perseverance to read a very long document and check it carefully. They just scan it very quickly.

5.6 Topic 436: Railway Accidents

Document ID: APW19990914.0022

Title	Germany to Replace Rail Chief
NIST Score	0
Length (#Words)	307
ProbRelevance	0.474
Has Key Words	Yes "Railway" , "Accidents"
No. of Judgments	2 Non-Relevant - 7 Relevant

Document's Description

The document is about Replacing the Rail Chief in Germany. It is focuses on how workers and others are not happy with him/her because of his/her way of handling things in the German Railway system. There is mention of a crash that has killed 101 people, but the accident is not descried in detail.

Morale at the railway has been spiraling since June 1998 crash of a high-speed train that killed 101 people.

Railway workers complained about Ludewig's handling of the aftermath of the crash, blamed on a faulty wheel, saying he did not make clear that it was an equipment problem.

Figure 5.9: Attractive Sentences for DocID: APW19990914.0022

Causes of Not Relevant Judgments

- **Absence of Specific Evidence.** Participant 2 said “Germany to replace rail chief. Okay. Accidents. Series of accidents. Okay. Did they go into detail. No. That doesn’t look like that they’re really talking about the accidents”. This participant started by looking for the term “Accidents” in the document. He/she located some but that did not convince him/her to judge it as relevant since the accident was not described. Participant 9 also scanned the document and found nothing that meets the criteria. He/she said “That’s not relevant. Although it does have accident stuff on it”.

Causes of Relevant Judgments

- **Presence of Specific Evidence.** Participant 4 said “It’s just barely relevant. It doesn’t really provide me a lot of information. I think it’s just barely fits the thing, the description that we’re looking for”. Participant 5 also noted information on a crash that killed 101 people in 1998. He/she said, “This is might be slightly relevant. So, I am just going to say it is relevant”. We notice from what Participant 5 said that

he/she did not find the document to be very relevant and that is due to insufficient information about the accident even though he/she judge it as relevant.

Reflection

The document is graded as 0 by NIST assessor. It is in mainly about Replacing the Railways Chief in Germany. However, it only mentions a train crash in one line of the document, and no additional details or description are given. Reference to a train accident in which people were killed, did attract assessors to consider it as relevant, though the accident itself was not described, as requested in the search topic's description. Participants differ in their understanding of the search topic. Some will focus only on one part of the description of the search topic and ignore the other parts while others will take into account all what is mentioned there.

Document ID: XIE19980303.0229

Title	Traffic in Three Italian Railway Stations Suspended by Fire
NIST Score	0
Length (#Words)	81
ProbRelevance	0.5
Has Key Words	Yes "Accidents" , "Railway"
No. of Judgments	2 Non-Relevant - 7 Relevant

Document's Description

This document is short and consisted of only a couple of sentences. It includes mention of people being injured in an accident in February, where six train accidents had been reported. However, the beginning of the document talks about how a fire had stopped

traffic in three small Italian railway stations, and one individual was injured because of the fire.

The fires are the latest in a series of accidents on Italian railways.
In February alone, six train accidents were reported, with more than 20 people injured.

Figure 5.10: Attractive Sentences for DocID: XIE19980303.0229

Causes of Not Relevant Judgments

- Lack of Topicality. Participants 5 believed that the document is about a fire and therefore considered it to be not relevant. He/she said, “This is about fire. So it’s not relevant. Not relevant”.
- Not Sufficient Information. Participant 4 thought there was not enough information in the document and that it is not about a train accident. He/she said “Traffic in three Italian railway stations suspended by fire. Well, it’s not a train. Not relevant. Not enough information, five sentences”.

Causes of Relevant Judgments

- Presence of Specific Evidence. For example, Participant 2 said, “Data on railway accidents. Fire stopped three railway stations. Nobody was injured. Fires in latest series. Where either the railroad system or the vehicle or pedestrian caused the accidents. Uhm. Does it really say who the cause. Okay. Is fire is once a system. Does there any? Okay. Of any sort. So probably... It’s relevant”. However, the document

did not describe the accident. Also, Participant 12 thought that the document described a specific railway accident. What is mentioned in the document however, is general information. Nothing specific is described. The participant here was focusing on the following sentence while making a judgment, “In February alone, six train accidents were reported. With more than 20 people injured”. However, this is a generality and not about describing a specific accident. Also, even though Participant 9 knew that the document is discussing a Fire in a railway station, the participant thought that might be relevant. He/she was reading “Fire stopped traffic in three small railway stations ...”. Therefore, he/she considered the fire which is described to be a type of accident.

Reflection

The document is graded as 0 by NIST assessor. It is about Italian railway stations being suspended by fire. In the last line of the document, there is a brief reference to six train accidents but no description of these accidents is given “In February alone, six train accidents were reported, with more than 20 people injured”. Considering whether the fire may be viewed as a type of accident is what accounts for differences in the assessors judgments.

Document ID: NYT19991206.0299

Title	Kenyan Government to Re-train Railway Workers
NIST Score	1
Length (#Words)	244
ProbRelevance	0.6
Has Key Words	Yes “Railway”, “Accidents”
No. of Judgments	1 Non-Relevant - 8 Relevant

Document's Description

This document is about the intention of the Kenyan government to retrain their railway workers. There is mention of casualties, "Train explosion that killed 27 and injured 30 others". Reference is also made to other train accidents in which 13 people were killed and 30 were injured. Description of the accident is also included but briefly.

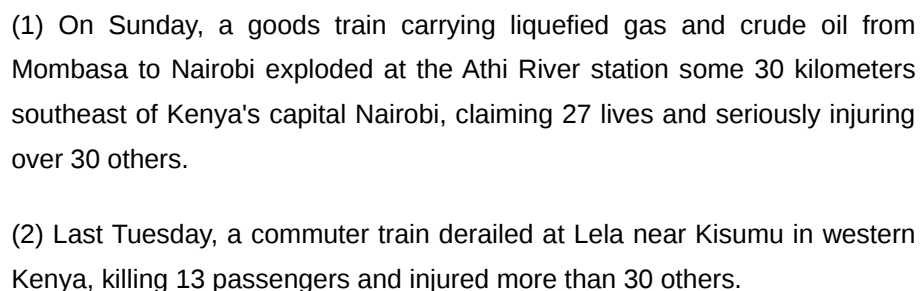
- 
- (1) On Sunday, a goods train carrying liquefied gas and crude oil from Mombasa to Nairobi exploded at the Athi River station some 30 kilometers southeast of Kenya's capital Nairobi, claiming 27 lives and seriously injuring over 30 others.
 - (2) Last Tuesday, a commuter train derailed at Lela near Kisumu in western Kenya, killing 13 passengers and injured more than 30 others.

Figure 5.11: Possible Relevant Sentences for DocID: NYT19991206.0299

Causes of Not Relevant Judgments

- Absence of Specific Evidence. Participant 4 said, "involved caused the accident. I mean it's kind of implying that it did because they need to re-train the railway workers. I'm assuming the rail workers screwed something up or could have done something but it's just not clear as to that. So I'd say that this point, although they identify accidents, it's not relevant. Uhm, since I don't know what's the cause of those accidents were". Even though the participants could locate a number of train accidents in the document, the cause of those accidents were not clear.

Causes of Relevant Judgments

- Presence of Specific Evidence. Participant 2 was reading random lines of the document and then he/she thought that it is relevant. He/she said, “So trains carrying liquefied crude oil from Mombasa exploded. Derailed. Killing. Okay. It looks pretty detailed”. Also, Participant 3 said, “This is more about training. But it does mention an actual accident, two actual accidents. So I guess that makes it relevant”. Locating a number of accidents was enough to judge this document as relevant to Participant 3. He/she did not look for the causes of the accidents. Therefore, this is a good example of how the behavior of secondary assessors differ. In fact, the behavior of the same assessor will differ during the the same judgment process.

Reflection

The document is graded as 1 by NIST assessor. It may be a source of confusion for assessors because the title gives a sense that it may be not relevant “Kenyan Government to Retrain Railways Workers”. There is mention of causalities, “Train explosion that killed 27 and injured 30 others”. There is also reference to other train accidents in which 13 people were killed and 30 were injured; the accident also is described. This type of document requires not only good understanding of the search topic but also careful processing of the information in the document.

Chapter 6

High Consensus Documents

The previous chapter considered analysis of the low consensus documents. In this chapter, on the other hand, we study the high consensus documents. The organization of the chapter is similar to that of Chapter 5, with the exception that two separate sections are included, one for each group of documents: *Relevant Documents and Non-Relevant Documents*. Aside from this difference, the structure of each section is similar to Chapter 5, where we start with a description of the document, follow with Causes of Relevant Judgments and/or not relevant to the search topic, and conclude by reflecting on the nature of the documents and participants' experiences with each one.

6.1 Relevant Documents

6.1.1 Summary

Black Bear Attacks (Topic 336)

The documents here do fit the criteria of relevance. However, one of the documents for this search topic discusses the number of black bear attacks (frequency) and also lists suggestions that should be followed in order to control aggressive behavior of black bears. The other two documents might appear to be more difficult to judge since they only focus on a single case, which to some assessors is not sufficient to qualify them as relevant documents. Those documents are judged 1 by the NIST assessor and do not discuss the frequency of the attacks.

Radio Waves and Brain Cancer (Topic 310)

This topic includes some technical terminology that might affect the behavior of secondary assessors. One should be familiar with this terminology in order to judge a document correctly. High consensus documents for this topic contain information on radio waves (cell phones) and brain cancer, brain activities, or alterations in brain cells. However, when the topic of brain cancer is not clearly mentioned, assessors tend to be reluctant to judge the document as relevant.

Mental Illness Drugs (Topic 383)

Locating relevant material in high consensus documents is not difficult for topics such as Mental Illness Drugs. The drug names are clear to identify and they sometimes appear even in the title of a document. The title therefore, occasionally gives strong indication of the

relevance of a document. Even if not stated in the title, drug names can readily be located in the document. Some issues that assessors might encounter when judging this type of document include lack of sufficient understanding of the search topic and unknown or unfamiliar words and terms in the document. These issues cause an assessor to experience confusion, and therefore he/she will make unreliable relevance judgments. Assessors might think that they are making the right decisions, while in fact they are entirely incorrect.

Railway Accidents (Topic 436)

Relevant material in this type of search topic is usually identifiable on a more surface level. Therefore, assessors will not find it difficult to judge this type of document. However, assessors' own interpretations of the search topic might cause avoidable errors in relevance judgments. Sometimes, assessors make incorrect presumptions and subsequently incorrect judgments. However, these types of interpretations are rare and do not reflect a wide spectrum of assessors' relevance judgments. We believe if assessors are provided with adequate examples in the tutorial part of the task, this type of misinterpretation will be minimal.

6.1.2 Key Findings

Number of Documents	12
NIST Score	2 (8 Documents) – 1 (4 Documents)
Correct Relevance Judgments	Presence of Specific Evidence (Most Common) Presence of Specific Evidence (The Impact of the Title) Presence of Keywords Trouble Understanding the Search Topic Lack of Concentration Lack of Familiarity or Knowledge
Incorrect Relevance Judgments	Absence of Specific Evidence Lack of Familiarity or Knowledge Absence of Keywords Insufficient Information

6.1.3 Topic 336

Document ID: APW19990809.0179

Title	Boy Attacked by Bear at Scout Camp
NIST Score	2
Length (#Words)	199
ProbRelevance	0.875
Has Key Words	Yes “black bears” , “bears”
No. of Judgments	4 Non-Relevant - 5 Relevant

Document’s Description

This document discusses a bear attack on a boy at a Scout Camp. It mentions that a black bear attacked a boy while he was sleeping in his tent. The document describes the attack

and mentions information about what may have attracted the bear to the boy's tent, a probable cause of the attack. It states, "Candy wrappers were found in the tent".

A black bear attacked a teen-age boy Monday while he slept in a tent at a Boy Scout camp, seriously injuring him.
Candy wrappers were found in the tent, but officials did not know if that was related to the attack.

Figure 6.1: Relevant Sentences For DocID: APW19990809.0179

Causes of Not Relevant Judgments

- Absence of keywords/Lack of Concentration. Participant 2 in particular was focusing on keywords in the document. He/she searched for all possible relevant words he/she could consider while judging the document, in particular the keywords "cosmetic" and "food". When no information was found relevant to the keywords, he/she judged the document as not relevant. He/she said, "Okay. Boy attacked by a bear. So. Need to look for a why in the attack ", then he/she was reading part of the description of the search topic "as it relevant would include aforementioned causes. Okay". He/she continued to talk and said "Search. So food or cosmetics Okay, that wasn't the reason why he is attacked. Bite to his shoulder. Candy. Okay. Does not explain why he was attacked". In fact, this participant read the word "candy" and had not realized or noticed that it is a type of food.
- Absence of Specific Evidence. For instance, Participant 3 was searching for causes of the attack, and he/she was convinced after scanning the document that it is

not relevant. He/she said, “I do not see causes, so, I am going to say that is not relevant”. Participants 11 and 12 also judged the document as not relevant for the same reason. Participant 11 said, “It does not talk about the causes or frequency. It is just one incident. So, it is not relevant”. Participant 12 said, “Boy attacked by bear at scout camp. First line says black bear. So this one may be relevant based on that. Well, this article actually doesn’t seem to be relevant because it doesn’t discuss the frequency. It’s just talking about one isolated incident and doesn’t really talk about any causes of the uh, the attacks. And there’s no uh, there’s no scientific uh, speculation or anything like that. So, not relevant”. Here, it seems that the participant was somehow that the document is relevant since he/she found couple of keywords such as “attack” and “black bear”. However, after reading the document more, he/she decided to judge it as not relevant for not mentioning the frequency of attacks, cause of the attack, and the lack of mentioning of scientific speculation about the attacks as stated in the description of the search topic.

- Not Sufficient Information. Participant 4 was looking for the causes of the attack and after reading the document, he/she believed that the document is not relevant because it does not provide enough information. He/she said, “This is not relevant. It is just short, we are not getting enough information about anything”.

Causes of Relevant Judgments

- Presence of Specific Evidence (The Impact of the Title). Participant 5 judged the document as relevant after reading the title only. He/she said, “So boy attacked by bear. This even from the title seems relevant. Ahh. So I’m gonna say relevant”. He/she did not further read the document to check in detail whether or not it contains

the relevant material. Here, the title had a great impact on the judgment of the participant.

- Presence of Specific Evidence. Participant 8 believed it is relevant because it meets one of the criteria that is mentioned in the description of the search topic, which is about mentioning the cause of the attack. He/she said, “ This talks about a black bear attack. Candy wrappers were found in the tent, but officials did not know if it was related. So, Yes ... from the scientific community. This is relevant”. Participant 10 believed also that the document is relevant because it discusses the cause of the attack. He/she said, “And the possible causes of the behavior. So, it gives the cause of the behavior, even though it does not discuss the frequency world wide but that is one of the criteria. So, I would say yes”.

Reflection

The document is graded as 2 by NIST assessor. It discusses one incident of black bear attacks. The entire document is devoted to discussing this. Though there is mention of a possible black bear attack (candy wrappers), some participants as we noted above, did not observe this. The candy wrappers which are associated with a type of food product, are the possible reason for the attack, which the scientific community is speculating on in the document. Therefore, participants who judged the document as not relevant were not able to infer that candy wrappers are a type of food associated item that attracts black bears and causes them to attack people. The ability of inference in this type of document is critical to finding the relevant material.

Title	Animal Attacks in North Woods Frighten Nature Lovers
NIST Score	1
Length (#Words)	736
ProbRelevance	0.864
Has Key Words	Yes "black bear"
No. of Judgments	1 Non-Relevant - 8 Relevant

Document ID: NYT20000706.0242

Document's Description

This document talks about how people now are being frightened by animal attacks (specially bears) in North Woods. It is graded as 1 by NIST assessor. It addresses the cause of one of the attacks. This cause is speculated by a biologist with the Canadian Wildlife Service. He said, "Her (the athlete) running motion my have frightened him, He (refers to the bear) may have attacked in what he considered self-defense".

Mary Beth Miller, a 24-year-old athlete from Yellowknife, Northwest Territories, was on a biathlon training run near the Valcartier military base when she was attacked by what officials deduce was a 200-pound male black bear.

"She defended herself and tried to escape" said Dr. Yvan Turmel, the coroner who performed the autopsy. "But the bear attacked and attacked. She had no chance of getting away alive, absolutely no chance".

"Her running motion may have frightened him" said Dick Russell, a biologist with the Canadian Wildlife Service. "He may have attacked in what he considered self-defense".

Figure 6.2: Possible Relevant Sentences For DocID: NYT20000706.0242

Causes of Not Relevant Judgments

- Absence of keywords. Participant 2 again was focusing on keywords in the document. He/she searched for all possible relevant words he/she could think of when judging the document, in particular “cosmetic” and “food”. When nothing about these keywords was found, he/she judged the document as not relevant. He/she said “Cause. Food. Let’s see cosmetics. Excellent. Attacked specially some fear. Campers. Okay. There’s a number of black. There is an inhabitant but it doesn’t talk about the cause. Okay. I don’t see anything. Just waiting the cause”.

Causes of Relevant Judgments

- Presence of Keywords. Participant 9 was searching for the word “attack” in the document. He/she found several instances of this word, and read the texts that contain the word. He/she then judged the document as relevant. Though the participant judged the document as relevant, it is apparent while watching the video recording of this participant, that he/she was not particularly interested in finding the relevant material.
- Presence of Specific Evidence. Participant 4 was certain that the document is relevant. He/she said, “I believe this is relevant because you know we are talking about a number of different cases. It talks about the last case occurred in Quebec which gives us an idea to the incidents. It seems we have lots of statistics from wildlife officials. So, I would say that this is certainly a relevant document”. Participant 10 also judged the document as relevant because he/she was convinced after reading the document that it discusses possible causes of black bears attacks. He/she said, “So there was

a bear attack. And a black bear. So it's talking of bear attack. So I'm looking if talks about, Uhm, any reasons. They think maybe the bear was startled. Uhm. So it's talking about, I guess the possible causes. So I'd say it's relevant". Additionally, Participant 11 said, "Speculation by scientific community" and he/she judged the document as relevant. Participant 12 also judged the document as relevant, but from the transcribed data and the video recording we determined that he/she just was guessing while making this judgment "Lethal attacks. I guess, in Canada and US, the 20th century. So, I guess that would make this a relevant document". Here, Participant 12, though he/she did not find a cause, he/she determined the document to talk about "lethal attacks" in U.S. and Canada, which he/she thought it fits the criteria in the description of the search topic which states, "A relevant document would discuss the frequency of vicious black bear attacks worldwide and the possible causes for this savage behavior".

Reflection

The document is graded as 1 by NIST assessor. Again, this document is mainly focusing on one incident, though there is brief mention of other attacks. The cause of the attack that is the document focuses on, is speculated on by the scientific community. These types of documents are difficult in which to find relevant material if you try to search just for keywords without reading and making inferences about what has been read. For example, Participant 2 judged it as not relevant because he/she did not find what he/she was searching for, such as the words, "food" and "cosmetic".

Title	Bears Become Uncomfortably Close Neighbors in Northern New Jersey
NIST Score	2
Length (#Words)	1979
ProbRelevance	0.846
Has Key Words	Yes
	“bear attack”, “black bear”, “attacked” , “bear”
No. of Judgments	2 Non-Relevant - 7 Relevant

Document ID: NYT20000602.0371

Document’s Description

This document is about how residents of northern New Jersey are currently not comfortable with seeing bears near their homes. It is a very long document which contains several stories about black bear invading peoples’ homes and backyards. It also mentions speculation from the scientific community about what makes the homes an attractive place for bears. Additionally, the end of the document includes a number of suggestions for dealing with bears near homes and camps. These tips could be used to help with the control or management of aggressive black bears.

Causes of Not Relevant Judgments

- Absence of Specific Evidence. Participant 5 said, “So, I am going to say not relevant because it is just about scaring people. They are not actual attacks”. Here, the participant was only focusing on finding the causes of the attack. He/she did not consider the other information in the description of the search topic. Participant 2 also judged the document not relevant for the same reason. He/she said, “Bears become uncomfortably close neighbors in Northern Jersey. Okay. What is this? Are there where attack involved. Although none of the complains involved bears

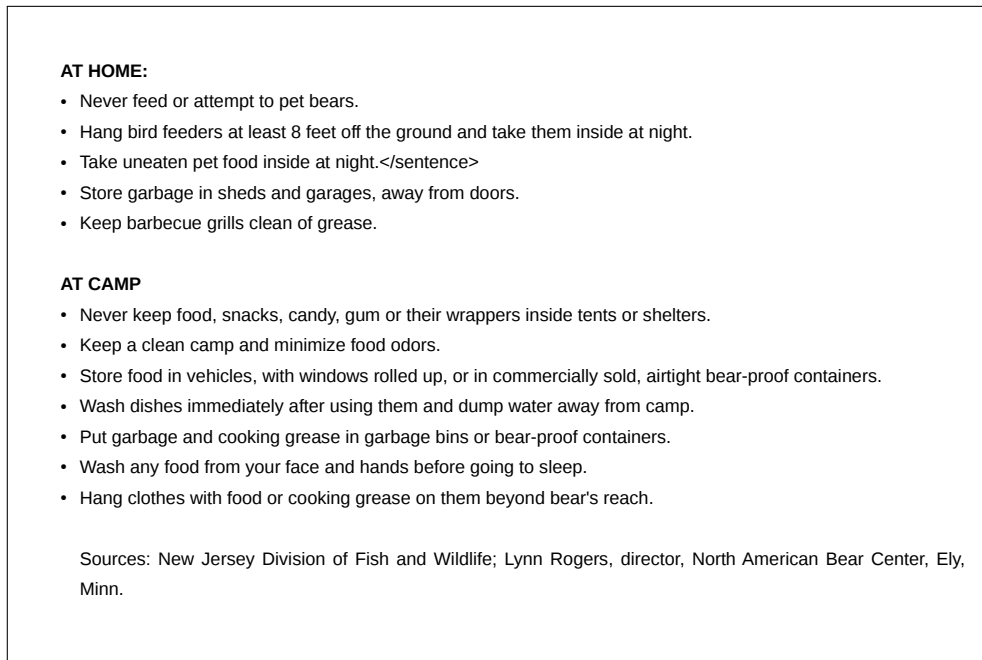


Figure 6.3: Relevant Sentences For DocID: NYT20000602.0371

attacking or injuring people. Okay. extremely rare. Tips for dealing with them. Okay. But a relevant and detail step. No. Causes. Ah! No. Not relevant”.

Causes of Relevant Judgments

- Trouble Understanding the Search Topic. Participant 12 for example said, “Although none complaints involved the bears attacking or injuring people. So, I guess that is sort of reporting the frequency as zero. So, guess this is relevant”. Here, the participant made a judgment based on his/her own understanding of what is mentioned in the search topic. He/she assumed that since there are no complaints, this means that the frequency of attacks is zero. Regardless of the correctness of the partici-

pants judgment, this kind of understanding shows that he/she did not comprehend the search topic very well.

- Presence of Specific Evidence. Participant 9 here judged the document as relevant simply because he/she found information on an expert who discusses black bear attacks. He/she said, “Here is the expert, Lyn, whatever her name was, Okay, Could harms neighbors. Relevant”.

Reflection

The document is graded as 2 by NIST assessor. It might be slightly confusing for assessors because it does not mention black bear attacks as we expect they may occur, with injuring or killing humans. However, it deals more with bears invading peoples’ homes and backyards. Assessors who search for information on actual incidents of black bear attacks will not find it relevant. However, this document talks about what attracts the bears to invade peoples’ houses, and also how to manage/control their savage behavior, including steps to deal with them. This is the reason that secondary assessors sometimes simply focus on one part of the description and overlook the others, since they are not the originators of the search topic. They merely follow the guidelines in the description of the search topic.

6.1.4 Topic 310

Document ID: APW20000608.0153

Document’s Description

The document is about Cell Phones and how the emissions which they produce are not safe and could harm users. There are lines in the document that make it relevant, and

Title	FDA To Oversee Cell Phone Safety
NIST Score	2
Length (#Words)	569
ProbRelevance	0.853
Has Key Words	Yes "brain" , "cancers"
No. of Judgments	5 Non-Relevant - 4 Relevant

these are: "A few animal studies have suggested that cell phones' low-level radiation could accelerate cancer growth, and some research suggests it also causes subtle alterations in signals from brain cells". However, this passage does not discuss brain cancer in particular, but rather points out how low-level radiation emitted from cell phones could accelerate the spread of cancer.

A few animal studies have suggested that cell phones' low-level radiation could accelerate cancer growth, and some research suggests it also causes subtle alterations in signals from brain cells.

Figure 6.4: Possible Relevant Sentences For DocID: APW20000608.0153

Causes of Not Relevant Judgments

- **Absence of Specific Evidence.** Participant 5 was reading various lines of the document looking for a link between car phones and brain cancer but did not find anything regarding this. He/she said, "This does not provide link but it is kind of ... Just raises the issue. No proof that cell phones are totally risk-free but there is also no connection ... Relevant but it does not look like. So, it is not relevant". It is clear from this part of the transcription that the participant was very challenged in finding a link between cell phones and brain cancer. Participant 11 said, "Talks

about the setup of the study and money but it does not really say results. So, it is not relevant". Participant 10 also said, "It is not talking about brain cancer. It talks about changes in brain cells. So, I would say it is not relevant". Participant 12 skimmed the document quickly and commented, "Well, this article does not seem to be relevant because it does not really talk about any studies, and it does not talk about the incidents of cancer being higher, lower, the same for people live near the towers or use phones. So, I don't think this is relevant".

- Unknown Reason. No clear reason was given. Participant 4 while judging this document, did not give any specific reason.

Causes of Relevant Judgments

- Presence of Keywords. Participant 2 was looking for specific keywords such as "brain" and "brain cancer". Once he/she located the keywords, he/she was reading the lines that contain them. The participant judged the document as relevant because he/she believed that it mentions animal studies and this makes it fit the criteria. Participant 9 also was searching for the word "cancer" and when found it said, "I guess it is relevant".
- Presence of Specific Evidence. Participant 3 said, "So, it is so far ... this does not mention any correlation. It mentions some studies saying that there is no connection ... So, this does seem relevant". Participant 8 also said, "Cell phone safety. Cause cancer and other problems. Radio waves. Biological effects ... This is relevant". He/she here believed that the document is relevant since it mentions information about radio waves and cancer. However, in spite that he/she judged the document correctly, he/she was not able to identify the relevant material in the document.

He/she judged the document as relevant because it does mention cancer and radio waves, as well as cell phone safety in general.

Reflection

The document is graded as 2 by NIST assessor. However, participants in our study are secondary assessors, and this means they did not create the search topic. The information which is needed is given to them in a description. This type of search topic is difficult for secondary assessors since it requires that inference occasionally be made about the information, in order to decide whether the document is relevant. If we consider the above reasons a document may be considered relevant, we notice that some assessors experience difficulty in finding a link between cell phones and brain cancer (specific evidence) in the document.

Document ID: NYT19991003.0452

Title	Mixed Signals on Cell Phones and Cancer
NIST Score	1
Length (#Words)	1143
ProbRelevance	0.909
Has Key Words	Yes
	“brain”, “cancer” , “cell phones”
No. of Judgments	2 Non-Relevant - 7 Relevant

Document's Description

This document is about the effort made in finding a connection between cancer and the use of cell phones. A number of scientists and researchers were cited in the article talking about this. At the end of document, mention is made of the statistically significant increased risk of rare human brain tumours in patients who used cell phones.

In studies by Muscat and colleagues at the American Health Foundation in New York this year, there was a statistically significant increased risk of rare human brain tumors, known as neuroepithelial tumors, in patients who used cell phones.

Carlo acknowledged that there was no link between this rare form of cancer and the frequency and duration of cell-phone calls.

Also, researchers found no association between overall brain cancer risk and cell phone use.

Figure 6.5: Possible Relevant Sentences For DocID: NYT19991003.0452

Causes of Not Relevant Judgments

- Lack of knowledge or Familiarity. Participant 5's lack of familiarity with key terms that are related to the search topic caused him/her to judge the document as not relevant. He/she said, "So, phone, wireless phone radiation causes genetic damage. This is not radio waves though. So, nothing about radio waves. So, I am going to say it is not relevant".
- Absence of Specific Evidence. Participant 11 said, "It does not report being lower, it just reports what the person who did the study says. I don't think it is relevant".

Causes of Relevant Judgments

- Presence of Specific Evidence. Participant 3 was convinced that the document is relevant because it is about cell phones and cancer in general. He/she said, "So the title seems relevant. So, this is relevant to cell phones and cancer and it mentions brain cancer specifically. So, that is very relevant". Participant 4 also judged the document as relevant because it meets the criteria mentioned in the description of the search topic. He/she said, "I am not seeing anything here about studies or results of

studies. He defends his view about the possible biological hazards of cell phones by describing a handful of studies. Okay, here we go. This is summarizing a whole bunch of different studies. So, yes, at the end it became relevant. Glad I read it till the end”. Participant 12 also judged the document as relevant after reading the last part of the document, similar to Participant 4. Participant 10 also judged the document as relevant for the same reason after reading the following line, “rare human brain tumours and they ... it says rare form of cancer. So, based on that, I would say that is relevant”.

- Presence of Keywords. Participant 9 was searching for the keyword “cancer” and was reading the lines that contains this word. The following line includes this keyword: “Even Joshua Muscat, a New York epidemiologist whose brain-cancer studies have been cited by Carlo as raising public health concerns, think Carlo’s warnings are extreme”.

Reflection

The document is graded as 1 by NIST assessor. One of the problems with this type of document is that the relevant material is not explicitly stated or mentioned only in part of it. This makes the assessment process difficult; good knowledge of the search topic and a patient attitude play an important role in finding/discovering the relevant material. For example, one of the participants (Participant 5) was struggling to judge this document only because he/she lacked the knowledge about radiation in general and whether radio waves are different or similar to radiation.

Title	Cell Phones: Questions but No Answers
NIST Score	1
Length (#Words)	1059
ProbRelevance	0.93
Has Key Words	Yes
	“brain”, “cancer”
No. of Judgments	3 Non-Relevant - 6 Relevant

Document ID: NYT19991025.0333

Document’s Description

The document is about the safety of using cell phones and the potential health risks associated with them. In the middle of the document, mention of a study that has determined some association between the use of cell phone of one rare form of brain cancer is made. It states, “A hospital study that compared brain cancer patients and a similar group without brain cancer found no statistically significant association between cell phone use and a group of brain cancers known as glioma”.

A hospital study that compared brain cancer patients and a similar group without brain cancer found no statistically significant association between cell-phone use and a group of brain cancers known as glioma.

Figure 6.6: Possible Relevant Sentences For DocID: NYT19991025.0333

Causes of Not Relevant Judgments

- Absence of Keywords. Participant 10 was looking for keywords “animal”, “cancer” and then “brain cancer”. He/she said, “So, I am seeing that this one talks about mobile phones and cancer. It does not say brain cancer in specific”. When he/she

found no information on this, he/she decided it not to be relevant. The participant merely read the first several lines of the document and started to look for keywords in the document after this.

- **Lack of Familiarity or Knowledge.** Participant 12 here skimmed the document looking for the relevant material, but after reaching the end of it, he/she commented, “It does not really talk about radio waves though. So, I guess ... I will say not relevant. Although, I am not really sure”.
- **Absence of Specific Evidence.** Participant 5 was convinced after skimming the document that it is not relevant. He/she said, “It does not provide any link between the radio waves and brain cancer. It is just about how much we do not know. So, just going to skim through it but ... I am going to go with not relevant because it does not provide any links. Questions with no answers, like the article says”. Participant 10 judged it as not relevant as well. He/she said, “So, I am seeing this one talks about mobile phones and cancer. It does not say brain cancer in specific”.

Causes of Relevant Judgments

- **Presence of Specific Evidence.** Participant 3 said, “Alright, so it seems like very specifically about brain cancer and cell phones so we could say it is relevant right of the bat”. The participant here was reading the sentences, “Almost since there have been cellular phones, there have been worries that the radio waves they emit might cause brain cancer. Yet despite years of studies, no one has established a solid link, and the industry has long sought to reassure the public that the technology is perfectly safe.”. He/she judged the document as relevant after reading these two sentences. Participant 4 also judged it as relevant, but was struggling to reach a deci-

sion. He/she said, “This document is relevant because it includes an experiment that report ... you know ... I think that just barely fits the thing”. Finally, Participant 11 found the document to be relevant because it talks about results of a hospital study. He/she said, ”It talks about the results of a hospital study. That is relevant”.

- Lack of Concentration. Even though participant 9 judged the document as relevant, he/she was talking about something off-topic, not related to the search topic.

Reflection

The document is graded as 1 by NIST assessor. It requires good ability to identify the relevant material. Also, the amount of detail mentioned in the description makes the process of assessment not an easy one. There are some details about what makes the document relevant that assessors need to focus on while looking for relevant information.

6.1.5 Topic 383

Document ID: NYT20000925.0105

Title	Are We Overmedicating our Kids?
NIST Score	2
Length (#Words)	1384
ProbRelevance	0.808
Has Key Words	Yes
	“Mental” , “Drugs”, “Illness”
No. of Judgments	1 Non-Relevant - 8 Relevant

Document’s Description

This document is about how we over-medicate our children, and whether or not this is

right It focuses on mental illness and mentions several drugs that are used for treating mental illness, such as Ritalin, Prozac, and Zoloft.

Another factor underlying the rise in prescriptions, says Rushton, is the emergence of a new class of drugs called selective serotonin reuptake inhibitors, including Prozac, Zoloft and Paxil, which have fewer, less serious side effects than the previous generation of antidepressants. They have been increasingly prescribed over the last decade by pediatricians for treating depression, obsessive compulsive disorder and anxiety in children.

Figure 6.7: Relevant Sentences For DocID: NYT20000925.0105

Causes of Not Relevant Judgments

- Lack of Familiarity or Knowledge. Participant 2 here did not know whether Attention Deficit Hyperactivity Disorder (ADHD) is a form of mental illness. He/she said, “ADHD ... mental illness. I do not know. I am not sure”. At the end he/she judged the document as not relevant.

Causes of Relevant Judgments

- Lack of Familiarity or Knowledge. Though Participant 5 judged the document as relevant, he/she indicated a low level of certainty when making his/her judgment. His/her confusion about what defines mental illness, and if anxiety and ADHD are considered mental illness caused him/her to be unsure about the judgment he/she made. He/she said, “I am a bit confused about what is meant by mental illness? I do not think anxiety or some depression was considered. ADHD, again hyper activity

does not sound to me like it is an issue here. This maybe slightly relevant because it does mention drugs that are used in treatment but again I am not sure I get the link between ... Maybe my understanding of what mental illness is should be ...". Then, he/she judged it as relevant. It is clear from this text that the participant is completely hesitant and confused about what is meant by mental illness. Participant 10 shared the same confusion with Participant 5 about whether ADHD is a form of mental illness. He/she was confused at the beginning and continued to read until he/she reached the following line, "Mental and emotional health problems". Then, he/she said "since it mentions that, I guess I would say it is relevant".

- Presence of Specific Evidence. Participant 8 said, "Ritalin, ADD, yes, it is relevant". Participant 12 also skimmed the document looking for specific drug names and noted "Ritalin" and he/she said, "This article mentions Ritalin. So, this article is relevant". The other participants expressed similar reasons for judging the document as relevant.

Reflection

The document is graded as 2 by NIST assessor. There are several drug names in this document such as Ritalin, Prozac and Zoloft. Though 8 of the participants judged it as relevant, Participant 5 was merely guessing its relevance, since he/she was not certain whether ADHD is a type of mental illness. Only one participant, Participant 2, judged it as not relevant, not only because he/she was not able to find a specific drug name, but also because of the confusion he/she experienced about the condition of ADHD. He/she was not certain whether ADHD is considered a type of mental illness. Therefore, familiarity with the search topic and the related terms are helpful with judging the document correctly.

Document ID: NYT20000319.0216

Title	White House Seeks to Curb Use of Ritalin, Prozac, Other Such Drugs by Children
NIST Score	2
Length (#Words)	1070
ProbRelevance	0.882
Has Key Words	Yes “Mental” , “Illness”
No. of Judgments	1 Non-Relevant - 8 Relevant

Document’s Description

This document is about intention of the White House to stop use of drugs used in the treatment of mental illness, such as Ritalin, Prozac, and others for treatment of children. These kinds of drugs should not be used as the first option to treat psychiatric disorder in children. Reading the title of the document would play an important role in convincing assessors to judge it as relevant since a number of drug names are included.

Causes of Not Relevant Judgments

- Lack of Familiarity or Knowledge. Again, Participant 5 was confused whether depression or anxiety could be considered a form of mental illness. He/she said, “So, I am going to say not relevant. I do not think that was mental illness to me. Anxiety or depression do not count like mental illness. Just about regulating those drugs”.

Causes of Relevant Judgments

- Presence of Specific Evidence (The impact of the title). A number of participants decided that the document is relevant simply because the title contains specific drug names. For instance, Participant 2 judged the document as relevant only after reading

its title. He/she had not even read a single line of the document. He/she said, “White house seeks to curb use of Ritalin, Prozac, and other such drugs by children. That is specific. It is relevant”. Participant 3 did the same and was certain that the document is relevant. However, he/she read the first several lines in order to confirm his/her judgment. Participant 12 judged it as relevant only after reading the title. He/she said, “Well, it is right in the title. So, this article is relevant”.

- Presence of Specific Evidence. All of the remaining participants judged the document as relevant because they found specific drug names that are used to treat mental illness. Participant 11 for instance said, “Zoloft has been approved to treat obsessive-compulsive disorder. So, it has the name it says it will be used to treat the mental illness. So, it is relevant”.

Reflection

The document is graded as 2 by NIST assessor. It also is highly relevant since it mentions not only one drug name for treating mental illness but rather a number of them. The title gives a clear and strong indication about the document’s relevancy to the search topic, since it includes several drug names such as “Prozac” and “Ritalin”. Also, the entire body of the document focuses on these drugs and how parents should be aware of the impact they might have on their childrens health. All of the participants here did not find it difficult to judge it as relevant except Participant 5, who was confused about whether depression and anxiety are a form of mental illness.

Title	Prozac's Reign as Top Drug Ending
NIST Score	2
Length (#Words)	600
ProbRelevance	0.833
Has Key Words	Yes "Illness", "Mental"
No. of Judgments	0 Non-Relevant - 9 Relevant

Document ID: APW20000307.0001

Document's Description

The title of this document mentions the drug name of "Prozac", which is used in the treatment of mental illness. Therefore, this is a very good indication that the document is relevant. Moreover, the entire document discusses Prozac and that it is presently losing its value and reputation against other competitive drugs used in the treatment of mental illness, such as Zoloft and others.

None of the participants judged the document as not relevant

Causes of Relevant Judgments

- Presence of Specific Evidence (The impact of the title). Again, a number of participants decided that the document is relevant entirely because the title contains specific drug names. Participant 12 for instance said, "Prozac's Reign as top drug ending. Again, it is right in the title. So, it is relevant". Participant 2 indicated the same point and judged the document as relevant based on the title. He/she said, "Prozac's Regin as top Drug Ending. That is definitely a drug use". The participants here did not read the body of the document to confirm their initial judgments.

- Presence of Specific Evidence. The remaining participants judged the document as relevant because they found a specific drug name. For example, Participant 10 said, “Okay. Prozac. That is a specific drug name. So, I am just scanning to see if it mentions treatment for mental illness. So, depression is a mental illness. So, I would say then it is relevant”. However, though Participant 4 judged the document as relevant, he/she was not convinced and stated, “Kind of interesting that I don’t see mental health identified. And I think it implied because already know what Prozac is. An “antidepressant”, oh, I don’t know if there are other parts of the body that gets depressed. It’s telling me about the sales of the drug and telling me about sales invoice done very well but I guess it’s relevant”.

Reflection

The document is graded as 2 by NIST assessor. It did mention a drug name, “Prozac”, in its title, and the entire body of the document discusses it in greater detail. Though this document is directed more toward discussing the marketing challenges of this drug, it meets the criteria of the search topic by mentioning a specific drug name that is used to treat mental illness. All participants did not find any difficulty in judging the document as a relevant with one exception. Participant 4 was not certain whether an “antidepressant” is linked only to mental illness in the human body. However, he/she judged it as relevant at the end since he/she found mental illness drug names.

Title	Railway Engineers Sentenced for Major Accident
NIST Score	2
Length (#Words)	261
ProbRelevance	0.967
Has Key Words	Yes
	“Railway” , “Accidents”
No. of Judgments	0 Non-Relevant - 9 Relevant

6.1.6 Topic 436

Document ID: XIE19991207.0246

Document’s Description

The document is about two engineers who were charged for their role in a deadly railway accident in China. It discusses the accident in detail and also mentions the cause of it. It is stated, as illustrated in Figure 6.8.

The two failed to apply the brakes and reduce their high speed of 111 km per hour in time while passing a turnout near Hengyang, which allows for a speed of only 45 km per hour. This failure resulted in derailment of the train.

Figure 6.8: Relevant Sentences For DocID: XIE19991207.0246

None of the participants judged the document as not relevant

Causes of Relevant Judgments

- Presence of Specific Evidence. Participant 3 said, “So this describes a specific accident. So, that would be relevant”. Also, participant 4 judged it as relevant because he/she found the cause of the accident which is included in the criteria of the search topic. He/she said, “The two failed to apply the brakes. Alright, well. That is

relevant. That is easy”. The other participants expressed similar reasons for judging the document as relevant. Participant 2 also considered the document to be relevant, though he/she was not completely convinced that it is relevant. He/she said, “Kind of. Not very much information but it is okay”.

Reflection

The document is graded as 2 by NIST assessor. It is highly relevant since it describes an accident in which a number of people were killed. None of the participants judged the document incorrectly. The accuracy was 100%. The reason for the high accuracy rate is because the participants indicated that the entire document describes an accident in detail, and that the cause of the accident is clearly stated; Participant 2 did not refer to these reasons, though he/she judged it correctly.

Document ID: XIE19990802.0027

Title	Writethru: Toll In India’s Rail Accident Up to 250
NIST Score	2
Length (#Words)	565
ProbRelevance	1
Has Key Words	Yes “Accidents” , “Railway”
No. of Judgments	0 Non-Relevant - 9 Relevant

Document’s Description

This document is about a railway accident in India where the death toll reached up to 250. The title of the document gives a good indication regarding its relevancy. Moreover, the document itself describes the accident in great detail and also identifies the cause.

None of the participants judged the document as not relevant

At least 250 passengers were killed and over 460 injured early Monday morning in one of the worst rail accidents in India.

The tragedy occurred when a Delhi-bound mail train collided head-on with an east-bound express in the wee hours Monday in Gaisal of the Northern Siligury District of the West Bengal state.

A source in the Railway Ministry said the accident could be due to a human error while a railway police officer said it was caused by a signal failure.

The worst train accident before the latest happened in August 1995 when 302 people were killed as a Delhi-bound express train rammed into a stationary passenger train near Firozabad in North Uttar Pradesh.

PTI reported at the site that at least 14 compartments beside the engines of both trains bore the brunt of the collision and four compartments telescoped into each other.

The victims of the mail train were mostly army men and border security forces, according to PTI.

Earlier reports claimed that preliminary information suggested the accident was a case of a bomb blast, but after seeing the actual position of the engine and compartments of the trains, it seemed to be a case of head-on collision.

Figure 6.9: Relevant Sentences For DocID: XIE19990802.0027

Causes of Relevant Judgments

- Presence of Specific Evidence. For example, Participant 12 said, “Well, this 250 passengers were killed. Then, worst rail accident in India. That falls under the criteria. So, it is relevant”. Participant 8 also said, “This one talks about hundreds of people were killed in an accident. This one ... Okay ... This one is relevant for sure”. The other participants expressed similar reasons for judging the document as relevant.

Reflection

The document is graded as 2 by NIST assessor. Judging this document as relevant is not at all difficult, not only because the title gives a good indication about its relevancy, but also since most of the document discusses the accident in greater detail and identifies the

cause of the accident. Again, the relevant information can be identified with no significant effort.

Document ID: XIE19981020.0034

Title	29 Bodies in Egypt's Railway Accident Identified
NIST Score	1
Length (#Words)	597
ProbRelevance	0.943
Has Key Words	Yes
	"Railway", "Accidents"
No. of Judgments	0 Non-Relevant - 9 Relevant

Document's Description

This document's title gives a good indication about its relevancy. The document discusses one accident in detail and two other accidents in brief. It talks in detail about the first accident and describes it very clearly. The cause of the accident is also mentioned as shown in Figure 6.10.

The train, which was traveling south from Egyptian northern port city of Alexandria, derailed and crashed into a square in Kafr al Dawar, a town about 30 kilometers southeast of Alexandria at 5:10 p.m. local time (1510 GMT) Sunday.

The Interior Ministry said in a statement that the driver lost control of the train when he changed tracks at high speed because the brakes between the locomotive and the carriages were broken by some hitchhikers on the train.

Head of the Railway Authority Mahmoud Marei, who was also on the site, said the train's driver was unable to stop the train when he used the brakes, because the air tanks operating them were damaged.

Figure 6.10: Relevant Sentences For DocID: XIE19981020.0034

None of the participants judged the document as not relevant

Causes of Relevant Judgments

- Presence of Specific Evidence. Participant 4 was highly certain that this document is relevant. He/she said, “The driver lost control of the train when he changed tracks of high speed. Alright, super relevant”. Participant 9 also said “Here is hard data. yes, this is more hard data. This probably better than the other one. Okay ... I will say it is relevant”. The other participants expressed similar reasons for judging the document to be relevant.

Reflection

The document is graded as 1 by NIST assessor. An accident is mentioned directly in the title and most of the body is devoted to discussing this accident and describes it in more detail. All of the participants judged it correctly.

6.2 Non-Relevant Documents

6.2.1 Summary

Black Bear Attacks (Topic 336)

The documents under this search topic are not directly relevant to the topic, even if reference is made to black bears or attacks, as this is out of context and does not fit the criteria given in the description of the search topic. Lack of concentration might be a cause of error when judging documents of this type. Some assessors might find a reference to a bear attack, but it is not related to black bears specifically.

Radio Waves and Brain Cancer (Topic 310)

Again, lack of topicality is the theme that describes documents for this search topic. The documents do not contain information which links radio waves and brain cancer. Therefore, they are easy to judge as not relevant.

Mental Illness Drugs (Topic 383)

Documents under this search topic might include numerous instances of keywords such as ‘Mental’, ‘mental illness’, ‘drugs’. However, they do not include any drugs names for treating mental disorders. Therefore, assessors need to pay careful attention, to read them line by line in order to identify the relevant material, since a drug name might appear in just a single line.

Railway Accidents (Topic 436)

Documents here are not difficult to judge as not relevant since they do not contain any information on the topic of railway accidents. However, assessors’ incorrect interpretation

of the words and phrases might cause simple errors that can be avoided if assessors are provided with sufficient examples of criteria required for relevant document. However, these kind of errors are rare.

6.2.2 Key Findings

Number of Documents	12
NIST Score	All Documents were graded as '0'
Correct Relevance Judgments	Lack of Topicality (Most Common) Absence of Specific Evidence Absence of Keywords
Incorrect Relevance Judgments	Lack of Familiarity or Knowledge Lack of Concentration Trouble Understanding the Search Topic

6.2.3 Topic 336

Document ID: NYT19990927.0436

Title	For Sasquatch Believers There's no Turning Back
NIST Score	0
Length (#Words)	1375
ProbRelevance	0.222
Has Key Words	Yes
No. of Judgments	"black bears" , "Attacked" 9 Non-Relevant - 0 Relevant

Document's Description

The document discusses the possibility of the existence of the sasquash. It is not about

black bear attack whatsoever, though black bears are mentioned four times in the document; however, these occur in reference to unrelated situations.

Causes of Not Relevant Judgments

- **Lack of Topicality.** Participants consider that the document is off-topic. For instance, Participant 5 was searching for black bear attacks and he/she was convinced after scanning the document that it is not relevant because it does not talk about black bear attacks. He/she said, “But this doesn’t seem to be, Ahh, about bear attacks at all. So I’m gonna say not relevant”. Participant 10 also said, “I’m looking for black bear. This is talking about sasquatch. So I’m thinking that’s not related. No. That’s not related”. Participant 3 said also, “Sasquatch believers? That seems not related”. One participant in particular (Participant 9) was focusing on searching for keywords in the document. He/she searched for all possible relevant words he/she could think of in particular “bear” and “attack”. Though he/she identified a number of instances of these keywords in the document, he/she was not convinced that the document is relevant and said, “But we want bear attack. The UFO people aren’t the only ones who get attacked. Ahh. It’s like attacking their idea or arguments. I say it’s not relevant”.

None of the participants judged the document as relevant

Reflection

The document is graded as 0 by NIST assessor. Though these types of documents are not hard to judge, an assessor needs to be careful and read the entire document before he/she gives his/her final decision. None of the participants judged it as relevant.

Document ID: NYT20000718.0206

Title	Behind Conflict Over New Grizzly Program, an Endangered Species
NIST Score	0
Length (#Words)	2134
ProbRelevance	0.25
Has Key Words	Yes "bear"
No. of Judgments	8 Non-Relevant - 1 Relevant

Document's Description

This document is about the topic of grizzly bears, and it has nothing to do with black bear attacks. There is no mention of black bears or attacks in the document. Therefore, it is not at all related to the search topic.

Causes of Not Relevant Judgments

- Lack of Topicality. For example, Participant 3, after reading several lines of the document, realized that the document is talking about grizzly bears, and it has nothing to do with black bears. He/she said, "Uhm, so grizzlies are not the same as black bears, see if it mention black bears. Uhm yeah, this seems to be entirely about grizzlies so I'm going to say it's not relevant". Participant 10 also said, "So I'm seeing grizzly bear. Well that's different from black bears. So I'm scanning if it mentions black bears. So it's only about grizzlies. So I would say it's not relevant".

Causes of Relevant Judgments

- Lack of Concentration. We describe it as a kind of user behavior. It was the result of simple error when a participant judged a document as relevant, while it was entirely

unrelated to the search topic. For instance, Participant 2’s lack of concentration on the task caused the incorrect judgment. The entire document discusses the topic of grizzly bears without any mention of black bears, and the participant nevertheless judged the document to be relevant. He/she said, “Okay it was talking about food that’s a cause. Threatened animals. See if there is any recommendations. See if there’s any ways to control it. Oh. Okay. That’s good”. It is clear from the transcribed data here, that Participant 2 was focusing on finding the cause of the attack and not paying attention to what this document is talking about in general.

Reflection

The document is graded as 0 by NIST assessor. Again, this document is unrelated to the search topic and discusses the subject of grizzly bears instead of black bears. Eight (8) participants found it not relevant, while only one participant judged it as relevant. The participant who judged it as relevant lacked concentration since he/she was looking for the cause of the attack and entirely forgot that it this attack was about grizzly bears and not black bears.

Document ID: XIE19980113.0247

Title	Chinese Expert Says It’s Not Bigfoot, It’s Just Big Bears
NIST Score	0
Length (#Words)	262
ProbRelevance	0.222
Has Key Words	Yes “black bear”
No. of Judgments	9 Non-Relevant - 0 Relevant

Document's Description

This document is talking about the myth of bigfoot and how a Chinese expert is not supporting this myth by saying that these are in fact white-haired bears. The document has nothing to do with black bears attacks whatsoever.

Causes of Not Relevant Judgments relevant

- Lack of Topicality. Participant 4 said, “Brown bears, brown bears is pretty close to black bears. Okay, not relevant again. I don’t know why, I didn’t pay attention to big foot thing again”. Participant 12 also said, “seems like they talking about brown bears, not black bears. So, not relevant”.

None of the participants judged the document as relevant

Reflection

The document is graded as 0 by NIST assessor. This is not a long document and all participants judged it as not relevant. The entire document talks about the myth of bigfoot and what a Chinese expert states about this topic.

6.2.4 Topic 310

Document ID: NYT19990405.0532

Document's Description

The document is about confronting a cluster of brain tumours at the Amoco Research Center. It does not have anything to do with cell phones or car phones and brain cancer.

Title	Oil Company Confronts Cluster of Brain Tumors
NIST Score	0
Length (#Words)	608
ProbRelevance	0.05
Has Key Words	Yes "brain" , "cancers"
No. of Judgments	9 Non-Relevant - 0 Relevant

It discusses the steps that the company had taken to determine whether the reported cases of the tumours are work-related.

Causes of Not Relevant Judgments

- Lack of Topicality. Participant 3 said, "So again there's no mention of cell phones, specifically more about oil. So I'm going to say it's not relevant". Participant 8 also said, "This does not to do with cell phone or radio towers. So. It's just water and oil. So not relevant". Also, Participant 11 said, "This is all about chemicals and not anything about a radio tower from what I can see. Handling toxic chemicals were the cause. So, it's not relevant".

None of the participants judged the document as relevant

Reflection

The document is graded as 0 by NIST assessor. Though the whole document is about brain tumours, it is unrelated to the given search topic since it does not have any information about cell phones and radio towers. It is entirely about a number of employees in an oil company who were suffering from a rare type of brain tumour, and suspicion existed that this was related to their employment, since these employees were working in chemical labs

and exposed to chemical materials. The participants did not illustrate any difficulty in judging this document and all judged it correctly (as not-relevant).

Document ID: NYT19990805.0436

Title	Rare Cancer in Amoco Employees is Probably Work Related
NIST Score	0
Length (#Words)	874
ProbRelevance	0.071
Has Key Words	Yes “cancer” , “brain”
No. of Judgments	9 Non-Relevant - 0 Relevant

Document’s Description

This document also talks about the reported brain tumor cases at the Amoco Research Center. However, it mainly focuses on a report that was written by experts, who were hired by the company to investigate the causes of the brain tumours. The report talks in details about the results of this investigation. Nothing in this document is related to cell phones or car phones.

Causes of Not Relevant Judgments

- Lack of Topicality. Participant 5 said “So I’m just gonna think that this isn’t relevant. I’m just gonna skim through it if I see anything about radio waves but, ah, it doesn’t seem to be related”. Participant 11 said, “It does not report being lower, it just reports what the person who did the study says. I don’t think it is relevant”.
- Absence of Keywords. Participant 2 was searching for keywords “car phones” and “radio”. When he/she found nothing about them, he/she decided to judge the document as not relevant. He/she said, “It doesn’t look like it’s related. Nope. I would

say not relevant”. Participant 9 also said, “So that’s brain cancer. There’s nothing to do with phones in their. Okay, nothing. Alright. Not relevant”. Again, Participant 9 was searching the document for the word “cancer” and found a number of instances of this word, and then he/she searched for the word “phone” and found no instances. Based on this, he/she judged the document as not relevant.

None of the participants judged the document as relevant

Reflection

The document is graded as 0 by NIST assessor. Again, this document is similar to the previous, except this one discusses the report that has been written by a number of experts, as requested by the oil company, in order to determine whether the detected brain tumours are work-related. Similar to the results in the previous document, all participants judged it as not relevant.

Document ID: NYT19990907.0397

Title	DETAILS ON CARDINAL’S BRAIN TUMOR ARE SCANT
NIST Score	0
Length (#Words)	658
ProbRelevance	0
Has Key Words	Yes
	“brain”, “cancer”
No. of Judgments	9 Non-Relevant - 0 Relevant

Document’s Description

The document is about a cardinal’s diagnosis of a brain tumour. The document also gives some specific elaboration on brain cancer, and discusses more generally about who can be treated. The document contains no information on either cell phones or car phones.

Causes of Not Relevant Judgments

- **Absence of Keywords.** Participant 9 said, “But is there anything about cell phone out here. Ahh. This is about brain tumour in general. But is there anything about cell phones? Or car phones or any devices? No. So I say it’s not relevant”. He/she was searching for the word “cell” and found a number of instances of this word in the document, and then searched for “phone” and found no instance. Then, he/she judged the document as not relevant.
- **Lack of Topicality.** Participant 8 said, “It talks about brain tumours in general and same thing. Not radio waves or cell phones yet. Uhm. It is not relevant”. Participant 10 also said “So I’m seeing brain tumours all over the article. Uhm. So I’m looking if then it says anything about, Uhm, waves. And, I would say it’s not relevant. It’s just talking about brain cancer but nothing to do with waves”.

None of the participants judged the document as relevant

Reflection

The document is graded as 0 by NIST assessor. The document is about a single case which a cardinal experienced. Nothing in the document was about cell phones or car phones, though it discusses several types of brain cancer. None of the participants judged it as relevant.

Title	Capitol Shooting Exposes Cracks in Mental-Health Care System
NIST Score	0
Length (#Words)	1169
ProbRelevance	0.143
Has Key Words	Yes
	“Mental” , “Drugs” , “Illness”
No. of Judgments	8 Non-Relevant - 1 Relevant

6.2.5 Topic 383

Document ID: NYT19980727.0419

Document’s Description

This document contains no specific or generic drug names used for treating mental illness. The entire document is about the importance of treating individuals suffering with mental illness and making certain they receive the required help and treatment in order to control their behavior and their unexpected episodes of violence.

Causes of Not Relevant Judgments

- Absence of Specific Evidence. Participant 2 said, “That’s general. Okay. I haven’t found anything specific. Okay. It’s looks like it’s detailing me incidents. I would say not relevant”. Participant 12 also said, “So far, it seems it just be talking in general and not really mentioning any names of drugs about halfway through, though. So. Well, doesn’t seem relevant, based on not finding any drug names”.

Causes of Relevant Judgments

- Trouble Understanding the Search Topic. Participant 10 did not understand the search topic correctly. He/she said, “So it’s talking about mental illness. Uhm. So

it says antipsychotic drugs. Uhm. But now, I'm not sure if the criteria ... is that specific enough for the criteria. Specific or generic type of drug. Uhm. So I guess that's a generic type of drug. Antipsychotic. Uhm. So I guess I would say it's relevant". He/she was guessing and thought that the word "antipsychotic" in the document was sufficient to make it relevant, considering it to be a generic type of drug.

Reflection

The document is graded as 0 by NIST assessor. Though it is on the subject of mental illness, and the words "mental" and "illness" appear numerous times, a specific or generic drug name for treating mental illness is not included. When assessors are given such a document, they need to pay more attention since the document is specifically on mental illness, and there is a high probability it might have what they are searching for. Reading the entire document is the best approach an assessor should follow when assessing these types of documents.

Document ID: NYT19990711.0089

Title	Experts Say Study Confirms Prison's New Role as Mental Hospital
NIST Score	0
Length (#Words)	1159
ProbRelevance	0.04
Has Key Words	Yes
	"Mental" , "Illness" , "Drugs"
No. of Judgments	8 Non-Relevant - 1 Relevant

Document's Description

The document is about how prisons and jails presently also become mental illness hospitals,

since a considerable portion of the prisoners are mentally ill. Though the document talks in details about mentally ill prisoners, it does not include reference to any specific or generic drug. The phrase “antipsychotic drugs” is included, but this does not refer to either a specific nor a generic drug name.

Causes of Not Relevant Judgments

- **Absence of Specific Evidence.** Participant 5 said, “So I’m gonna say this isn’t relevant because I haven’t seen anything about drugs that used to treat mental illnesses. Ahh. It’s just about general”. Participant 9 also said, “This is not relevant. No, I don’t see any specific drug named. This is just generally talking about mental illness and to all the prisons obviously”.

Causes of Relevant Judgments

- **Trouble Understanding the Search Topic.** Participant 8 did not understand the search topic correctly. He/she said, ”So this is talking about bizarre behavior in jails and, Delusions, hallucinations. New antipsychotic drugs made medicating patients. Seem a human alternative to long-term hospitalization. Okay. So I think this is relevant. It talks about generic drugs”.

Reflection

The document is graded as 0 by NIST assessor. This document is also on the subject of mental illness, as the previous one, since it discusses mentally ill individuals who are in prisons and jails. However, nothing on specific or generic drugs is sated. Assessors might find it difficult not to read the entire document to search for relevant material, since it talks

specifically about mental illness of prisoners. There is mention of the phrase “antipsychotic drugs”, but this is not considered either a specific or a generic drug name.

Document ID: NYT20000717.0206

Title	Depictions Of Violent, Erratic Behavior Warp Perceptions of The Mentally Ill
NIST Score	0
Length (#Words)	2138
ProbRelevance	0.053
Has Key Words	Yes “Illness”, “Mental” , “Drugs”
No. of Judgments	9 Non-Relevant - 0 Relevant

Document’s Description

The document includes no reference to specific or generic drug names. It is about how mentally ill people are presently depicted in the culture. They are stigmatized as violent and erratic. The document includes many words on the topic of mental illness, but nothing about drug names.

What makes it non relevant

- Absence of Specific Evidence. Participant 3 said, “It’s the perception of the mentally ill. I don’t see any mentions so far of drugs for treating it just the perception of mental illness so that’s not relevant”. Participant 11 also said, “Talks about psychiatric terms. But not about treatment. No drug information again. Still no information. It doesn’t say any drug names. Still don’t see any treatments or drug names. And still don’t see any, Any treatment or drug names. No. I don’t see anything at all. So I’m gonna say not relevant”.

None of the participants judged the document as relevant

Reflection

The document is graded also as 0 by NIST assessor. However, as has been stated with reference to the two previous documents, though it is about mental illness, no specific or generic drug names are stated. Assessors need to be cautious when judging these types of documents and not judge them quickly as not relevant.

6.2.6 Topic 436

Document ID: XIE19960718.0212

Title	China Sets Railway Safety Record
NIST Score	0
Length (#Words)	185
ProbRelevance	0
Has Key Words	Yes “Railway” , “Accidents”
No. of Judgments	8 Non-Relevant - 1 Relevant

Document’s Description

The document is about how safety of Chinese Railways is presently better than in previous years. It discusses in brief the situation of the railways, and safety in general. It does not mention or describe anything on railways accidents.

Causes of Not Relevant Judgments

- Absence of Specific Evidence. Participant 3 said, “Railway safety record doesn’t sound relevant. Yeah, not relevant”. Also, participant 4 said, “No. It is not relevant

because it is not identifying any accidents. It is just general”. Participant 9 also said, “China Sets Railway Safety Record. Yes, that is more general safety stuff” and he/she judged the document as not relevant.

Causes of Relevant Judgments

- **Trouble Understanding the Search Topic.** Participant 2 judged the document incorrectly because he/she thought it includes the words “human deaths”, and this in his/her opinion to be a type of reporting on accidents, as mentioned in the description of the search topic. He/she said, “So does it have information on accidents? Yes. Human deaths. Okay. An then does it talk about prevention? Or safety? So documents that discuss railroading in general, new lines, new technology for safety. Safety and accident prevention. Okay. But does it have prevention? So I would say it’s still okay”.

Reflection

The document is graded as 0 by NIST assessor. This document neither mentions nor describes specific railway accidents. The entire document is on the subject of railway safety in China and how the safety record in that year was better than of the previous year. However, assessors’ difficulty in understanding the search topic might lead them to incorrectly assume that they have found relevant material in the document. For example, Participant 2 thought that mention of “human deaths” does qualify the document to meet the criteria mentioned in the description of the search topic.

Title	Shanghai Railway Bureau Sets Safety Record
NIST Score	0
Length (#Words)	107
ProbRelevance	0
Has Key Words	Yes “Accidents” , “Railway”
No. of Judgments	8 Non-Relevant - 1 Relevant

Document ID: XIE19970503.0122

Document’s Description

This document is also on the subject of safety and does not describe or mention any railways accidents. It talks about Shangahi railways safety records, where no major accidents occurred for 1200 consecutive days.

Causes of Not Relevant Judgments

- Absence of Specific Evidence. Participant 5 said, “Again, this is not relevant because it is more about safety”. Also, participant 8 said, “This talks about no accidents happening and safety. I don’t think it is relevant”. Participant 9 also said, “Shanghai Railway Bureau Sets Safety Record. Data on accidents. No, it is more general. So, it is not relevant”.

Causes of Relevant Judgments

- Difficulty Understanding the Search Topic. Here, Participant 12 judged the document as relevant, though it does not describe any accidents. The participant thought since the document mentions that there have been no accidents (zero accident), then this makes it relevant and fit the criteria of the search topic. He/she said “Shanghai Railway Bureau sets safety record. Title again doesn’t seem like this will be relevant.

Well, this one does give data, I guess, of zero accidents. So, I guess it's relevant".
From the transcribed data, we notice that Participant 12 was not certain about his/her judgment.

Reflection

The document is graded as 0 by NIST assessor. It is about safety in Shanghai's railway in general, and that they did not record any major railway accident in almost 3 years and 3 months. However, only one participant, Participant 12, judged the document incorrectly due to his/her own interpretation of the information in the search topic. He/she read the phrase "no accident" and assumed that this meant "zero accident" and that therefore the document is relevant, since it reports data on railways accidents.

Document ID: XIE20000724.0250

Title	Railway Inspections to Ensure Safety
NIST Score	0
Length (#Words)	174
ProbRelevance	0.042
Has Key Words	Yes "Railway", "Accidents"
No. of Judgments	8 Non-Relevant - 1 Relevant

Document's Description

The entire document discusses the inspection that the ministry of transportation in China wanted to conduct in order to ensure safety in all Chinese railways. The document does not describe any railway accidents.

Causes of Not Relevant Judgments

- Absence of Specific Evidence. Participant 4 said, “No ... not relevant. No accidents identified”. Also, Participant 10 said “So I’m just scanning for accidents. Uhm, This one doesn’t mention a specific accident. Uhm, So I’m gonna say it’s not relevant”. Participant 12 also said “Alright. Title is railway inspections to ensure safety. Well, based on that, I don’t think this article is uhh, is relevant. Lets just see if it talks about safety because if some kind of accident. Well, it doesn’t seem to mention a specific accident. So, not relevant”.

Causes of Relevant Judgments

- Trouble Understanding the Search Topic. Participant 9 was reading part of the topic description which states, “a relevant document will provide data on railway accidents ...”. Subsequently, he/she was reading a line of the document which states that, “the inspection are part of a nationwide push for transportation safety after several serious accidents in the past few months” and said, “Well, it is data. Well, okay, that is a tough one”. He/she then judged it as relevant. However, the participant here did not understand the search topic clearly. The description of the search topic required him/her to locate data about railways accidents and the cause of these accident. This means that it is necessary for the document to provide a description of the accident.

Reflection

The document is graded as 0 by NIST assessor. It discusses the inspection that the China’s ministry of transportation is willing to conduct. There is no mention of railways accidents

or description of these. However, one of the participants (Participant 9) judged it as relevant.

Chapter 7

Certainty Interfaces

In our recent qualitative study ([Al-Harbi and Smucker, 2014](#)), we observed assessors behaviour while making relevance judgments. We found that assessors judge the relevance of documents with different levels of certainty. These levels of certainty range from very uncertain judgments to very certain ones.

Firstly, the low level certainty relevance judgments are those which are made when assessors are entirely unsure of their judgements. When assessors produce low level certainty relevance judgments, they are merely guessing. Secondly, medium certainty relevance judgments are those which are made with a degree of certainty, but lack complete certainty. Lastly, high certainty relevance judgments are those made with complete or near complete certainty.

In this study, we used the same set of documents and set of search topics we used in our Think-Aloud Study. The selected documents were placed into two groups: low

consensus and high consensus documents. Low consensus documents are those that lack a majority agreement on their relevance, while high consensus documents are those that reach a majority agreement on their relevance. For more details about the selection of topics and documents, we refer you to Chapter 3 of this thesis. Based on the parameters mentioned in Chapter 3, there was an initial assumption that assessors tend to be uncertain when they disagree about relevance judgments. Therefore, we conducted a study to study the assessors' behaviour when judging both low and high consensus documents.

7.1 Methods and Materials

We designed a 2x2 factorial experiment. The first factor covers levels of certainty that an assessor could have had when judging the relevance of documents. The second factor covers the words or phrases that an assessor might have said or considered while judging the relevance of documents. The two sets of words/phrases were collected from our previous study (Al-Harbi and Smucker, 2014), where we observed participants to express their judgments by using these words and phrases.

7.1.1 Study Protocol

The study consisted of two parts: a practice part and the main task. On average, the required time to complete the entire study was an hour. Prior to working on the main search tasks, assessors practised on one search topic (SearchoTopic 427: UV Eye Damage). The objective here was to give the assessors the experience of what they would work on

1. Please judge the document below as relevant or not relevant to the search topic on the right.

Relevant	Not Relevant
----------	--------------

2. Please determine your level of certainty about your judgment of the document's relevance to the search topic.

Guessing	Definitely
----------	------------

Figure 7.1: First Certainty Relevance Judgments Interface

during the main task. In the practice part, assessors judged the relevance of four documents to the given search topic. Two of the documents were relevant to the search topic, while the other two were not relevant. Assessors were informed whether or not their judgments were correct, and they were permitted to ask questions or seek more clarification when necessary. In the main task part, assessors were given 4 search topics. For each search topic, there were 9 documents to be judged. In total, every assessor produced 36 relevance judgments. Also, since this part was the main element of our data collection process, the research facilitator did not intervene. Assessors were working on their own.

7.1.2 User Interfaces

We designed four user interfaces for this study. These interfaces fall into two groups, with two interfaces in each group: binary certainty interface group, and the ternary certainty interface group. All of the interfaces (in spite of their certainty levels) have the same answers for the first question about relevance judgments, either relevant or not relevant. However, the answers for the second question, which is about the level of certainty, differ

1. Please judge the document below as relevant or not relevant to the search topic on the right.

Relevant	Not Relevant
----------	--------------

2. Please determine your level of certainty about your judgment of the document's relevance to the search topic.

	Relevant for sure	Definitely relevant
	Pretty sure	Super relevant
	Relevant/Not relevant	Certainly relevant
I guess that serves	Not related	Very relevant
I do not see how this is relevant	It does not count	Not related at all
I am not sure	Pretty relevant	Definitely not relevant
Low	High	

Figure 7.2: Third Certainty Relevance Judgments Interface

1. Please judge the document below as relevant or not relevant to the search topic on the right.

Relevant	Not Relevant
----------	--------------

2. Please determine your level of certainty about your judgment of the document's relevance to the search topic.

Guessing	Maybe	Definitely
----------	-------	------------

Figure 7.3: Second Certainty Relevance Judgments Interface

1. Please judge the document below as relevant or not relevant to the search topic on the right.

Relevant	Not Relevant
----------	--------------

2. Please determine your level of certainty about your judgment of the document's relevance to the search topic.

	I say it is	I would say	
	I am going with	I still think	
	Slightly relevant	I am going with	Relevant for sure
	That would be	That sounds	Pretty sure
	Should be relevant	That might be	Relevant/Not relevant
I guess that serves	There is some relevancy	Maybe slightly relevant	Not related
I do not see how this is relevant	I will say	a bit relevant	It does not count
I am not sure	It looks like	That makes it	Pretty relevant
Low	Medium		High

Figure 7.4: Fourth Certainty Relevance Judgments Interface

in each interface. In the binary certainty interface group, assessors were given only two choices. However, these choices were different in the two interfaces in the binary group. In the first interface, Figure 7.1, we gave assessors only two words to consider: “Guessing” and “Definitely”. Each word represents a different certainty level. Guessing refers to a low certainty level, while Definitely refers to a high certainty level. For example, if an assessor chooses the term Guessing when answering the second question, this means that the assessor does not know the answer, and the answer is merely a guess. In contrast, if Definitely is chosen, this means that the assessor is highly certain about the produced relevance judgment. The certainty-choice answers in the other interface in the binary certainty interface group, which is the third interface, Figure 7.2, are represented by either low or high levels of certainty. We grouped a number of words and phrases under each level. The objective here is to give all possible words/phrases that might come to assessors’ minds while judging relevance of documents. Therefore, the third interface offered more

expressions than the first interface. On the other hand, the ternary certainty interface group also contained two interfaces: the second interface and the fourth interface. Both of these interfaces are composed of three levels of certainty. However, the second interface gave assessors three words in each level from which to choose. Guessing referred to the low certainty level, Maybe to the medium certainty level and Definitely to the high certainty level. Instead of giving single words, as in the second interface, Figure 7.3, in the fourth interface, Figure 7.4, assessors were given levels : low, medium, and high. Under each level, we grouped all of the words and phrases we found assessors to use in a previous study to express their relevance judgments.

7.1.3 Participants

Since our study involved human participation, an ethics clearance was obtained from the Office of Research Ethics at the University of Waterloo. Once clearance was received, we recruited participants via different means: emails, posters, and in-person invitations. We recruited 48 participants from different departments and programs. We targeted both graduate and undergraduate students. The participants' range of age was 18 to 34 years old, and the average age was 25 years old. For most participants, use of search engines is a daily activity. Eighty-four (84%) percent of the participants considered themselves experts in finding information. No training in information retrieval was given to any of the participants, with the exception of basic search training sessions offered by the university library.

7.1.4 Latin Square

In order to eliminate any influence of exposing the participants to all conditions of the experiment, we decided to use the between-subjects design where each participant is exposed to only one level of our two variables in the study. Therefore, we used the Latin Square to balance the order of the topics in the study. Since we used the between-subjects design, we only balanced the order of topics and this was done to ensure it would not have any impact on the way in which participants completed the study. Also, for each search topic and each participant, documents were randomized. We used 3 blocks of Latin Squares for each interface. The reason for this is that we have 4 topics and in order to create a Latin Square for 4 topics, it is necessary to create a 4x4 square. Consequently, 4 participants were required in each block.

7.1.5 Measuring Accuracy Rate

In this chapter, we are more interested to compute the accuracy rate, which represents the fraction of relevant and non-relevant documents which are judged correctly, at each level of certainty. The accuracy rate is measured as:

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |FN|} \quad (7.1)$$

where TP, FP, TN, and FN are from Table 7.1.

Participant	Relevant (Pos.)	Non-Relevant (Neg.)
Relevant	TP = True Pos.	FP = False Pos.
Non-Relevant	FN = False Neg.	TN = True Neg.

Table 7.1: Confusion Matrix “Pos.” and “Neg.” stand for “Positive” and “Negative” respectively.

7.2 Assessors’ Judging Behavior

7.2.1 Low consensus Documents

In this section, we show the results we obtained when analyzing assessors’ judging behavior when judging both high and low consensus documents. A low consensus document is one for which assessors do not have a majority agreement on its relevance, while a high consensus documents is one for which assessors have a majority agreement on its relevance. All of the results reported in the coming tables in this section are given with their standard errors.

The standard error of a statistic is defined as the standard deviation of the sampling distribution of the statistic; this statistic could be the mean, the proportion, or others (Walpole, 1974, p. 130). Since we deal with proportions in the tables in this chapter, we computed the standard error for proportions as described in (Triola, 2006) as the following:

$$\text{standard error} = \sqrt{\frac{p(1-p)}{n}} \quad (7.2)$$

where p is the sample proportion and n is the sample size. For example, to compute the standard error for the proportion(percentage) 28% at the low certainty level in the first interface in Table 7.2, we did the following:

$$\begin{aligned}
\text{Standard Error} &= \sqrt{\frac{p(1-p)}{n}} \\
&= \sqrt{\frac{0.28(1-0.28)}{144}} \\
&\approx 0.4,
\end{aligned}$$

where p represents the proportion for the low certainty relevance judgments, which is equal to 0.28 as illustrated in Table 7.2, and n represents the total number of relevance judgments for the low consensus documents in the first interface, which is equal to 144.

Binary Certainty Level Interfaces

Low Consensus Documents				
Interface Type	Percent of Judgments		Relevance Judgment Accuracy	
	Certainty Level		Certainty Level	
	Low	High	Low	High
First Interface	28% ±4	72% ±4	58% ±8	58% ±5
Third Interface	21% ±3	79% ±3	50% ±9	48% ±5

Table 7.2: Results. This table reports the percent of judgments made with binary certainty, as well as the accuracy of the relevance judgments. The standard error for each percentage is reported as well. Results are shown for low consensus documents.

When using a binary certainty interface, our results show that assessors tend to be certain even when they disagree about relevance judgments. We found that assessors to be at the same level of certainty when making high certainty relevance judgments (72% ±4

in the first interface and $79\% \pm 3$ in the third interface), regardless of whether they were correct or not correct in their judgments. Table 7.2 illustrates this finding. When we look at the high certainty relevance judgments in the binary certainty interfaces (first and third interfaces), we found the accuracy rate to be also similar $58\% \pm 5$ in the first interface and it is $48\% \pm 5$ in the third interface.

Similarly, the low certainty relevance judgments are only $28\% \pm 4$ in the first interface and $21\% \pm 3$ in the third interface. Similar percentages were found when calculating the accuracy rate for low certainty relevance judgments. In the first interface, the accuracy rate is $58\% \pm 8$ when making low certainty relevance judgments and it is $50\% \pm 9$ in the third interface. The accuracy rate when producing either high or low certainty relevance judgments is between 60% and 40%. This shows us that when judging low consensus documents, the level of certainty does not impact the accuracy rate.

Ternary Certainty Interfaces

Low Consensus Documents						
Interface Type	Percent of Judgments			Relevance Judgment Accuracy		
	Certainty Level			Certainty Level		
	Low	Medium	High	Low	Medium	High
Second Interface	$8\% \pm 2$	$37\% \pm 4$	$55\% \pm 4$	$67\% \pm 14$	$47\% \pm 6$	$53\% \pm 6$
Fourth Interface	$13\% \pm 3$	$26\% \pm 4$	$61\% \pm 4$	$37\% \pm 11$	$51\% \pm 5$	$52\% \pm 5$

Table 7.3: Results. This table reports the percent of judgments made with binary certainty, as well as the accuracy of the relevance judgments. The standard error for each percentage is reported as well. Results are shown for low consensus documents.

When analyzing data from the second and fourth interfaces (ternary interfaces), we found that assessors tend to also be highly certain about their relevance judgments. They produced more relevance judgments at a high level of certainty than at low or medium certainty. Assessors are found to be highly certain in their relevance judgments; 55% \pm 4 in the second interface, and 61% \pm 4 in the fourth interface. The medium certainty level judgments (37% \pm 4 in the second interface and 26% \pm 4 in the fourth interface) are greater than those which are low certainty (8% \pm 2 in the second interface and 13% \pm 3 in the fourth interface). The results for the accuracy rate of the high and medium certainty relevance judgments are similar (between 47% and 53%). The accuracy rate for the low certainty relevance judgments is almost the same in the second interface of that of the fourth interface.

7.2.2 High Consensus Documents

As a reminder, we refer to a document that has a majority agreement on its relevance as a high consensus document. We devote this subsection to discussion of the results of high consensus documents and certainty levels.

Binary Certainty Level Interfaces

Assessors continued to show the same tendency which they showed when judging the low consensus documents when judging high consensus documents. They were highly certain 78% \pm 2 in the first interface, and 83% \pm 2 in the third interface.

High Consensus Documents				
Interface Type	Percent of Judgments		Relevance Judgment Accuracy	
	Certainty Level		Certainty Level	
	Low	High	Low	High
First Interface	22% \pm 2	78% \pm 2	70% \pm 6	90% \pm 2
Third Interface	17% \pm 2	83% \pm 2	62% \pm 7	90% \pm 2

Table 7.4: Results. This table reports the percent of judgments made with binary certainty, as well as the accuracy of the relevance judgments. The standard error for each percentage is reported as well. Results are shown for high consensus documents.

Not surprisingly, there is also considerable increase in the accuracy rate when judging high consensus documents. This finding is anticipated since high consensus documents are found to be as such in the previous research studies conducted by our research group, as mentioned earlier in this chapter. Moreover, even when making the low certainty judgments, the accuracy rate has also increased, but slightly from the 50s in the low consensus documents to between the 60 and 70 percent in the high consensus documents.

Ternary Certainty Interfaces

When using the ternary certainty interfaces, we found the percentages of the high certainty relevance judgments in the second and fourth interfaces to be seventy-six (76%) \pm 3 and 68% \pm 3. What is worthwhile noting here, is that the low certainty relevance judgments decreased to 1% \pm 1 and 9% \pm 2 in both interfaces. The large majority of relevance judgments were at either high certainty level or at a medium certainty level, leaving only a very small percentage at a low certainty level, as illustrated in Table 7.5. This indicates to

High Consensus Documents						
Interface Type	Percent of Judgments			Relevance Judgment Accuracy		
	Certainty Level			Certainty Level		
	Low	Medium	High	Low	Medium	High
Second Interface	1% \pm 1	23% \pm 3	76% \pm 3	67% \pm 27	67% \pm 1	95% \pm 1
Fourth Interface	9% \pm 2	23% \pm 3	68% \pm 3	67% \pm 9	83% \pm 2	89% \pm 2

Table 7.5: Results. This table reports the percent of judgments made with binary certainty, as well as the accuracy of the relevance judgments. The standard error for each percentage is reported as well. Results are shown for high consensus documents.

us that assessors prefer to avoid indicating that they are entirely uncertain when making relevance judgments.

The accuracy rate is also high with high consensus documents. It is 95% \pm 1 in the second interface and 89% \pm 2 in the fourth interface. However, for medium or low certainty relevance judgments, a decline in the accuracy rate occurred. For example, it was 67% \pm 1 in the second interface and 83% \pm 2 in the fourth interface when making medium relevance judgments. Similarly, it was 67% \pm 27 in the second interface and 67% \pm 9 in fourth interfaces for low certainty relevance judgments.

7.2.3 Low vs. High Consensus Documents

We also studied the differences in assessors' relevance judgments when judging low and high consensus documents. We wanted to know if assessors are more certain when they judge low consensus documents or high consensus documents; also, if there is a difference, we wanted to know whether or not this difference is statistically significant. To answer

all of these questions, we computed a 95% confidence interval for difference between two population proportions at each level of certainty (low and high) for the binary certainty interfaces and (low, medium, and high) for the ternary certainty interfaces. We used the following equations as described in (Daniel, 1999) to compute the confidence interval for the difference between two population proportions at each level of certainty (note: C.I. refers to Confidence Interval):

$$\text{C.I.} = \text{Difference Between the Sample Proportions} \pm z \times (\text{Std Error for Difference}) \quad (7.3)$$

where z is a number that is taken from the normal curve and it indicates the level of confidence. When the level of confidence is 95%, z is equal to 1.96.

$$\text{Difference Between the Sample Proportions} = p_1 - p_2 \quad (7.4)$$

and the standard error for difference is computed as follow:

$$\text{Standard error for the difference} = \sqrt{\left(\sqrt{\frac{p_1(1-p_1)}{n_1}}\right)^2 + \left(\sqrt{\frac{p_2(1-p_2)}{n_2}}\right)^2} \quad (7.5)$$

where p_1 and p_2 are *sample*₁ and *sample*₂ proportions respectively and n_1 and n_2 are *sample*₁ and *sample*₂ sizes respectively.

Binary Interfaces

We computed a 95% confidence interval for the difference between two population proportions at each level of certainty (low and high), as illustrated in Table 7.6. We combined the relevance judgments from both single and multi-words interfaces at each level of consensus.

Our results show that though there is a slight increase in the percentages of high certain judgments when judging the high consensus documents (4%), this increase is not statistically significant. From the confidence interval column (CI for Difference) in Table 7.6, we note that the confidence interval for the differences between these two proportions (percentages) contains zero. Therefore, we conclude that there is no statistically significant difference in the two population values at 0.05 level of significance. Subsequently, we also cannot state that assessors were more certain when judging high consensus documents by using the binary certainty interfaces.

Percent of Judgments				
Consensus Level				
Certainty Level	High	Low	High - Low	CI for Difference
High	80% \pm 2	76% \pm 3	4%	4 \pm 6
Low	20% \pm 2	24% \pm 3	-4%	-4 \pm 6

Table 7.6: Results. This table reports the percent of judgments made with binary certainty interfaces, as well as the standard error at each level of certainty. Results are shown for low and high consensus documents.

Ternary Interfaces

We also made the same calculations for the ternary certainty interfaces (second and fourth interfaces). We computed a 95% confidence interval for the difference between two population proportions at each level of certainty (low, medium and high) as illustrated in Table 7.7.

Percent of Judgments				
Certainty Level	Consensus Level		High - Low	CI for Difference
	High	Low		
High	72% \pm 2	58% \pm 3	14%	14 \pm 7
Medium	23% \pm 2	31% \pm 3	-8%	-8 \pm 6
Low	5% \pm 1	11% \pm 2	-6%	-6 \pm 4

Table 7.7: Results. This table reports the percent of judgments made with ternary certainty interfaces, as well as the standard error at each level of certainty. Results are shown for low and high consensus documents.

Our results show that there is an increase in the percentages of high certain judgments in judging high consensus documents when using the ternary certainty interfaces (Single and Multi-Word interfaces combined); this increase is statistically significant as we see in Table 7.7. If we look at the confidence interval column in Table 7.7, we notice that the confidence interval for the differences between these two proportions (percentages) does not contain zero.

Therefore, we conclude that there is a statistically significant difference in the two population values at a 0.05 level of significance. We found the same results for the medium and low certainty judgments. There is a statistically significant difference between the

medium certainty judgments when judging the low consensus documents and the high consensus documents. The same observation can also be stated for the low certainty judgments and thus, there is some evidence that with a ternary interface that certainty is lower on low consensus documents.

7.3 Analysis of Time

We analyzed assessors' relevance judging time at each level of certainty. The numbers reported in Tables 7.8 and 7.9 represent the average relevance judging times at each level of certainty and their standard errors.

Since we were dealing with time data, we used the following equation to compute the standard error at each level of certainty (Triola, 2006):

$$\text{standard error} = \frac{s}{\sqrt{n}} \tag{7.6}$$

where s is the standard deviation and n is the number of relevance judgments at the desired certainty level. For example, to compute the standard error for the average judging time 76 at the low certainty level in the first interface in Table 7.8, we did the following:

$$\begin{aligned}
\text{Standard Error} &= \frac{s}{\sqrt{n}} \\
&= \frac{57}{\sqrt{40}} \\
&\approx 9,
\end{aligned}$$

where s represents the standard deviation for the low certainty relevance judgments, which is equal to 57, and n represents the number of relevance judgments at the low level of certainty, which is equal to 40.

7.3.1 Binary and Ternary Certainty Interfaces

When looking at the average relevance judging times in Tables 7.8 and 7.9, we notice that assessors in general appear to take less time to make their judgments as certainty increases. For instance, as it can be seen in Table 7.8, the average judging time for high certainty judgments is 51 ± 3 seconds compared to 81 ± 10 seconds when making low certainty judgments in the third interface at the high level of consensus. Likewise, when using the first interface, it is 53 ± 5 seconds when making high certainty relevance judgments and 76 ± 9 seconds when making low certainty relevance judgments at the low consensus level.

Similarly, when using the ternary certainty interfaces to judge the high consensus documents as shown in Table 7.9, assessors spent 89 ± 27 seconds when making low certainty judgments, 73 ± 8 seconds when making medium certainty judgments, and only 51 ± 3 seconds when making high certainty judgments in the second interface. Similar observation

Interface Type	Average Judging Time	
	Certainty Level	
	Low	High
Low Consensus Documents		
First Interface	76 ±9	53 ±5
Third Interface	70 ±8	46 ±3
High Consensus Documents		
First Interface	69 ±8	59 ±4
Third Interface	81 ±10	51 ±3

Table 7.8: Results. This table reports the average judging time in seconds with different levels of certainty, as well as the standard error. Results are shown for the binary certainty interfaces.

can be noticed when looking at the average judging times in the fourth interface when judging the low consensus documents.

Interface Type	Average Judging Time		
	Certainty Level		
	Low	Medium	High
Low Consensus Documents			
Second Interface	59 ±15	69 ±7	47 ±4
Fourth Interface	69 ±13	56 ±5	50 ±4
High Consensus Documents			
Second Interface	89 ±27	73 ±8	51 ±3
Fourth Interface	68 ±8	77 ±8	51 ±3

Table 7.9: Results. This table reports the average judging time in seconds with different levels of certainty, as well as the standard error. Results are shown for the ternary certainty interfaces.

That being said, we cannot conclude that there are statistically significant differences in the time it takes to judge documents in many cases due to the small size of the samples.

7.4 Conclusion

As our results show in this chapter, assessors tend to make high certainty relevance judgments, despite the consensus level of documents. Also, we found when judging high consensus documents, assessors' accuracy to be lower when making low certainty relevance judgments, and the judgments to be more accurate and tended to agree with NIST assessors when making high certainty relevance judgments. However, when judging low consensus documents, assessors' accuracy tends to be low regardless of their certainty level. In regard to the difference between the judgments made at each level of consensus, we did not find the difference to be statistically significant when using the binary certainty interfaces; however, the difference was statistically significant when using the ternary certainty interfaces. In other words, assessors made more certain judgments when judging the high consensus documents by using the ternary certainty interfaces. In consideration of the average judging time, in general assessors appear to spend less time making high certainty relevance judgments; however, due to the small sample sizes we are not able to conclude that there are statistically significant differences in the time it takes to judge documents in many cases.

Chapter 8

Conclusion and Future Directions

8.1 Conclusion

Study of the behavior of secondary assessors when making relevance judgments is important to the field of information retrieval, as it provides insight into user behavior. An understanding of this entire process will be helpful with developing superior retrieval systems that are capable of satisfying users information needs. In this thesis, we have conducted two user studies; the first is a qualitative study, and the second study is quantitative. The qualitative study (Think-Aloud Study) is supported by quantitative data. The two studies revealed many interesting findings about secondary assessors and the process by which they make relevance judgments. In the Think-Aloud Study, we investigated secondary assessors' judging behaviors in general. In the design of the study, considerable attention was given to ensure that we would select documents that help to best reflect what secondary

assessors experience when they judge documents, either in study labs or in crowdsourcing experiments. Therefore, we considered to divide the study documents into two groups: low and high consensus documents. The goal was to observe the behavior of secondary assessors when they are exposed to the two different groups. All of the conditions that ruled the selection of these documents, as well as the selection of the search topics, are discussed in detail in Chapter 3 of this thesis. In general, a high consensus document is that in which we have agreement on its relevance, while low consensus is one in which the agreement on its relevance is absent.

One of the major findings of our work deals with the level of certainty in secondary assessors' relevant judgments. We found that assessors are at different levels of certainty when judging documents and this is caused by a number of factors such as lack of familiarity, lack of understanding of the search topic and other factors that are discussed in detail in Chapters 5 and 6. An assessor might be at a high, medium, low level of certainty when he/she is making decisions about a document. Different levels of certainty were evident in the text that we had collected. Assessors may be completely perplexed while judging, and this will lead towards an arbitrary decision. In other instances, their level of confusion may be lower. Therefore, we may conclude that the certainty level is not consistent, and varies based upon particular circumstances that the assessor encounters during the judgment process.

8.1.1 Certainty in Relevance Judgments

We found that assessors relevance judgments fall under three levels: low, medium, and high certainty. Each level represents a state that describes an assessor’s judging behavior. The low level of certainty refers to the level when an assessor is unsure and hesitant in his/her judgment. Relevance judgments at a low level of certainty are no more than a guess. The medium level of certainty is a higher level than the low one and relevance judgments are made with certainty, but lack of complete certainty. This was evident from the types of terms and expressions that assessors used while judging documents. Finally, the high level of certainty is at the top in which assessors are certain and confident in their judgments. In Chapter 4, we explain these levels in much greater detail and several examples are provided to illustrate each level.

8.1.2 Categories of Incorrect Relevance Judgments

In IR, we know that assessors make incorrect relevant judgments, and there are a number of studies that quantitatively investigated secondary assessor behavior when making relevance judgments. For example, in (Smucker and Jethani, 2012), they found that assessors tend to take much more time when making incorrect judgments. Our study, explains this type of behavior qualitatively rather than quantitatively. Therefore, we divided the incorrect relevance judgment based on four categories: search topic, document, assessor, and NIST assessor’s mistakes.

In the search topic category, we found that assessors were experiencing difficulty to find a good match between what is mentioned in the search topic’s description and the material

in the document. Assessors were found to have consulted the search topic several times in such cases, and were not certain whether they correctly identified criterial material relevant to the search topic. This type of behavior is not unexpected, since secondary assessors sometimes do not adequately comprehend what the person who wrote the search topic and its description means and wants to find. In fact [Kinney et al. \(2008\)](#) specifically points out this problem, and found that non-experts (referred to as generalists in the paper) show weak understanding of the true meaning of the search topics which they have been given. The role of an expert or primary assessor here would solve such an issue because he/she would be a better judge to identify whether or not the material would fit the criteria. We call this category “Difficulty in Applying the Search Topic”. This type of category in our opinion would still remain an issue in the relevance judgments produced by secondary assessors since they are not the originators and may necessarily not have interest in the search topic. However, careful design of a search topic might be helpful with reducing errors frequency in this category.

Our second category is the document. The categories of document and search topic differ in that the difficulty is not in applying the search topic, but rather in processing the document material. An assessor sometimes may show good understanding of the search topic; however, he/she might not be able to locate or find relevant material in the document. In some search topics, the relevant material is contained in the document; however, it is necessary that the assessor process the document carefully and not to overlook the relevant material. We believe that errors under this type of category will be more apparent in search topics that require locating specific information such as a drug name or an individuals

name. Moreover, if a document is long, this will make the process more challenging, and a patient attitude is required for accurate judgment.

The Assessors category is our third category discussed in this thesis. When recruiting assessors, we are aware that they have different backgrounds and varying levels of skill and expertise. From the collected data, we have found that they are different in a number of aspects, such as relevance judging completion time, levels of concentration, the expression of ideas and thoughts and in being liberal and conservative in their behavior when judging documents. In a number of instances of our collected data, assessors at times demonstrate a lack of concentration when making relevance judgments. Also, an assessor may encounter unfamiliar terminology and as a result he/she will be in doubt about the decision to be made. In other cases, he/she will judge a document based on his/her own understanding of the terminology. Again, an expert in the search topic would not experience the same degree of difficulty as a secondary assessor. This is due to his/her higher level of familiarity with concepts and terminology in the search topic target area.

The fourth and the last category is what we refer to as NIST assessor mistakes. Secondary assessors may be correct in their judgment; however, the system might count these judgments as incorrect only because they do not comply or match the NIST assessor judgment. [Grossman and Cormack \(2011\)](#), also address such an issue when discussing the primary assessor (they refer to him/her as Topic Authority) making incorrect relevance judgments.

8.1.3 Low Consensus Documents

These type of documents might be a source of confusion not only to a secondary assessor, but also occasionally to the NIST assessor. Documents of this type would typically contain partial relevant material. However, since the search topic is not written or generated by the secondary assessor, he/she might sometimes be more reluctant and experience difficulty to judge such a document. A NIST assessor however, should not have the same difficulty in understanding the search topic and deciding whether what is stated in the document does fit the criteria of the search topic partially or totally. He/she only has to decide on is the level of relevance of the document, as highly relevant or at a lower level of relevance to the search topic. This level of understanding is not the same to a secondary assessor. He/she might locate a piece of information in the document, but he/she will experience difficulty deciding whether it fits the criteria of the search topic that was given to him/her. Several questions might be posed when he/she is in a situation such as this. For example, he/she may ask a self-question, such as does the person who wrote the search topic means “such and such” or does he/she wants to find about “such and such”. If he/she does not fully understand the meaning of the search topic, he/she will use his/her own intuition to interpret what the author of the search topic means. Our findings in this respect are consistent with the ones were found by (Kinney et al., 2008). However, our results here address more the low consensus documents. This of course will lead to different results sometimes, since humans’ understanding varies, and the role of subjectivity will be more dominant here. Moreover, the type of search topic also affects the degree of difference in relevant judgments. Some search topics are either difficult to judge or

require a good level of understanding and familiarity with the terminology in the subject field. The Think-Aloud data confirms this to us by providing several examples where assessors experience more challenges with some topics than others. From the assessors own words, we discovered that lacking familiarity could be a factor in the confusion of their relevance judgments. In fact, the Think-Aloud data revealed to us more than this. In many instances, we found assessors to make judgments correctly; however, the judgments were based on inaccurate rationale and reasoning. The quantitative methods would not be able to give such a profound understanding of secondary assessors behavior. Typically, when researchers conduct experiments and relevance judgments are obtained, we presume that all correct relevance judgments to be based on a good level of understanding of the search topic. However, this is not always true, as we have seen in the data collected in the think-aloud study.

8.1.4 High Consensus Documents

In high consensus documents, assessors of our study were generally able to locate the relevant material more easily than in low consensus documents. Relevant materials are closer to the surface and a major effort is not required to determine whether a document is relevant. However, challenging topics may require more effort from secondary assessors to judge, since they include a higher degree of unfamiliar terminology or necessitate good ability in inferencing to find the relevant material. In our study, Topic 310 (Radio Waves and Brain Cancer) and Topic 336 (Black Bear Attacks) are examples of such topics. Assessors in Topic 310 and Topic 336 were given a number of points in the description they

needed to consider when searching for relevant materials in documents. Documents of this type might be highly relevant to the search topic because they meet one criterion of the search topic, but unfortunately assessors might overlook this and merely focus on one part of the description, ignoring the other parts, and as a result they judge the document as not relevant. We think that if the description of the search topic is divided into smaller parts and these are individually assigned to assessors, this may be helpful with not overlooking relevant material in the document. In some search topics such as Topic 383 (Mental Illness Drugs) and Topic 436 (Railway Accidents), the relevant material is almost on the surface, if not even included in the document's title itself. Therefore, assessors would not find judging such documents a difficult task, unlike Topic 310 and Topic 336 which we mentioned earlier. The type of errors assessors make when judging high consensus documents for topics, such as Topic 383 and 436 are uncommon and rare since they are the result of almost a complete lack of concentration or a totally inaccurate interpretation of the search topic.

8.1.5 Certainty Interfaces

In our second user study, we investigated secondary assessors behavior when they are provided with interfaces that collect uncertainty relevance judgments. We built four certainty interfaces in which we asked assessors two main questions in each interface. The first question is the same in all interfaces, which is about judging the document as either relevant or not relevant. The second question differs in each interface. However, the objective of the second question in all interfaces is to collect uncertainty information about the rele-

vant judgments assessors make when judging each documents. Two of the interfaces were built to collect binary certainty responses, while the other two were built to collect ternary certainty responses. The difference between the two interfaces that were used to collect binary certainty responses, is that we provided assessors with only one word in each level of certainty, while we provided several words and phrases in the second one. The same procedure was done with the ternary interfaces. The goal with providing assessors with single-words and multi-words interfaces is to investigate if providing additional words at each level of certainty would change their behavior in making more uncertain or certain judgments on one level more than the other(s).

The results of our second study show that assessors tend to be certain in their relevance judgments despite the level of consensus of the documents. In other words, they will judge low and high consensus documents with a high level of certainty. This type of behavior is observed in all the interfaces we used in the study. Moreover, assessors were more accurate when using single-word interfaces compared to the multi-word interfaces. When analyzing the relevance judging time, we found also that assessors spend less time when making high certainty relevance judgments, compared to medium and low certainty relevance judgments.

The results we obtained from the second user study show us clearly that the tendency to make high certainty judgments is the most dominant and this may reflect the nature of how assessors think and behave when making relevance judgments. Secondary assessors have different beliefs, ways of thinking and perceptions. In addition, when they are certain, they will not spend much time on a document, unlike when they think they are not certain and require additional time to carefully think and read further in order to reach a decision.

8.2 Contributions

In summary, our key contributions were:

1. We found secondary assessors to judge relevance documents at different levels of certainty. These levels of certainty vary from low certainty to high certainty.
2. We showed the various categories of incorrect relevance judgments. These categories include: difficulty in applying a search topic, difficulty in processing a document, secondary assessor factors, and true error in primary assessor judgment.
3. We found through preliminary evidence how the number of details in a search topic's description could impact a secondary assessor's judging behavior.
4. We showed how the ability to articulate thoughts is different among assessors; however, this does not impact their overall performance in judging relevance of documents.
5. We found the age of assessors to impact their understanding of certain search topics. Some younger assessors may not be able to recognize older technologies and this results in not satisfactorily understanding a given search topic.
6. A document length and location of relevant material in the document were found to play an essential role in making it low consensus. This finding is not specific to secondary assessors but also to primary assessors who might encounter the same issue.

7. We showed how different factors such as lack of familiarity/knowledge, lack of understanding of the search topic, absence of keywords and other factors could be a source of variations in relevance judgments.
8. We showed how variations in the judgments could be caused by the length of a search topic. Some assessors focus on or consider only part of the description while forgetting or overlooking other sections.
9. We showed that when judging a document correctly, this does not mean an assessor's decision is necessarily based on correct perception. Misunderstanding of the search topic or the content of the document might occur.
10. We found secondary assessors' tendency to be certain is dominant when making relevance judgments regardless of whether or not the document high consensus.
11. When judging high consensus documents, assessors' accuracy was found to be lower when making low certainty relevance judgments, and to be higher and tended to agree with NIST assessors when making high certainty relevance judgments.
12. When judging low consensus documents, assessors' accuracy was found to be low, regardless of their certainty level.
13. When judging low consensus documents using the ternary certainty interface, we found certainty to be lower.

8.2.1 Limitations

The limitation in our second user study is a the missing of a controlling interface, where we asked assessors only to judge the document as relevant and not relevant without collecting uncertainty responses. The goal of a controlling interface is to measure whether collecting uncertain responses will impact the performance of assessors. This of course would add interesting findings to our results. There are a number of studies such as ([Wakeling et al., 2016](#); [Ruthven et al., 2007](#)) which ask assessors about their confidence level in the given search topics. In these studies, the degree of confidence was found to impact the quality of the results. However, our approach would be different since we will ask assessors to report their certainty when judging each document, instead of declaring the confidence level before or after each search task. In other words, each time an assessor judges a document as relevant or not, he/she also is required to report his/her certainty level. Our assumption is that reporting the degree of certainty would also impact the quality of the relevance judgments we obtain, since assessors need to carefully consider each time they judge a document if they are certain or not.

8.3 Future Work

Studying the impact of uncertainty on assessors performance:

We have an interest to study the impact of certainty on assessors' performance. Therefore, we plan to conduct a study in the future where we use the same certainty interfaces which we used in our second user study; however, this time we will add a controlling interface

where we do not include a question about certainty. In the controlling interface, the aim is simply to ask assessors to judge documents as relevant or not relevant. However, we will still employ the other certainty interface to collect the uncertainty level in assessors relevance judgments. The objective of this type of experiment is to investigate whether or not instructing assessors to report their level of certainty when making relevance judgments, will impact assessors' performance.

Evaluating Systems based on the high consensus documents:

Relevance assessments made by secondary assessors have been compared to those of primary assessors in several studies ([Wakeling et al., 2016](#); [Alonso and Mizzaro, 2012](#)). In these studies and others, secondary assessors were found to produce reliable relevance judgments. It is true that the overlap between these relevance judgments, those made by primary assessors and secondary assessors, varies. However, if a careful design, execution, and quality control are considered as stated by [Alonso and Mizzaro \(2012\)](#), we might use secondary assessors' relevance judgments to replace those of primary assessors when evaluating IR systems. Therefore, we think if we apply the concept of high and low consensus documents when comparing search systems by the relevance judgments made by secondary assessors, we would achieve comparable results, similar to if we use primary assessors' relevance judgments. In other words, we crowdsource the collection of relevance judgments where we recruit as many assessors as possible. This means that we ask secondary assessors to judge the documents retrieved by search systems that we wish to compare. After judging these, we classify those documents based on their consensus levels. We then only consider the high consensus documents when evaluating the search systems and exclude

the low consensus. In this manner, we only consider the highly relevant documents in the evaluation. However, the low consensus documents could be given to additional assessors (experts) to be readjudicated. Therefore, it will not be necessary for the experts to go through an entire set of documents, and need only to adjudicate the low consensus ones.

Impact of Search Topic’s Length:

We are also interested to study the impact of topic descriptions on assessors’ agreements. We will examine the behavior of assessors when they are provided with different types of topic descriptions. Some of topic descriptions will be general, while others will be specific. We decided to study this for two reasons. First, we observed that assessors seem to differ in their false positive rates and false negative rates when the length (detail) of topic descriptions varied. Secondly, there is no complete study on only the impact of topic descriptions on assessors’ agreement. The only work of which we are aware of are ([Webber et al., 2012b](#); [Wakeling et al., 2016](#)); however, we believe that [Webber et al. \(2012b\)](#) study is insufficient for the following reasons: (1) it included only two participants, and this is not sufficient to cover and study the issue and make meaningful generalizations about the results; and (2) it is focused on legal-track relevance assessments only. Also, [Wakeling et al. \(2016\)](#) have compared between primary and secondary assessors when making relevance judgments on what they call “real-life topics”. They claim that there is no relationship between the length of a topic and the number of documents judged relevant. The topics they used in their study were divided into two groups: open and closed topics. However, our intention is to divided long search topics into small parts and give assessors these parts

one by one. We want to study how secondary assessors' performance will be when they are given different search topics length and different parts of the description of a search topic.

References

- Hervé Abdi. Signal detection theory (sdt). *Encyclopedia of measurement and statistics*, pages 886–889, 2007.
- Aiman L Al-Harbi and Mark D Smucker. User expressions of relevance judgment certainty. 2013.
- Aiman L. Al-Harbi and Mark D. Smucker. A qualitative exploration of secondary assessor relevance judging behavior. In *Proceedings of the 5th Information Interaction in Context Symposium, IiX '14*, pages 195–204, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2976-7. doi: 10.1145/2637002.2637025. URL <http://doi.acm.org/10.1145/2637002.2637025>.
- Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 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*, pages 683–684. ACM, 2008.
- Omar Alonso. Crowdsourcing for information retrieval experimentation and evaluation. In *Multilingual and Multimodal Information Access Evaluation - Second International Conference of the Cross-Language Evaluation Forum, CLEF 2011, Amsterdam, The Netherlands, September 19-22, 2011. Proceedings*, page 2, 2011. doi: 10.1007/978-3-642-23708-9_2. URL http://dx.doi.org/10.1007/978-3-642-23708-9_2.
- Omar Alonso and Ricardo A. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, pages 153–164, 2011. doi: 10.1007/978-3-642-20161-5_16. URL http://dx.doi.org/10.1007/978-3-642-20161-5_16.

- Omar Alonso and Matthew Lease. Crowdsourcing for information retrieval: principles, methods, and applications. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 1299–1300, 2011. doi: 10.1145/2009916.2010170. URL <http://doi.acm.org/10.1145/2009916.2010170>.
- Omar Alonso and Stefano Mizzaro. Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, volume 15, page 16, 2009.
- Omar Alonso and Stefano Mizzaro. Using crowdsourcing for trec relevance assessment. *Inf. Process. Manage.*, 48(6):1053–1066, November 2012. ISSN 0306-4573. doi: 10.1016/j.ipm.2012.01.004. URL <http://dx.doi.org/10.1016/j.ipm.2012.01.004>.
- Omar Alonso, Daniel E. Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008. doi: 10.1145/1480506.1480508. URL <http://doi.acm.org/10.1145/1480506.1480508>.
- Amazon. Amazon mechanical turk, 2016. URL <https://www.mturk.com/mturk/welcome>. [Online; accessed 4-May-2016].
- Javed A Aslam, Virgil Pavlu, and Emine Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 541–548. ACM, 2006.
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.
- Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P de Vries, and Emine Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 667–674. ACM, 2008.
- Stephen Bales and Peiling Wang. Consolidating user relevance criteria: A meta-ethnography of empirical studies. *Proceedings of the American Society for Information Science and Technology*, 42(1), 2005.
- G.C. Barhydt. A comparison or relevance assessments by three types of evaluator. In *Proceedings of the American Documentation Institue*, pages 383–385. American Documentation Institue, 1964.

- G.C. Barhydt. The effectiveness of non-user relevance assessments. *Journal of Documentation*, 23:146–149, 1967.
- Carol L. Barry. User-defined relevance criteria: an exploratory study. *JASIS*, 45(3):149–159, 1994.
- Pia Borlund. Experimental components for the evaluation of interactive information retrieval systems. *Journal of documentation*, 56(1):71–90, 2000.
- Pia Borlund. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10):913–925, 2003a.
- Pia Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research. An International Electronic Journal*, 8(3), 2003b.
- Pia Borlund and Peter Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of documentation*, 53(3):225–250, 1997.
- Jennifer L Branch. The trouble with think alouds: Generating data using concurrent verbal protocols. *Proc. of CAIS*, 2000.
- Robert Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing & Management*, 28(5):619–627, 1992.
- Stefan Büttcher, Charles LA Clarke, and Gordon V Cormack. *Information retrieval: Implementing and evaluating search engines*. Mit Press, 2010.
- Chris Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using amazon’s mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP ’09, pages 286–295, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL <http://dl.acm.org/citation.cfm?id=1699510.1699548>.
- Ben Carterette and Ian Soboroff. The effect of assessor error on ir system evaluation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’10, pages 539–546, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0153-4. doi: 10.1145/1835449.1835540. URL <http://doi.acm.org/10.1145/1835449.1835540>.

- Praveen Chandar, William Webber, and Ben Carterette. Document features predicting assessor disagreement. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 745–748. ACM, 2013.
- Elizabeth Charters. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education Journal*, 12(2), 2003.
- Alexandra Chouldechova and David Mease. 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, WSDM '13*, pages 103–112, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1869-3. doi: 10.1145/2433396.2433411. URL <http://doi.acm.org/10.1145/2433396.2433411>.
- Heting Chu. Factors affecting relevance judgment: a report from trec legal track. *Journal of Documentation*, 67(2):264–278, 2011.
- Cyril W Cleverdon. The effect of variations in relevance assessments in comparative experimental tests of index languages. 1970.
- Paul Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, and Amy Warner. Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing*, 17(4): 32–38, July 2013. ISSN 1089-7801. doi: 10.1109/MIC.2012.95. URL <http://dx.doi.org/10.1109/MIC.2012.95>.
- Colleen Cool, Nicholas Belkin, Ophir Frieder, and Paul Kantor. Characteristics of text affecting relevance judgments. In *NATIONAL ONLINE MEETING*, volume 14, pages 77–77. LEARNED INFORMATION (EUROPE) LTD, 1993.
- Gordon V Cormack, Christopher R Palmer, and Charles LA Clarke. Efficient construction of large test collections. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 282–289. ACM, 1998.
- Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 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*, pages 758–759. ACM, 2009.
- Erica Cosijn and Peter Ingwersen. Dimensions of relevance. *Inf. Process. Manage.*, 36(4):533–550, July 2000a. ISSN 0306-4573. doi: 10.1016/S0306-4573(99)00072-2. URL [http://dx.doi.org/10.1016/S0306-4573\(99\)00072-2](http://dx.doi.org/10.1016/S0306-4573(99)00072-2).

- Erica Cosijn and Peter Ingwersen. Dimensions of relevance. *Inf. Process. Manage.*, 36(4): 533–550, July 2000b. ISSN 0306-4573.
- W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*, volume 283. Addison-Wesley Reading, 2010.
- W.W. Daniel. *Biostatistics: a foundation for analysis in the health sciences*. Wiley Series in Probability and Statistics. Wiley, 1999. ISBN 9780471163862.
- Efthimis N Efthimiadis and Mary A Hotchkiss. Legal discovery: Does domain expertise matter? *Proceedings of the American Society for Information Science and Technology*, 45(1):1–2, 2008.
- Carsten Eickhoff, Christopher G. Harris, Arjen P. de Vries, and Padmini Srinivasan. 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*, SIGIR '12, pages 871–880, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348400. URL <http://doi.acm.org/10.1145/2348283.2348400>.
- Michael Eisenberg and Carol Barry. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science*, 39(5):293–300, 1988.
- David Ellis. A behavioural model for information retrieval system design. *Journal of information science*, 15(4-5):237–247, 1989.
- K Anders Ericsson and Herbert A Simon. Verbal reports as data. *Psychological review*, 87(3):215, 1980.
- K Anders Ericsson and Herbert A Simon. Protocol analysis, 1993.
- Raya Fidel. Qualitative methods in information-retrieval research. *Library & Information Science Research*, 15:219–247, 1993 1993. URL <http://faculty.washington.edu/fidelr/RayaPubs/QualitativeMethodsInInformationRetrievalResearch.pdf>.
- Thomas J. Froehlich. Relevance reconsidered—towards an agenda for the 21st century: Introduction to special topic issue on relevance research. *J. Am. Soc. Inf. Sci.*, 45(3):124–134, April 1994. ISSN 0002-8231. doi: 10.1002/(SICI)1097-4571(199404)45:3<124::AID-ASI2>3.0.CO;2-8. URL [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(199404\)45:3<124::AID-ASI2>3.0.CO;2-8](http://dx.doi.org/10.1002/(SICI)1097-4571(199404)45:3<124::AID-ASI2>3.0.CO;2-8).

- Rebecca Green. Topical relevance relationships. I: Why topic matching fails. *J. Am. Soc. Inf. Sci.*, 46(9):646–653, October 1995. ISSN 0002-8231. doi: 10.1002/(SICI)1097-4571(199510)46:9<646::AID-ASI2>3.0.CO;2-1. URL [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(199510\)46:9<646::AID-ASI2>3.0.CO;2-1](http://dx.doi.org/10.1002/(SICI)1097-4571(199510)46:9<646::AID-ASI2>3.0.CO;2-1).
- Howard Greisdorf. Relevance thresholds: a multi-stage predictive model of how users evaluate information. *Information Processing & Management*, 39(3):403–423, 2003.
- Maura R Grossman and Gordon V Cormack. 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*, pages 1–11. Citeseer, 2011.
- C. D. Gull. Seven years of work on the organization of materials in the special library. *American Documentation*, 7(4):320–329, 10 1956. ISSN 1936-6108. doi: 10.1002/asi.5090070408. URL <http://dx.doi.org/10.1002/asi.5090070408>.
- Donna Harman. Overview of trec-1. In *Proceedings of the workshop on Human Language Technology*, pages 61–65. Association for Computational Linguistics, 1993.
- Donna Harman. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(2):1–119, 2011.
- Stephen P Harter. Psychological relevance and information science. *Journal of the American Society for information Science*, 43(9):602, 1992.
- Stephen P Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *JASIS*, 47(1):37–49, 1996.
- Sandra G Hirsh. Children’s relevance criteria and information seeking on electronic resources. *Journal of the American Society for information Science*, 50(14):1265–1283, 1999.
- Jeff Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, New York, NY, USA, 1 edition, 2008. ISBN 0307396207, 9780307396204.
- Joan E Hughes, BW Packard, and P David Pearson. Reading classroom explorer: Navigating and conceptualizing a hypermedia learning environment. *Reading Online*, 1998.

- Joseph W. Janes. 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):160–171, 1994. ISSN 1097-4571. doi: 10.1002/(SICI)1097-4571(199404)45:3<160::AID-ASI6>3.0.CO;2-4. URL [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(199404\)45:3<160::AID-ASI6>3.0.CO;2-4](http://dx.doi.org/10.1002/(SICI)1097-4571(199404)45:3<160::AID-ASI6>3.0.CO;2-4).
- Joseph W Janes and Renee McKinney. Relevance judgments of actual users and secondary judges: A comparative study. *The Library Quarterly*, pages 150–168, 1992.
- C.P. Jethani. Effect of prevalence on relevance assessing behavior. Master's thesis, University of Waterloo, 2011.
- Gabriella Kazai, Natasa Milic-Frayling, and Jamie Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 452–459, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572019. URL <http://doi.acm.org/10.1145/1571941.1572019>.
- Jaana Kekäläinen. Binary and graded relevance in ir evaluationscomparison of the effects on ranking of ir systems. *Information processing & management*, 41(5):1019–1033, 2005.
- Jaana Kekäläinen and Kalervo Järvelin. Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.
- Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(12):1–224, 2009.
- Kenneth A. Kinney, Scott B. Huffman, and Juting Zhai. How evaluator domain expertise affects search result relevance judgments. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 591–598, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458160. URL <http://doi.acm.org/10.1145/1458082.1458160>.
- Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357127. URL <http://doi.acm.org/10.1145/1357054.1357127>.

- Carol C Kuhlthau. Inside the search process: Information seeking from the user's perspective. *JASIS*, 42(5):361–371, 1991.
- Michael E Lesk and Gerard Salton. Relevance assessments and retrieval system evaluation. *Information storage and retrieval*, 4(4):343–359, 1968.
- Le Li and Mark D Smucker. Tolerance of effectiveness measures to relevance judging errors. In *Advances in Information Retrieval*, pages 148–159. Springer, 2014.
- Neil A Macmillan and C Douglas Creelman. *Detection theory: A user's guide*. Psychology press, 2004.
- Stefano Mizzaro. How many relevances in information retrieval? *Interacting with computers*, 10(3):303–320, 1998.
- Douglas W Oard, Jason R Baron, Bruce Hedin, David D Lewis, and Stephen Tomlinson. Evaluation of information retrieval for e-discovery. *Artificial Intelligence and Law*, 18(4):347–386, 2010.
- John O'Connor. Some independent agreements and resolved disagreements about answer-providing documents. *American Documentation*, 20(4):311–319, 1969. ISSN 1936-6108. doi: 10.1002/asi.4630200405. URL <http://dx.doi.org/10.1002/asi.4630200405>.
- Taemin Kim Park. The nature of relevance in information retrieval: An empirical study. *The library quarterly*, pages 318–351, 1993.
- Taemin Kim Park. Toward a theory of user-based relevance: A call for a new paradigm of inquiry. *J. Am. Soc. Inf. Sci.*, 45(3):135–141, April 1994. ISSN 0002-8231. doi: 10.1002/(SICI)1097-4571(199404)45:3<135::AID-ASI3>3.0.CO;2-1. URL [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(199404\)45:3<135::AID-ASI3>3.0.CO;2-1](http://dx.doi.org/10.1002/(SICI)1097-4571(199404)45:3<135::AID-ASI3>3.0.CO;2-1).
- Sarah Ransdell. Generating thinking-aloud protocols: Impact on the narrative writing of college students. *The American journal of psychology*, pages 89–98, 1995.
- Alan M Rees and Douglas G Schultz. A field experimental approach to the study of relevance assessments in relation to document searching. final report to the national science foundation. volume i. 1967.
- Ian Ruthven, Mark Baillie, and David Elswiler. The relative effects of knowledge, interest and confidence in assessing relevance. *Journal of Documentation*, 63(4):482–504, 2007.

- Mark Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010.
- Tefko Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6): 321–343, 1975.
- Tefko Saracevic. Relevance reconsidered. In *Information science: Integration in perspectives. In Proceedings of the Second Conference on Conceptions of Library and Information Science*, volume 1, pages 201–218, 1996.
- Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):2126–2144, 2007.
- Tefko Saracevic. Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective. *Library Trends*, 56(4):763–783, 2008.
- Linda Schamber. Relevance and information behavior. *Annual review of information science and technology (ARIST)*, 29:3–48, 1994.
- Linda Schamber, Michael Eisenberg, and Michael S. Nilan. A re-examination of relevance: Toward a dynamic, situational definition. *Inf. Process. Manage.*, 26(6):755–776, November 1990. ISSN 0306-4573. doi: 10.1016/0306-4573(90)90050-C. URL [http://dx.doi.org/10.1016/0306-4573\(90\)90050-C](http://dx.doi.org/10.1016/0306-4573(90)90050-C).
- Falk Scholer, Andrew Turpin, and Mark Sanderson. 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*, pages 1063–1072. ACM, 2011.
- Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S Lee, and William Webber. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 623–632. ACM, 2013.
- Mark D Smucker and Chandra Prakash Jethani. Human performance and retrieval precision revisited. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 595–602. ACM, 2010.

- Mark D. Smucker and Chandra Prakash Jethani. Measuring assessor accuracy: A comparison of nist assessors and user study participants. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1231–1232, New York, NY, USA, 2011a. ACM. ISBN 978-1-4503-0757-4. doi: 10.1145/2009916.2010134. URL <http://doi.acm.org/10.1145/2009916.2010134>.
- Mark D Smucker and Chandra Prakash Jethani. 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, Beijing*, volume 194, 2011b.
- Mark D. Smucker and Chandra Prakash Jethani. Time to judge relevance as an indicator of assessor error. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1153–1154, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348515. URL <http://doi.acm.org/10.1145/2348283.2348515>.
- Mark D Smucker, Gabriella Kazai, and Matthew Lease. Overview of the trec 2013 crowd-sourcing track. Technical report, TREC, 2013.
- Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL <http://dl.acm.org/citation.cfm?id=1613715.1613751>.
- MW van Someren, Yvonne F Barnard, Jacobijn AC Sandberg, et al. *The think aloud method: a practical approach to modelling cognitive processes*. Academic Press, 1994.
- Eero Sormunen. Liberal relevance criteria of trec: Counting on negligible documents? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 324–330. ACM, 2002.
- Amanda Spink, Howard Greisdorf, and Judy Bateman. From highly relevant to not relevant: examining different regions of relevance. *Information Processing & Management*, 34(5):599–621, 1998.
- Rong Tang and Paul Solomon. Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *Information processing & management*, 34(2):237–256, 1998.

- Arthur R Taylor, Colleen Cool, Nicholas J Belkin, and William J Amadio. Relationships between categories of relevance criteria and stage in task completion. *Information Processing & Management*, 43(4):1071–1084, 2007.
- Mario F Triola. *Elementary statistics*. Pearson/Addison-Wesley Reading, MA, 2006.
- Andrew Trotman and Dylan Jenkinson. IR evaluation using multiple assessors per topic. *Proc. of ADCS*, 2007.
- Pertti Vakkari. 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*, pages 2–9. ACM, 2000.
- Pertti Vakkari and Nanna Hakala. Changes in relevance criteria and problem stages in task performance. *Journal of documentation*, 56(5):540–562, 2000.
- Robert Villa and Martin Halvey. 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*, pages 765–768. ACM, 2013.
- Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 315–323, New York, NY, USA, 1998. ACM. ISBN 1-58113-015-5. doi: 10.1145/290941.291017. URL <http://doi.acm.org/10.1145/290941.291017>.
- Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000.
- Ellen M. Voorhees. The philosophy of information retrieval evaluation. In *Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, CLEF '01, pages 355–370, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-44042-9. URL <http://dl.acm.org/citation.cfm?id=648264.753539>.
- Ellen M Voorhees, Donna K Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005.
- Simon Wakeling, Martin Halvey, Robert Villa, and Laura Hasler. 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*, CHIIR '16, pages 173–182, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3751-9. doi: 10.1145/2854946.2854968. URL <http://doi.acm.org/10.1145/2854946.2854968>.
- Ronald E. Walpole. *Introduction to Statistics*. Macmillan Publishing Co., 1974.
- Jianqiang Wang. Accuracy, agreement, speed, and perceived difficulty of users relevance judgments for e-discovery. In *Proceedings of SIGIR Information Retrieval for E-Discovery Workshop*, page 1, 2011.
- Jianqiang Wang and Dagobert Soergel. A user study of relevance judgments for e-discovery. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10, 2010.
- Peiling Wang and Marilyn Domas White. A cognitive model of document use during a research project. study II. decisions at the reading and citing stages. *Journal of the American Society for Information Science*, 50(2):98–114, 1999.
- William Webber and Jeremy Pickens. Assessor disagreement and text classifier accuracy. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 929–932, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. doi: 10.1145/2484028.2484156. URL <http://doi.acm.org/10.1145/2484028.2484156>.
- William Webber, Praveen Chandar, and Ben Carterette. Alternative assessor disagreement and retrieval depth. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 125–134. ACM, 2012a.
- William Webber, Bryan Toth, and Marjorie Desamito. Effect of written instructions on assessor agreement. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1053–1054. ACM, 2012b.
- Wikipedia. Amazon mechanical turk — wikipedia, the free encyclopedia, 2016. URL https://en.wikipedia.org/w/index.php?title=Amazon_Mechanical_Turk&oldid=718450152. [Online; accessed 4-May-2016].
- HONG IRIS XIE and Colleen Cool. Online searching in transition: The importance of teaching interaction in library and information science education. *Journal of education for library and information science*, 39(4):323–331, 1998.

Yunjie Calvin Xu and Zhiwei Chen. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, 57(7):961–973, 2006.

Shu Ching Yang. Information seeking as problem-solving using a qualitative approach to uncover the novice learners' information-seeking processes in a perseus hypertext system. *Library & Information Science Research*, 19(1):71–94, 1997.

APPENDICES

Appendix A

Forms Used in the Studies

A.1 Think-Aloud Forms

A.1.1 Information and Consent Forms

Title of Project: Users Behaviour when Assessing Full Documents

Investigators: Aiman Al-Harbi, a2alharb@uwaterloo.ca,
Prof. Mark Smucker: mark.smucker@uwaterloo.ca

Summary of the Project:

This project is part of a research program aimed at improving information retrieval (text search) through interaction with the search user. In order to improve text search systems, we need to be able to better understand how people use these systems. A key part of this is that we know that everyone uses these systems in different ways and has different opinions about what is relevant for a given search topic.

This project will collect information about what is considered relevant for a search as well as the computer systems used while making these decisions. The information collected in this study may reveal valuable information about the decision-making process when evaluating full documents.

Procedure:

Your participation in this study is voluntary. Participation involves judging the relevance of full documents to a given search topic. In addition to completing several brief questionnaires, you will be asked to judge the relevancy of full documents to a given search topic for 4 topics. The questionnaires that you will be asked to complete consist of a demographic

questionnaire and a questionnaire concerning the search topic before each search topic task and a questionnaire about the task after each search topic task. You will be asked to think aloud when answering these questionnaires and during the judgement process as well.

To participate, you must be a native speaker of English and require no assistance with using a computer with a keyboard, mouse, and LCD monitor.

Participating in the study will take approximately 2 hours of your time. We will video record both your judgements and your interaction with the computer. We will also make note of and record anything we observe, including what you say, while you are participating in the study.

You may decline to answer any question that you prefer not to answer. You may stop participating in the study at any point and withdraw your consent without penalty.

Confidentiality and Data Security:

You will be issued an anonymous identifier (ID) as a participant in this study. The mapping from your 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 you to the data. All computer usage will be with University of Waterloo computers and not with personally computers, i.e. you will not use your own computer.

All data collected will be retained indefinitely and will be used for research purposes. We may refer to individual participants when describing the results or the study, and in these cases, we will always refer to “participant 1” or some other similar anonymous name. Your name will never appear in any publication that results from this study.

The document test collection that we use comes from the U.S. National Institute of Standards and Technology (NIST). This is a publicly available dataset. By our very use of this dataset, we will “link” with it, but we will not be linking your information collected here to any other information that concerns you personally.

We may choose to distribute the data collected to other researchers.

All data will be anonymized at the conclusion of the study and prior to any distribution, but each participants data will remain identifiable as coming from an individual, i.e. “participant 1”, “participant 2”, etc. We will not publicly share this data, i.e. the data would only be made available to other researchers for research purposes. Video recordings will be kept confidential, accessed only by the researchers, used only for analysis, and securely stored on a password protected computer.

Remuneration for Your Participation:

You will be paid \$20 for the whole study. Should you stop before completing the study, you will be paid on a pro-rated basis at a rate of \$5 per search topic completed.

The amount received is taxable. It is your responsibility to report the amount received for income tax purposes.

Risks and Benefits:

There is minimal risk to you from participation in this study. Computer use and searching for relevant documents 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 either major newswire services (Associate Press, etc.) or from U.S. governmental agencies.

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 Ethics Clearance:

We would like to assure you that this study has been reviewed and received ethics clearance through the Office of Research Ethics at the University of Waterloo. However, the final decision about participation is yours. Should you have comments or concerns resulting from your participation in this study, please contact Dr. Maureen Nummelin in the Office of Research Ethics at 519-888-4567, Ext. 36005 or maureen.nummelin@uwaterloo.ca. Thank you for your assistance in this project.

Consent Form:

I agree to participate in a study being conducted by Aiman Al-Harbi, a PhD student in the University of Waterloo's School of Computer Science and Dr. Mark Smucker, an assistant professor in the University of Waterloo's Department of Management Sciences. I have made this decision based 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 the study, I will be asked to complete several brief questionnaires and to judge the relevancy of full documents to a given search topic for 4 topics. Also, I will be asked to think aloud when answering these questionnaires and during the judgement process as well.

As a participant in this study, I am aware that I may decline to answer any question that I prefer not to answer. I am also 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 pro-rated basis at a rate of \$5.00 per search topic completed.

I am aware that all information that I provide will be anonymized with no identifiers retained to connect it to me.

I am aware that this study has been reviewed by, and received ethics clearance through, the Office of Research Ethics at the University of Waterloo, and that I may contact Dr. Maureen Nummelin at 519-888-4567 ext. 36005 or maureen.nummelin@uwaterloo.ca if I have any concerns or comments resulting from my participation in this study.

I agree to participate in this study

[Self-report, questionnaires, and relevance judgments (approximately 2 hours)]

YES NO (Please circle your choice)

A.1.2 Questionnaire

General Questionnaire:

1. What is your age?
2. Are you male or female?
3. If you are a student, are you:
 - An arts student.
 - A science, technology, engineering, or math student.
 - Other.
4. 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).
5. How much do you agree with the following statements? (strongly agree, agree, neutral, disagree, strongly disagree)
- I am an expert at finding information using search engines like Google, Yahoo, and Microsoft Bing.
 - Friends and family turn to me to help them search the internet for answers to their questions.
 - I enjoy using search engines like Google, Yahoo, and Microsoft Bing.
 - I consider myself a fast reader of web pages, magazines, and books. f. When Im in a group and a handout is given for us to read, Im one of the last to finish reading the handout.
 - Rarely (less than one search a month on average).
6. Have you ever had special training or education in searching or information retrieval? (yes/no) If yes, please describe the training or education.

Before Each Search 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)

After Each Search 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? (bored, engaged)
4. How hard was it to concentrate while judging the documents? (very hard, hard, neutral, easy, very easy)
5. Did you encounter any issues while completing this task? If yes, please describe.

A.2 Certainty Interfaces Study Forms

A.2.1 Information and Consent Forms

Title of Project: An Investigation into Relevance Assessing Behavior

Investigators: Aiman Al-Harbi, a2alharb@uwaterloo.ca,

Prof. Mark Smucker: mark.smucker@uwaterloo.ca

Summary of the Project:

This project is part of a research program aimed at improving information retrieval (text search) through interaction with the search user. In order to improve text search systems, we need to be able to better understand how people use these systems. A key part of this is that we know that everyone uses these systems in different ways and has different opinions about what is relevant for a given search topic.

This project will collect information about what is considered relevant for a search as well as the computer systems used while making these decisions. The information collected in this study may reveal valuable information about the decision-making process when evaluating full documents.

Procedure:

Your participation in this study is voluntary. Participation involves judging the relevance of documents to a given search topic. In addition to completing several brief questionnaires, you will be asked to judge the relevance of documents to a given search topic for 4 topics. The questionnaires that you will be asked to complete consist of a demographic questionnaire and a questionnaire concerning the search topic before each search topic task and a questionnaire about the task after each search topic task. We will record both your judgements and your interaction with the computer. We may also make note of and record anything we observe, including what you say, while you are participating in the study.

To participate, you must be a fluent in English and require no assistance with using a computer with a keyboard, mouse, and LCD monitor. Participating in the study will take approximately 1 hour of your time. You may decline to answer any question that you prefer not to answer. You may stop participating in the study at any point and withdraw your consent without penalty. Participation in the study requires that you follow some rules. The study's rules are as follows:

- You need to balance speed and accuracy. Please work as quickly as possible while making as few mistakes as possible. It is important to accurately judge the relevance of documents while being efficient in making your judgments.

- Some participants may finish before other participants. Please focus on your work and continue to judge documents as accurately and as quickly as possible.
- Please work on a given search topic task from start to finish. If you need to take a break, please do so between tasks. We will inform you when it is appropriate to take a break.
- Once you have made a judgment, do not attempt to go back and change your judgment. All judgments are final.
- This scientific research study requires your full attention. If you are unable to give this research your full attention, please excuse yourself from the study. In particular:
 - Please turn off your mobile phones. Phones may not be used during the study.
 - Please put all iPods and music players away. You may not listen to music during the study.
 - Do not use the computer for checking email, viewing web pages, or other activities during the study.
- If you use the computer for non-study activities or use a phone or music player, we will end your participation and ask you to leave.

Confidentiality and Data Security:

You will be issued an anonymous identifier (ID) as a participant in this study. The mapping from your 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 you to the data.

All computer usage will be with University of Waterloo computers and not with personal computers, i.e. you will not use your own computer. All data collected will be retained indefinitely and will be used for research purposes. We may refer to individual participants when describing the results or the study, and in these cases, we will always refer to “participant 1” or some other similar anonymous name. Your name will never appear in any publication that results from this study.

The document test collection that we use comes from the U.S. National Institute of Standards and Technology (NIST). This is a publicly available dataset. By our very use of this dataset, we will “link” with it, but we will not be linking your information collected here to any other information that concerns you personally.

We may choose to distribute the data collected to other researchers. All data will be anonymized at the conclusion of the study and prior to any distribution, but each participants data will remain identifiable as coming from an individual, i.e. “participant 1”, “participant 2”, etc. We will not publicly share this data, i.e. the data would only be made available to other researchers for research purposes.

Remuneration for Your Participation:

You will be paid \$12 for the whole study. Should you stop before completing the study, you will be paid on a pro-rated basis at a rate of \$3 per search topic completed. The amount received is taxable. It is your responsibility to report the amount received for income tax purposes.

Risks and Benefits:

There is minimal risk to you from participation in this study. Computer use and searching for relevant documents 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 either major newswire services (Associate Press, etc.) or from U.S. governmental agencies.

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 Ethics Clearance:

We would like to assure you that this study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee. However, the final decision about participation is yours. Should you have comments or concerns resulting from your participation in this study, please contact Dr. Maureen Nummelin in the Office of Research Ethics at 519-888-4567, Ext. 36005 or maureen.nummelin@uwaterloo.ca. Thank you for your assistance in this project.

Consent Form:

I agree to participate in a study being conducted by Aiman Al-Harbi, a PhD student in the University of Waterloos School of Computer Science and Dr. Mark Smucker, an associate professor in the University of Waterloos Department of Management Sciences. I have made this decision based 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 the study, I will be asked to complete several brief questionnaires and to judge the relevance of documents to a given search topic for 4 topics.

As a participant in this study, I am aware that I may decline to answer any question that I prefer not to answer. I am also 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 pro-rated basis at a rate of \$3 per search topic completed.

I am aware that all information that I provide will be anonymized with no identifiers retained to connect it to me.

I am aware that this study has been reviewed by, and received ethics clearance through a University of Waterloo Research Ethics Committee, and that I may contact Dr. Maureen Nummelin at 519-888-4567 ext. 36005 or maureen.nummelin@uwaterloo.ca if I have any concerns or comments resulting from my participation in this study.

I agree to participate in this study

[Self-report, questionnaires, and relevance judgments (approximately 1 hour)]

YES NO (Please circle your choice)

A.2.2 Questionnaire

General Questionnaire:

1. What is your age?
2. Are you male or female?
3. If you are a student, are you:
 - An arts student.
 - A science, technology, engineering, or math student.
 - Other.
4. 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).
5. How much do you agree with the following statements? (strongly agree, agree, neutral, disagree, strongly disagree)
- I am an expert at finding information using search engines like Google, Yahoo, and Microsoft Bing.
 - Friends and family turn to me to help them search the internet for answers to their questions.
 - I enjoy using search engines like Google, Yahoo, and Microsoft Bing.
 - I consider myself a fast reader of web pages, magazines, and books. f. When Im in a group and a handout is given for us to read, Im one of the last to finish reading the handout.
 - Rarely (less than one search a month on average).
6. Have you ever had special training or education in searching or information retrieval? (yes/no) If yes, please describe the training or education.

Before Each Search 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)

After Each Search 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 difficult was it to determine your level of certainty about your judgment? (very difficult, difficult, neutral, easy, very easy)

3. How would you rate your experience of judging the relevance of documents for this topic? (very unenjoyable, unenjoyable, neutral, enjoyable, very enjoyable)
4. How would you rate your mood while judging the documents? (bored, engaged)
5. How hard was it to concentrate while judging the documents? (very hard, hard, neutral, easy, very easy)
6. Did you encounter any issues while completing this task? If yes, please describe.