

Efficient User Interaction for High-Recall Retrieval: Model Priming

by

Abdul Manaam

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Sciences

Waterloo, Ontario, Canada, 2025

© Abdul Manaam 2025

Author's Declaration

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

High Recall Information Retrieval (HRIR) tasks, including legal e-discovery, information retrieval test collection construction and systematic review, require finding all, or nearly all relevant documents with the least amount of effort. Past research has shown that Technology Assisted Review (TAR) generally outperforms traditional e-discovery tools, and established that Continuous Active Learning (CAL) performs better than other commonly used TAR tools like Simple Active Learning and Simple Passive Learning. Prior research has also shown that adding search in a CAL-based HRIR tool can slow users down, and restricting access to full documents can speed up the document review process. Our goal was to design a system that provides more autonomy to users without affecting performance. Specifically, we wanted to investigate ways in which search can speed up the document review process. Systems like CAL often go through an initial training phase. We hypothesized that this training phase can be significantly shortened if search is used to seed the model. Moreover, we also created a novel interface that combines search with CAL on a single page. To test our hypothesis and the newly created user-interface, we conducted a user study with 40 participants to investigate five different configurations of an information retrieval system. We found that the addition of search, when preceded with proper user training, can significantly improve precision, performance, user experience and perceived effectiveness of the system. We also found that the newly designed interface, “Integrated CAL” performs comparably to the traditional interface, while providing a more familiar search based interface for users to interact with. Our findings reinforce the importance of hybrid High Recall Information Retrieval systems built on both search and CAL, that provide maximum control to users.

Acknowledgements

I would like to sincerely thank my supervisor, Dr. Mark D. Smucker, whose guidance, feedback, support, and expertise were irreplaceable throughout this research. This thesis would not have been possible without him.

I would like to thank my readers Dr. Charles L. A. Clarke and Dr. Olga Vechtomova for dedicating their time and effort to read and provide feedback on my thesis.

I would also like to thank my family and friends, who supported me through every step of this journey, offered invaluable advice, and lifted me up during the difficult times.

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (ALLRP 590219 - 23, RGPIN-2020-04665), Mitacs (IT38121), and Shinydocs, Inc. Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsors.

Dedication

For my parents, whose unwavering support, sacrifices, and belief in me have been the foundation of everything I have achieved.

Table of Contents

Author's Declaration	ii
Abstract	iii
Acknowledgements	iv
Dedication	v
List of Figures	x
List of Tables	xii
1 Introduction	1
1.1 Thesis Organization	5
2 Background on High Recall Information Retrieval (HRIR)	7
2.1 Pooling	8
2.2 E-discovery	8

2.2.1	Keyword search	9
2.2.2	Boolean Search	10
2.3	Technology Assisted Review (TAR)	11
2.3.1	Continuous Active Learning	12
2.3.2	Neural models in Continuous Active Learning	13
3	Related Work	15
3.1	Introduction	15
3.2	Methods	16
3.3	Results	17
4	Simulation	20
4.1	Introduction	20
4.2	Methods	21
4.2.1	Dataset and Topics	21
4.2.2	Implementation	21
4.2.3	Simulation Design	22
4.2.4	Evaluation Measures	22
4.3	Results	23
4.4	Discussion	25

5	User Study Methods	29
5.1	Test Collection	29
5.2	Participants	30
5.3	System	31
5.3.1	Search Engine Implementation	31
5.3.2	Paragraph CAL Implementation	32
5.4	User-Interface and configurations	34
5.5	User study procedure	40
5.5.1	Tutorial	40
5.6	Evaluation	46
5.6.1	Effort Curves	48
6	Results	49
6.1	Quantitative Analysis	49
6.1.1	Self-reported relevant documents	50
6.1.2	NIST marked relevant documents	53
6.1.3	Recall	55
6.1.4	Precision	58
6.1.5	Precision at minimum number of relevant documents	59
6.1.6	F1 score	59
6.1.7	Time to find relevant documents	61

6.1.8	Average number of search judgments made in the first five judgments	63
6.1.9	Switches per judgment	65
6.2	Usability Analysis	66
6.3	Qualitative Analysis	69
6.3.1	Most useful feature	70
6.3.2	Most liked	71
6.3.3	Most disliked	71
7	Conclusion	73
7.1	Summary of Contributions	77
	References	79
	APPENDICES	86
A	Forms	87
B	Tutorial	96
C	Poster	105

List of Figures

3.1	Architectural diagram for the system implemented by Abualsaud et al. , also included in their paper (Abualsaud et al., 2018).	16
4.1	Recall-effort curve for the simulation.	26
4.2	Total relevant docs found - effort curve for the simulation.	27
5.1	User-Interface for CAL.	34
5.2	User-Interface for search engine.	35
5.3	Modal to view full document.	36
5.4	User interface for CAL&SearchNudge once 5 relevant documents have been identified. The section highlighted in red shows the ‘ <i>Nudge</i> ’ message.	37
5.5	User-Interface for Search with Integrated CAL.	38
5.6	User-interface for iCALNudge once five relevant documents are found. The section marked in red shows the ‘ <i>Nudge</i> ’ message.	39
5.7	Website used to train participants on relevance judgments.	43

6.1	Average total number of self-reported relevant documents found by the participants.	52
6.2	Average total number of NIST-marked relevant documents found by the participants.	54
6.3	Recall-effort plot for all 5 configurations.	56
6.4	User found relevant documents over user found NIST relevant documents for the first 50 self reported relevant documents.	60
6.5	On a scale from 1 to 7, how would you rate the usability of this interface? .	67
6.6	How effective was this user interface in helping you find relevant documents?	68

List of Tables

1.1	2x2 factor design for the experiment and the available UI configurations.	4
3.1	2x2 factor design used in Zhang et al.'s experiment (Zhang et al., 2018).	17
3.2	Self-reported relevant documents found by users in Zhang et al.'s experiment.	18
4.1	Mean recall across 50 topics at various judgment cutoffs.	24
4.2	p-values for results from table 4.1	25
5.1	Algorithm used to generate query biased snippets.	32
5.2	Algorithm for C++ implementation of paragraph CAL by Abualsaud et al. based on AutoTAR (algorithm included in Zhang et al. 2018) (Abualsaud et al., 2018; Zhang et al., 2018).	33
6.1	Key/primer for reading Tables.	50
6.2	Self-reported relevant documents found by users.	51
6.3	NIST marked relevant documents found.	55
6.4	Recall.	57

6.5	Precision.	58
6.6	Precision at minimum number of self reported relevant documents. One user found 0 documents during one of their 1-hour sessions. Their precision was reported as 0 for that session for the calculation.	61
6.7	F1 score.	62
6.8	Time taken to find the first relevant document (in minutes).	62
6.9	Time taken to find 5 relevant documents (in minutes).	63
6.10	Average total number of search documents judged for the first five documents marked by users. The average for “ <i>iCALNudge</i> ” is 4.95 instead of 5 because one participant changed their judgment made through review tab (a page that shows past judgments).	64
6.11	Average total search documents judged in the first 10 judgments.	65
6.12	Total switches between search and CAL, normalized by total number of judgments ($totalSwitches/totalJudgments$).	66
A.1	Participant Responses to: <i>Which feature was the most useful in finding documents? Explain.</i>	87
A.2	Participant Responses to: <i>What did you like about the study?</i>	90
A.3	Participant Responses to: <i>What did you dislike about the study?</i>	93

Chapter 1

Introduction

In today's digital era, vast amounts of information exist in electronic form, making efficient retrieval of relevant documents from large collections a critical challenge. This is especially true in domains such as legal electronic discovery (e-discovery), systematic reviews, and the construction of Information Retrieval (IR) test collections. E-discovery involves identifying and retrieving electronic documents relevant to court proceedings, often in response to a request by the opposing party. Systematic reviews entail finding all available research for a certain topic to establish accurate scientific findings. Construction of IR test collections involves identifying all relevant documents in a test collection to evaluate and compare the performance of IR systems. Missing key documents can have significant consequences in every domain. Recall is the percentage of relevant documents found from the collection, and precision is the proportion of documents retrieved from the collection that are relevant. High-recall information retrieval systems aim to maximize recall while maintaining a high level of precision.

The quality of IR systems has been improving with time. In its early days, for

high recall information needs, exhaustive manual reviews were used, which meant going over every document in the collection. This was later followed by approaches like Boolean queries and keyword searches, which often resulted in the loss of a large number of relevant documents (Blair and Maron, 1985). Machine learning has increasingly been adopted for high-recall retrieval, with many law firms using Technology Assisted Review (TAR) for e-discovery. Despite its numerous variations, the process typically involves using a machine-learning model with labeled seed documents to classify a document collection. Cormack and Grossman established Continuous Active Learning (CAL) as the best-performing TAR protocol amongst the most commonly used ones in a controlled experiment (Cormack and Grossman, 2014). Later, topic and dataset-specific tuning parameters were removed from CAL, leading to the development of AutoTAR (Cormack and Grossman, 2015), a protocol that allows the document review process to begin with just a seed query. AutoTAR was used as the baseline in TREC Total Recall track that ran from 2015 to 2016 and consistently outperformed all other manual and automatic approaches for most of the topics (Roegiest et al., 2015; Grossman et al., 2016). The AutoTAR process starts with a seed query and is followed by an iterative feedback loop where the machine learning classifier scores all the documents in the collection that are not judged, the highest scoring document is then provided to the user, upon judging which, the classifier takes the relevance feedback into account. The process continues repeatedly where users are shown documents one by one until the desired recall level is reached.

Since CAL follows an iterative approach, users are limited to reviewing only the documents the system deems potentially relevant. Thus, if the system's notion of relevance is flawed, it can start recommending a series of non-relevant documents, leaving users with no option but to judge these documents. Abualsaud et al. introduced a system for High Recall Retrieval that included both search and CAL components (Abualsaud et al., 2018),

and was later used in the experiment conducted by [Zhang et al.](#) ([Zhang et al., 2018](#)). The addition of a search component mitigates this issue, allowing users to switch between components when needed. [Zhang et al.](#) in a controlled experiment compared the performance of CAL with and without the search tool available, however, contrary to the expectations, they found that users were able to find the highest number of relevant documents when using CAL by itself and the availability of search harmed performance. They suggested that carefully limiting the use of search to the early stages of a high-recall task might offer a performance benefit, but left this hypothesis untested. The goal of this thesis is to build upon the experiment by [Zhang et al.](#) and to investigate the aforementioned hypothesis, in order to create a system that seamlessly integrates search and CAL together while offering full flexibility to the users for the tools they want to use, without sacrificing performance.

Experiments conducted in the past have shown that CAL often has an initial training phase, where the first few returned documents are usually not relevant. However, once enough relevant documents have been marked, the classifier does an excellent job of finding most of the remaining documents. We hypothesize that if the model is primed with relevant documents in addition to a seed query, the training phase would finish earlier than expected, showing more relevant documents to the users and improving the overall experience.

Our initial simulations have shown that adding even one relevant document to the seed query before starting the document review process significantly shortens the model-building phase, providing more relevant documents to the user from the beginning. Our experiment aims to test a user interface that hides the CAL model from the users until they have found a few relevant documents through search.

For further improvements, the experiment aims to integrate CAL into search in a user-centered manner by designing an interface that aligns with users' existing search experiences. The system designed by [Abualsaud et al.](#) and later used in the experiment by [Zhang](#)

et al. used separate web pages for both search and CAL, requiring users to switch between interfaces (Abualsaud et al., 2018; Zhang et al., 2018). Hick’s law dictates that the more choices users have, the longer it will take them to make a decision (Hick, 1952). Thus, although the availability of search provides increased flexibility, it is often accompanied by an increased burden on the user, not only in terms of navigating different tools but also in deciding which tool to use and when to switch between them. This cognitive load can lead to inefficiencies, as users may spend additional time deciding what to do rather than focusing on the document review process.

By embedding CAL directly into the search, the goal is to create a seamless experience that minimizes disruption, reduces cognitive load, and allows users to focus on identifying relevant documents more efficiently. The proposed approach uses a CAL component as the top search result, ensuring that users encounter CAL documents first while maintaining access to a familiar ranked list of search results below. Thus, users can engage with CAL if the system provides useful recommendations. However, if the CAL results are unhelpful at any point, users can use search as they would in a traditional search system.

Base Case	CAL-Only	
CAL Disabled Initially	CAL types	
	Normal CAL	Integrated CAL
No	CAL&Search	iCAL
Yes	CAL&SearchNudge	iCALNudge

Table 1.1: 2x2 factor design for the experiment and the available UI configurations.

We hypothesize that manually seeding the Continuous Active Learning (CAL) model with five relevant documents from search before starting review will shorten the initial training phase for CAL, and will allow users to see more relevant documents quickly. We

further hypothesize that the new integrated CAL user interface will allow users to use both tools simultaneously, while reducing switching costs and providing a more familiar search-based user interface. To test these hypotheses, we conducted a controlled experiment with 40 participants using five different user interfaces to determine if the choice of the user interface has an effect, if any, on the time it takes a user to find relevant documents. Our experiment had a base case for comparison where the participants just had access to the CAL tool to understand the effects of removing search, and a 2x2 factor design to test our improvements (table 1). For one factor, we used a CAL component integrated into search, and the other factor constrained the user interface, so CAL was only available after the participants had found at least five relevant documents. The constrained user interface allowed us to test the effects of manually seeding the model, and the integrated CAL component allowed us to test the effects of a cleaner user interface, allowing for a quick switch between CAL and search. The experiment had five tasks, each associated with one user interface, assigned using Greco-Latin squares.

1.1 Thesis Organization

To test the effects of priming Continuous Active Learning model and to verify if changes in the user interface increase the rate of finding relevant information through an information retrieval system, we conducted a controlled user study with 40 participants, testing various configurations of the system.

In chapter 2, we cover the background for the field of High Recall Information Retrieval (HRIR). This chapter highlights the applications of HRIR across various domains and examines the tools commonly used.

This thesis builds on the experiment conducted by [Zhang et al.](#) and investigates an

untested hypothesis that they identified in their conclusion. Thus, in chapter 3, we go over the work of [Zhang et al.](#) in detail, highlighting their experiment design and results ([Zhang et al., 2018](#)).

Chapter 4, explains the design of the simulation that we created to observe the controlled effects of priming, and highlights the key results.

In Chapter 5, we detail the experimental design, including the rationale behind our design choices, the selection criteria for participants, the user interfaces implemented, and the tools utilized to conduct the study. We also discuss the evaluation measures used to analyze and compare our results.

In chapter 6, we present the results from our experiment while showing the performance under different configurations of the system across various evaluation measures. The chapter also compares the results from the user-study, with the results of simulation, and with the results from [Zhang et al.](#)'s experiment. Lastly, we present a qualitative analysis of the system for a more comprehensive evaluation.

In chapter 7, we further analyze the results, highlighting the main findings from the study and the implications for the future of development of HRIR systems.

Chapter 2

Background on High Recall Information Retrieval (HRIR)

In Information Retrieval (IR), recall is the fraction of relevant documents retrieved from a collection relative to the total number of relevant documents. High Recall Information Retrieval (HRIR) is a problem in information retrieval where the goal is to find all, or nearly all, relevant documents related to a topic with the least effort. HRIR tools are often used in applications like e-discovery, systematic review, and the construction of IR test collections. E-discovery consists of electronically stored documents that can be used as evidence in legal proceedings (Oard et al., 2010). Systematic review entails collecting and summarizing all information regarding a specific research topic. The construction of information retrieval test collections involves finding all relevant documents for a variety of topics. For all these processes, finding all relevant documents is highly important, as otherwise, we may risk not finding critical evidence in e-discovery, compiling a systematic review with incomplete information, and building a test collection where the majority of

relevant documents are not found and marked, leading to non-representative evaluating measures. There are a variety of information retrieval tools available that can be used for HRIR.

2.1 Pooling

Pooling is an important process in HRIR, usually through which test collections are constructed for the evaluation of information retrieval systems. The most prominent example of this approach is the one used by NIST. For their Text Retrieval Conference, NIST sends a document collection and a set of queries to the participating teams. The teams then run their retrieval algorithm on the queries and send a ranked list back to NIST for each query (Voorhees, 2002). NIST forms a pool of documents from these ranked lists by going till a certain depth for each list. These documents are then judged by assessors, and this method has proven to be reliable but does not always work for recall-sensitive tasks. Zobel estimated that even with a depth of 100, pooling likely only finds about 50%-70% of the relevant documents (Zobel, 1998). Yilmaz et al. discovered that test-collections created through shallow depth pooling (depth-10) using traditional systems like BM-25 often result in unreliable evaluation measures when evaluating systems built on other methods e.g. neural-based systems, highlighting the shortcomings of shallow-depth pooling in finding relevant documents (Yilmaz et al., 2020).

2.2 E-discovery

In the legal field, HRIR is also an extremely important tool for e-discovery. During legal proceedings, parties are required to share all information potentially relevant to the issue;

this information is also known as discovery. As more information becomes digital than ever before, most of these documents are stored electronically (hence e-discovery) (Roitblat et al., 2010). Although exhaustive manual review strategies that rely on reading all documents in the collection are still widely used, they are far from perfect. They can be extremely time-consuming and can result in significant discovery costs for clients. Review costs can range from two hundred and fifty to as much as five hundred dollars per hour (Tingen, 2012-2013). With increasing document collection sizes, the cost to extract relevant documents also increases and in some cases can be as high as 4-8 million dollars (Bace, 2007). Thus, various information retrieval tools are utilized to reduce the total number of documents to be reviewed and speed up the review process.

2.2.1 Keyword search

Keyword search is the most common tool available that is utilized in e-discovery. The process involves running queries on a collection to find relevant documents. However, it is limited to finding documents where there are exact keyword matches with the query. Thus, makes it difficult to find documents if the relevant documents do not contain the keywords searched for, or when the user is not adequately able to express their information need (Tingen, 2012-2013). As a result, keyword searches often carry the risk of missing relevant documents, especially if it's difficult to predict the kind of words a relevant document might have. Moreover, the presence of keywords in a document also does not necessarily equate to relevance, as the context of words is not considered, and therefore keywords with multiple meanings can result in many false positives (Baron et al., 2007). For example, the word "Novel" can be a book, or it can be something newly discovered, and "Apple" can be either the fruit or the company. Fuzzy search is an effective solution that overcomes

certain limitations of keyword search (Tingen, 2012-2013). Although keyword search relies on strict matching, fuzzy search can also retrieve misspelled words or different spelling variants than the ones that were searched for. Despite this, users still have to come up with good queries using words that are used in the document collection to achieve good recall. In a famous experiment by Blair and Maron, participants who were lawyers and paralegals used a keyword search-based information retrieval system and were instructed to stop only when they were satisfied that they had found at least 75% of the relevant documents from the collection (Blair and Maron, 1985). They stopped after finding only about 20% of the relevant documents, highlighting the gap between perceived and the actual effectiveness of keyword searches. Blair and Maron concluded that the results were primarily caused by the variety of words used in the relevant documents, that were difficult to predict. They further noted that, even if participants were aware they had only identified a fraction of the documents, they would still have struggled to locate additional ones due to the difficulty of predicting the specific words used in relevant documents. Thus, keyword search relies heavily on knowing what relevant documents might look like, raising questions about its reliability.

2.2.2 Boolean Search

Boolean search is another useful tool that is frequently utilized to quickly shortlist the number of documents. Roitblat et al. discuss a query used in the US vs. Philip Morris case containing more than a hundred words, highlighting that making these queries can often be difficult (Roitblat et al., 2010). Salton et al. report that constructing good boolean queries is difficult for ordinary people, and even professional searchers struggle to form consistently good queries (Salton et al., 1983). Cooper highlights four major problems

with the Boolean language (Cooper, 1983):

- It is very difficult to adapt, and novices often confuse ‘AND’ and ‘OR’ operations (there is a misconception that searching for A AND B would produce more results than searching for just A).
- Boolean queries are difficult to formulate correctly in the first go. Usually, the first search query yields too many or too few results, and if an ‘AND’ is used incorrectly, the total number of results can even be 0. This can be problematic if the query has to be provided beforehand to the opposing party for discovery.
- The set of results returned from a Boolean search is unranked, and if the result set is extremely large, it can take considerable time to go over each document.
- The set of results is limited by the user’s query. Although it is possible to retrieve all the relevant information using a Boolean query, users are often not willing to formulate long queries that can find essentially all the information available. Similar to keyword searches, it is also usually impossible to predict the words relevant documents may contain.

2.3 Technology Assisted Review (TAR)

Technology-assisted review (TAR) involves collaboration between humans and computers to identify relevant documents. With TAR, instead of reviewing the whole document collection, humans first review a small set of seed documents, and then usually a machine learning model tries to identify the rest; humans then review the documents that the machine learning model returns (Grossman and Cormack, 2014). Grossman and Cormack,

in their experiment, found that TAR on average required human review of only 1.9% of the total documents compared to a manual review (Grossman and Cormack, 2010), reducing costs considerably. Blair and Maron have shown the shortcomings of keyword-based search systems, and thus machine learning based systems offer a more sophisticated and promising approach by eliminating the need to predict keywords. Early research by Cormack and Mojdeh used a combination of interactive search with active learning to find relevant documents (Cormack and Mojdeh, 2009). TAR was accepted as a suitable method for e-discovery as early as 2012 in Moore v. Publicis Groupe where the court approved the use of TAR as a reasonable means of conducting e-discovery, noting that technology-assisted review can be more accurate and error-prone (Da Silva Moore v. Publicis Groupe, 2012). In a study conducted by Grossman and Cormack, to establish the effectiveness of TAR over exhaustive manual review, TAR demonstrated significantly higher levels of precision for all topics tested, while achieving a higher recall for the majority of the topics (Grossman and Cormack, 2010).

2.3.1 Continuous Active Learning

Continuous Active Learning, a TAR protocol, used by Cormack and Mojdeh (Cormack and Mojdeh, 2009), and by Grossman and Cormack (Grossman and Cormack, 2010), demonstrated superior performance to traditional methods like boolean search and keyword search. Later, Cormack and Grossman conducted a study evaluating the efficiency of commonly used TAR protocols: Simple Active Learning (SAL), Simple Passive Learning (SPL), and Continuous Active Learning (CAL) (Cormack and Grossman, 2014). Although SAL had shown prior success in numerous TREC tasks, and both SAL and SPL protocols were commonly used in the legal marketplace for e-discovery, their performance was

not formally evaluated in a controlled setting. The experiment demonstrated that CAL performance was significantly superior to SPL and outperformed SAL for all four legal and four TREC topics tested. Later, to allow easier configuration and to run CAL with a wider number of topics and test collections, [Cormack and Grossman](#) removed topic and dataset-specific tuning parameters from CAL, allowing users to start CAL with just a seed query ([Cormack and Grossman, 2015](#)). This was called AutoTAR, and it showed superior performance to CAL and other TAR protocols like SPL and SAL across a variety of test collections. AutoTAR was used as the baseline in TREC Total Recall track that ran from 2015 to 2016 and consistently outperformed all approaches, even the ones utilizing manual seeding ([Roegiest et al., 2015](#); [Grossman et al., 2016](#)).

The AutoTAR process starts with a seed query and is followed by an iterative feedback loop where the user is shown a document, upon judging which, the tool takes feedback into account and the system shows the next best document. The process continues repeatedly, where users are shown documents one by one until the desired recall level is reached.

2.3.2 Neural models in Continuous Active Learning

Although, AutoTAR BMI uses logistic regression (linear model) classifier to find documents, the use of neural models for technology assisted review has also been explored in the past. [Yang et al.](#), in their experiment, used pre-trained BERT transformer for TAR on two corpora, and compared the performance against a logistic regression baseline ([Yang et al., 2022](#)). [Yang et al.](#) found that the review costs can be decreased by 10%-15% when TAR process is done on a corpus similar to which BERT was trained on, however, linear models still outperformed BERT on a collection with different textual characteristics. Moreover, they also found that the performance is also dependent on the amount of

fine-tuning, and both too little or too much fine-tuning can hinder performance. Lastly, while comparing the computational costs, the researchers reported exponential differences, whereby the time taken to train and score the collection using logistic regression, on average, took 20 to 25 seconds, and with BERT it took 18 hours. [Sadri and Cormack](#) used BERT and T5 transformers in a variety of CAL based settings but did not find statistically significant performance differences compared to a logistic regression model with TF-IDF vectorizer ([Sadri and Cormack, 2022](#)). [Mao et al.](#) conducted a study to reproduce and build on the experiment by [Yang et al.](#), and confirmed the general findings of the previous study ([Mao et al., 2024](#)). [Mao et al.](#) also found that if the right fine-tuning epoch is identified, BERT can outperform baseline even for collections with textual characteristics different from those it was trained on. They further found that selecting the appropriate domain-specific pre-trained BERT model to match the collection can lead to results that outperform both linear models and base BERT. Although the findings highlight an optimistic view of integrating neural models for HRIR, the computational costs are still significantly more, and may not be feasible in all cases.

Chapter 3

Related Work

3.1 Introduction

This thesis directly builds on the work done by [Zhang et al.](#), “Effective User Interaction for High-Recall Retrieval: Less is More”, who investigated a hybrid information retrieval system with both search and a CAL component. [Zhang et al.](#) found that giving users access to search and the ability to view full documents harmed performance, and proposed a more restrictive interface. However, they acknowledged that there may be situations in which users would like to see full documents and hypothesized that in these situations, strategic use of search may offer performance benefits. Nevertheless, they left this hypothesis untested ([Zhang et al., 2018](#)). This thesis fills the gap by investigating whether and how users can effectively combine search and CAL tools. This chapter, therefore, goes over [Zhang et al.](#)’s experiment, highlighting the methods and the results.

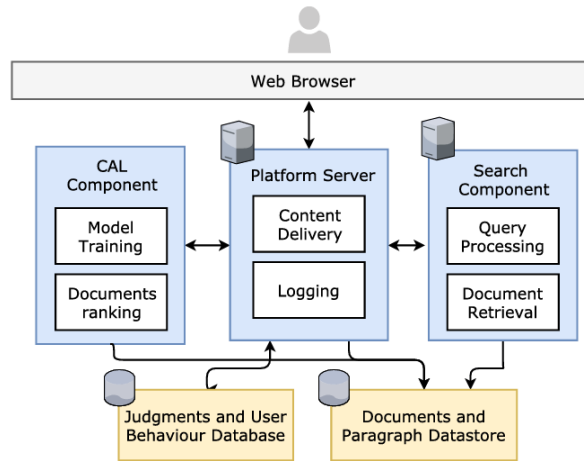


Figure 3.1: Architectural diagram for the system implemented by [Abualsaud et al.](#), also included in their paper ([Abualsaud et al., 2018](#)).

3.2 Methods

[Zhang et al.](#) conducted a user study with 50 participants to test the effectiveness of different user interfaces for an information retrieval system developed by [Abualsaud et al.](#) ([Abualsaud et al., 2018](#); [Zhang et al., 2018](#)). The experimenters used the test collection from TREC 2017 Common Core Track ([Allan et al., 2017](#)) and created a balanced study design using Greco-Latin squares.

Figure 3.1, highlights the architecture for the system developed by [Abualsaud et al.](#) ([Abualsaud et al., 2018](#)) (The diagram is also included in their paper). The system had a BM25 ([Robertson et al., 1994](#)) search and a CAL component, along with a platform server for presenting documents and logging responses. They implemented a C++ version of Baseline Model Implementation of AutoTAR for their high recall information retrieval system, allowing CAL to be used as a web service. The system also included a paragraph-

based CAL model that ranked paragraphs instead of full documents, and was the model mainly used in their study.

For the search engine, the New York Times Annotated Corpus (Sandhaus, 2008) was parsed using the provided java NYTCorpusDocumentParser class and Indri was used for indexing (Strohman et al., 2005), using a Krovetz stemmer and the default parameters for k1 and b. Query biased snippets were also utilized and the two top scoring sentences were concatenated together, truncated to 75 words and then shown.

Search Available	CAL types	
	Full document Available	Paragraph Excerpt only
No	CAL-D	CAL-P
Yes	CAL-D&Search	CAL-P&Search

Table 3.1: 2x2 factor design used in Zhang et al.’s experiment (Zhang et al., 2018).

Zhang et al. used a 2x2 factorial design with the availability of search one factor and the availability of full document as the other factor. Table 3.2, summarizes the four configurations and the 2x2 factorial design. For CAL, Zhang et al. used the paragraph based CAL model as described earlier for both configurations; its algorithm can be seen in figure 5.2. The only difference between CAL types was that when the full document was available, users, when presented the paragraph, could press a button and view the full document.

3.3 Results

Table 3.2 shows the main result from Zhang et al.’s study. It can be observed that allowing users to read full documents in CAL, and giving them access to search significantly

Search Available	CAL types		Marginal means (Search)	
	Full document available	Paragraph Excerpt only		
No	58.3	97.9	78.1*	p value < 0.001
Yes	51.4	65.4	58.4	
Marginal means (CAL types)	54.8	81.6*	68.2	
	p value < 0.001			

Table 3.2: Self-reported relevant documents found by users in [Zhang et al.](#)'s experiment.

impacted performance. They found that the decrease was primarily due to the rate at which participants judged documents under different configurations. Thus, the authors concluded that users could find more relevant documents when only the paragraph excerpt was available. They further found that search had little to no impact on performance measures like recall and F1 scores, and although they found that search improves precision, when they accounted for the variances in topics by finding ‘precision at k ’, where k is the minimum number of relevant documents found for a topic, the effects turned out to be non-significant. While the ability to view the full document made no difference in the value for precision at k , p -value for search was still relatively low (0.091) with the configurations that included search still being better. Therefore, the paper hypothesizes that perhaps a compromise could work wherein people could leverage the ability of search to find a few relevant documents quickly and then switch to CAL for the remainder of the review process. This thesis investigates a variant of the aforementioned hypothesis.

Using search before starting document review in CAL is not a new concept. For the TREC Total Recall track of 2015 ([Roegiest et al., 2015](#)), [Pickens et al.](#) showed that three reviewers, using different search strategies to manually seed CAL, were able to establish high recall ([Pickens et al., 2015](#)). For the TREC Total Recall Track of 2016 ([Grossman](#)

et al., 2016), Pickens et al. explored one-shot (single-query) versus iterative (multi-query) approaches for manual seeding (Pickens et al., 2016). Although they observed some differences in performance between different querying strategies on different topics, reviewers in their experiment were able to achieve high recall regardless of the querying strategy. Nevertheless, a comparative user-centered analysis on the usage of search alongside CAL, in controlled settings, has not been explored earlier.

In summary, Zhang et al. concluded that the inclusion of search in the interfaces resulted in a decrease in performance, thus, recommending an interface design that has no search, and where only the paragraph excerpts from documents are available. However, being restrictive, this interface takes freedom away from the users. Therefore, for future work, the authors indicated users may prefer greater control over the tools they would like to use for information retrieval, and pointed out that usage of search early in the retrieval task may speed up the process. Our experiment builds up on this hypothesis to investigate ways in which a more efficient hybrid information retrieval system can be designed that includes both search and CAL.

Chapter 4

Simulation

4.1 Introduction

Continuous Active Learning (CAL) often struggles with a learning phase where the initial documents returned by the model are often non-relevant. During this phase, the users might have to review a lot of non-relevant documents, whereas they could've gotten immediate success if they just used search. We believe that the model training phase can be shortened significantly if the users find few relevant documents using search, then switch to CAL for the remaining review process.

We carried out a simulation to investigate the effect of priming the CAL model with relevant documents, found through BM25 ([Robertson et al., 1994](#)) search, upfront before initiating the CAL review process. Through this simulation, we aim to understand the effect of priming by removing factors like user variability or interface limitations.

4.2 Methods

4.2.1 Dataset and Topics

We used the TREC 2017 Common Core Track dataset, containing approximately 1.8 million news-wire articles from the New York Times (Allan et al., 2017). For topics and relevance labels, we used the 50 NIST-assessed topics that were included in the collection.

4.2.2 Implementation

We indexed the test collection using Pyserini (Lin et al., 2021), a Python-based information retrieval toolkit, extracting each document’s ID, title, date, and content. We also used the BM-25 (Robertson et al., 1994) ranker included in the Pyserini retrieval toolkit, with the k_1 parameter set to 1.2 and b to 0.75, similar to Zhang et al. (Zhang et al., 2018; Lin et al., 2021). For the simulation, we did not use a stemmer.

For CAL, we used the same version of BMI of AutoTAR implemented in C++, as the one used by Zhang et al.. Since relevance labels were just available for documents, and not the individual paragraphs, we ran a document level Continuous Active Learning model. The documents from the test collection were loaded into CAL’s memory and for each topic, a session was started using both the topic title and the topic description as the seed query.

To interact with both the search engine and the CAL, we wrote a Python script. We used the built-in Pyserini functions for document ranking, qrels that came with the Common Core Track for determining relevance, and API calls to CAL web service to create sessions, send judgments and get recommended documents etc.

4.2.3 Simulation Design

We designed a two-pronged simulation to investigate our hypothesis. For the first phase, for each topic, the system searched for documents through Pyserini’s BM25 ranker using the topic title as the query. It then went down the ranked list until x number of relevant documents were found (where x is between 0 and 5). For each of these documents, the system determined relevance using the qrels from test collection. Any document not present in the qrels was considered non-relevant. The relevant and the non-relevant documents marked during this process were subsequently fed to CAL.

For the second phase, once all the required search judgments were made, the system then ran in a loop. For each iteration, the system fetched the document recommendation from CAL, judged it using the qrels, and then sent the judgment back to CAL. The loop was completed once a total of five hundred documents were reviewed by using both search and CAL. The counter for search documents was included to ensure that the effort spent on both search and CAL was counted in our final calculations.

For the simulation, to compare effectiveness of CAL as compared to normal search, we also observed a run where only BM25 search was used. For this, the system started with using the topic title as the query, and then went down the ranked list until 500 documents were seen.

4.2.4 Evaluation Measures

To evaluate the effectiveness of priming, we will use the following evaluation measures:

Recall@rank: Recall is the fraction of relevant documents found over the total number of known relevant documents ($Recall = |NIST.Rel|/|R|$, where NIST.Rel is the set of

documents found that were marked as relevant by NIST, and R is the set of all relevant documents in the collection), and is the measure commonly used to determine the percentage of relevant documents found. Measuring recall at various ranks will allow us to observe the effects of priming deeply at numerous points. For our experiment, we measured recall rank 50, 100, 250, and 500.

Recall-Effort curve: Recall-effort curve is one of the most common evaluation metric used in numerous prior publications (Cormack and Grossman, 2015, 2014; Zhang et al., 2020; Roegiest et al., 2015; Grossman et al., 2016). The curve plots average recall across all 50 topics for each of the priming levels, showing detailed trends. In this context, *effort* refers to the number of documents viewed (i.e., judged) by the system and thus, each additional judgment represents one unit of effort.

Relevant Documents found-Effort curve: Similar to recall-effort curve, the curve plots the average total number of relevant documents found at each effort level across all 50 topics, providing an additional perspective.

We used two-tailed paired student’s t-tests to determine whether the effects of priming are statistically significant. The per-topic Recall@rank results for when no priming was done were compared with the configurations where the model was primed with initial documents.

4.3 Results

Results in table 4.1 depict average recall across all 50 NIST-assessed topics at different effort levels (50, 100, 250, and 500). It can be clearly observed that finding and marking even a few relevant documents before starting up the document review process using CAL

can significantly improve recall, especially at lower efforts. For example, finding just one relevant document using search improves the average recall@50 by 62.5% (from 0.040 to 0.065). Moreover, we can also observe that the effect somewhat stays consistent once 3-5 relevant documents have been found.

Priming Level	Recall@50	Recall@100	Recall@250	Recall@500
No priming	0.040	0.090	0.218	0.311
1 doc	0.065	0.130	0.251	0.327
2 docs	0.065	0.133	0.249	0.320
3 docs	0.068	0.126	0.255	0.324
4 docs	0.064	0.123	0.248	0.322
5 docs	0.066	0.122	0.242	0.318
BM25 search	0.070	0.103	0.159	0.201

Table 4.1: Mean recall across 50 topics at various judgment cutoffs.

Table 4.2 shows the p-values from two-tailed paired t-test, highlighting the statistical significance of the increase in recall. It can be clearly observed that at lower levels of effort, the improvements with each level of priming is both incremental, and statistically significant ($p < 0.05$). However, at higher levels of effort, these effects start to fade away, with improvements in recall being small and the results being not significant at $p < 0.05$.

Figure 4.1 and figure 4.2, depict curves for recall and relevant documents found per each unit of effort, respectively. The legend on the curves highlight the number of documents used for priming (i.e., 0, 1, 2 etc.). There is also an additional curve for BM25 search. Both plots show similar trends; BM25 search alone outperforms CAL, both with and without priming, during the initial phase (roughly the first 50 document judgments). However, its advantage quickly diminishes, and the curve flattens thereafter. In contrast, the CAL-based

Priming Level	Recall@50	Recall@100	Recall@250	Recall@500
1 doc	0.000	0.000	0.000	0.064
2 docs	0.000	0.000	0.013	0.367
3 docs	0.000	0.000	0.000	0.171
4 docs	0.000	0.000	0.003	0.270
5 docs	0.000	0.000	0.013	0.449
BM25 search	0.001	0.003	0.000	0.000

Table 4.2: p-values for results from table 4.1

strategies continue to improve, surpassing BM25 in recall as more documents are reviewed. Furthermore, we can also observe that CAL without priming shows a lot of non-relevant documents in the beginning, and priming CAL with even just one document significantly shortens the learning phase. Lastly, we observe that the effects of priming are relatively consistent across different priming levels, with only slight differences in performance.

4.4 Discussion

The results show that priming CAL with relevant documents improves performance early on, and priming even with just one document helps in reaching higher levels of recall early. We can also observe that the improvement is a lot more noticeable and evident in the early stages, and the effects starts to fade away at higher levels of recall. The results suggest that a modest amount of priming can help with model building and improve the user experience, as users would have to view less non-relevant documents in the beginning.

The results provide evidence to pursue a priming based user-interface design through search, and is the basis for our Nudge based interface in the user study. The interface

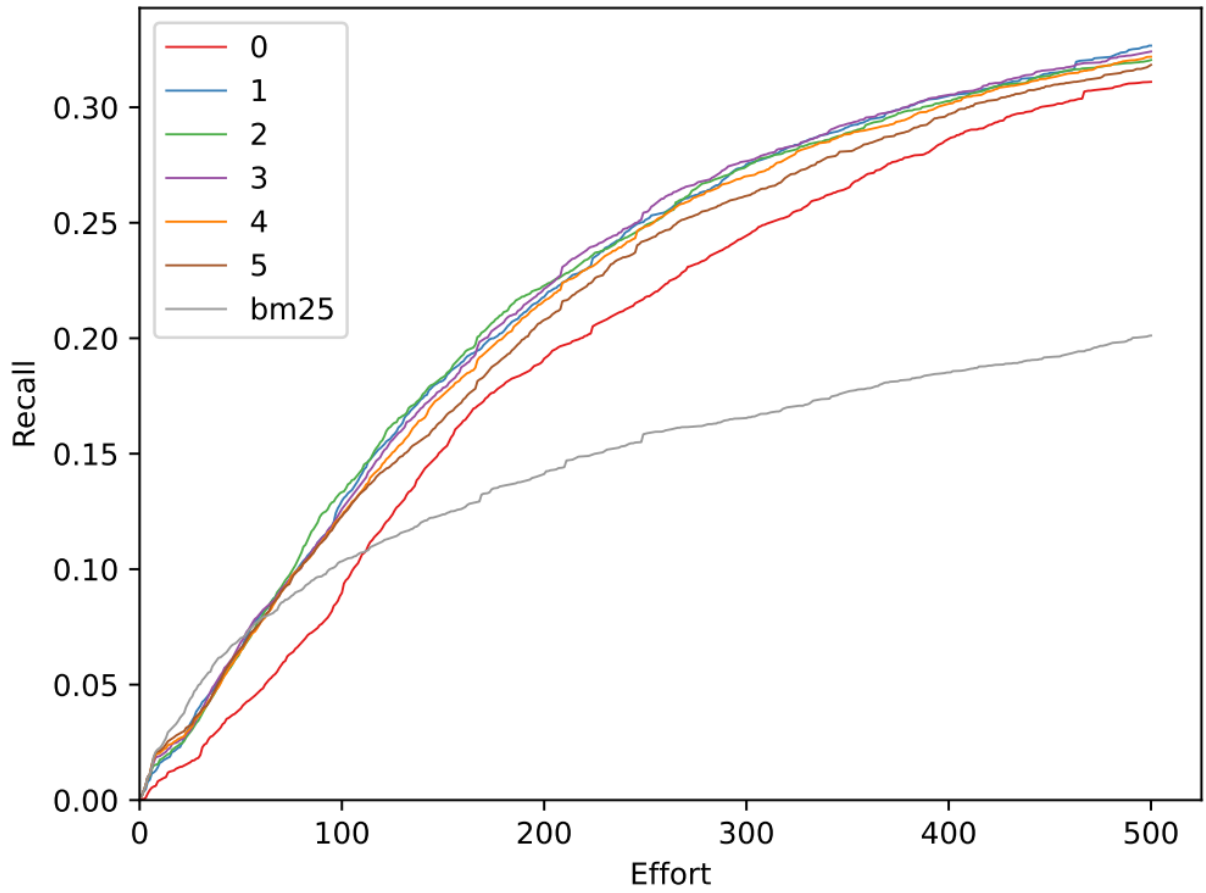


Figure 4.1: Recall-effort curve for the simulation.

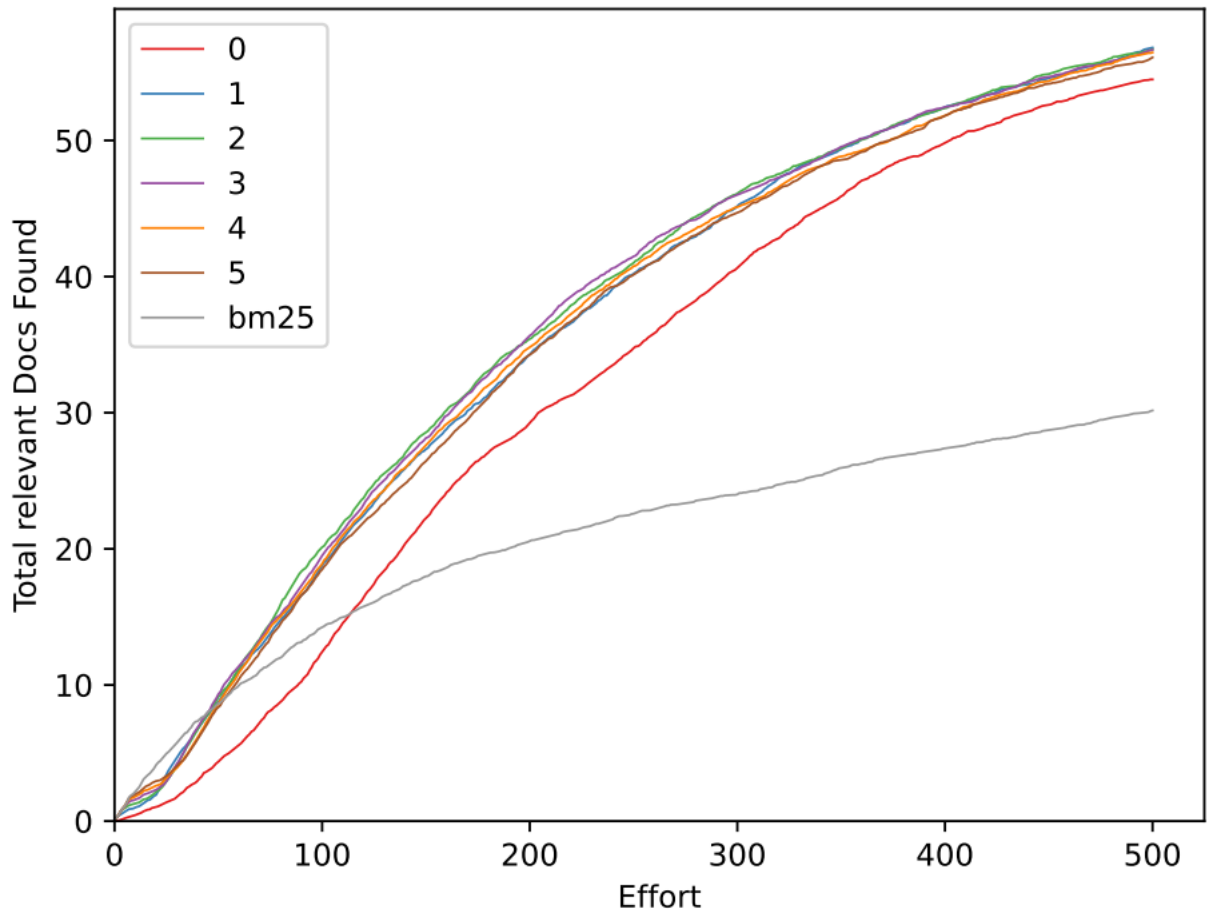


Figure 4.2: Total relevant docs found - effort curve for the simulation.

requires users to find a small number of judgments via search before engaging CAL. When we first formed this simulation analysis, there was an error in our initial analysis that had led us to believe that priming CAL with 5 documents would yield the best performance. Later, when the analysis was fixed, we did not find any significant difference between priming with either 1, 2, 3, 4, or 5 documents, meaning just finding one relevant document from search, before switching to CAL would have been sufficient. However, we had already decided on priming with 5 documents before we found the error in our initial analysis. The results also show that if the information need is to just find a few relevant documents, using search alone could provide a better return on effort. These insights provide the foundations for designing more effective and user-friendly high-recall retrieval systems that seamlessly combine both search and CAL.

Chapter 5

User Study Methods

5.1 Test Collection

For this experiment, we used the test collection from TREC 2017 Common Core Track (Allan et al., 2017). The collection contains 1.8 million newswire documents from the New York Times written between January 1st, 1987, and June 19th, 2007. We used the paragraph-based test collection created by Zhang et al.. Zhang et al. split the documents in the collection into paragraphs by extracting the contents between the `<p></p>` tags of documents. The collection contained about 30 million paragraphs, with an average of 16.7 paragraphs per document. The track also contains the updated version of 250 topics from TREC 2004 Robust Retrieval Track (Voorhees, 2004). Although it comes with crowdsourced judgments for all 250 topics, only 50 of the 250 topics were judged by NIST assessors. Thus, for our experiment, we used only the 50 topics with NIST relevance judgments.

5.2 Participants

A total of 40 individuals participated in the experiment. Of these, 21 participants identified as women (52.5%), 18 as men (45%), and 1 as transgender (2.5%). Their ages ranged from 18 to 37 years, with a mean age of 22.8 years. Moreover, 75% of the participants were under 25 years old, with 39 out of the 40 participants currently being students. Additionally, 14 participants (35.9%) had already completed a bachelor's degree, 14 participants (35.9%) were high school graduates, 7 participants (17.9%) held a Master's degree, 3 participants (7.7%) had completed some college credit (with no degree received), and 1 participant (2.6%) had obtained a doctoral degree. The participants came from various disciplines. 11 participants were from Management Science and Engineering, 10 from other engineering fields, 7 from Earth and Environmental Sciences, 5 from Health and Life Sciences, 3 from Computer Science, and 3 from Mathematics and Physical Sciences.

The participants also had diverse ethnic backgrounds, with the composition being as follows:

- South Asian (e.g. East Indian, Pakistani, Sri Lankan, Indo-Caribbean, etc.): 22 participants (55%)
- East Asian (Chinese, Korean, Japanese, Taiwanese descent): 13 participants (32.5%)
- Middle Eastern (Arab, Persian, West Asian descent, e.g. Afghan, Egyptian, Iranian, etc.): 3 participants (7.5%)
- Black (African, Afro-Caribbean, African-Canadian descent): 1 participant (2.5%)
- Southeast Asian (Filipino, Vietnamese, Cambodian, Thai, other Southeast Asian descent): 1 participant (2.5%)

5.3 System

Our experiment utilized a modified version of the information retrieval system that was used by the UWaterlooMDS team at the TREC 2021 Health Misinformation Track (Abualsaud et al., 2021; UWaterlooIR), and was very similar to the one used in the experiment by Zhang et al. (Zhang et al., 2018). Although there might have been small variations, We tried our best to provide a similar user-interface and features to the participants. The system contains two main tools: a paragraph-based Continuous Active Learning model (CAL) and a BM-25 based search engine. The system architecture was the same as Abualsaud et al.’s system and can be seen in figure 3.1.

5.3.1 Search Engine Implementation

We indexed the test collection using Pyserini (Lin et al., 2021), extracting each document’s ID, title, date, and content. We also used Pyserini’s BM25 (Robertson et al., 1994) ranker, with the k1 parameter set to 1.2 and b to 0.75. The parameters were set to match those used by Zhang et al. (Zhang et al., 2018). For further consistency with the prior experiment, we also used the Krovetz stemmer for the search. Although we did not use Indri (Strohman et al., 2005) as Zhang et al. did, we believe its impact on results will be insignificant as all other conditions were kept consistent with the experiment.

For our search snippets, we showed the two top-scoring sentences, separated by ellipses, trimmed to approximately the first 75 words. As shown in Table 5.1, we used a slightly modified version of the sentence scoring algorithm by Turpin et al. (Turpin et al., 2007). To highlight query terms within the snippet, we stemmed all query and snippet tokens using the Krovetz stemmer. If a stemmed snippet token matched a stemmed query token, we bolded the original (unstemmed) form of the snippet token in the displayed snippet.

Query-Biased Snippet Generation Algorithm

For each sentence in the document:

- a) Tokenize the sentence into individual terms.
 - b) Assign an initial score based on the sentence position:
 - First sentence: +2 points
 - Second sentence: +1 point
 - c) Count the occurrences of query tokens in the sentence (c).
 - d) Count the distinct query tokens present in the sentence (d).
 - e) If all query tokens appear in the sentence, add a bonus of +5 to d .
 - f) Compute the final sentence score as: $total_score = c + d + l$.
-

Table 5.1: Algorithm used to generate query biased snippets.

5.3.2 Paragraph CAL Implementation

The system uses the Baseline Model Implementation (BMI) software utilizing AutoTAR Continuous Active (Cormack and Grossman, 2015) method, also used in TREC Total Recall 2015 track (Roegiest et al., 2015). We used the same implementation of paragraph-based CAL as the one utilized by Abualsaud et al. (Abualsaud et al., 2018). The details of the algorithm are highlighted in table 5.2. As explained earlier (section 5.1), we split documents from test collection into paragraphs by extracting content from the `<p></p>` tags. The paragraphs were then loaded into memory and scored according to the algorithm in table 5.2.

Algorithm for Paragraph CAL

Step 1: Use the combination of topic title and topic description as a relevant document and add it to the training set.

Step 2: Temporarily augment the training set with 100 random documents from the corpus, assuming their label as “non-relevant”.

Step 3: Train a logistic regression classifier using the training set.

Step 4: Discard the 100 random documents added in Step 2 from the training set.

Step 5: Score all the paragraphs from all documents that have not been judged yet using the newly trained classifier.

Step 6: Present the highest-scoring paragraph p for assessment, and record the user’s judgment for document d to which the paragraph belongs.

Step 7: Add the labeled document d to the training set.

Step 8: Repeat steps 2 through 7 until some stopping criteria is satisfied.

Table 5.2: Algorithm for C++ implementation of paragraph CAL by [Abualsaud et al.](#) based on AutoTAR (algorithm included in [Zhang et al. 2018](#)) ([Abualsaud et al., 2018](#); [Zhang et al., 2018](#)).

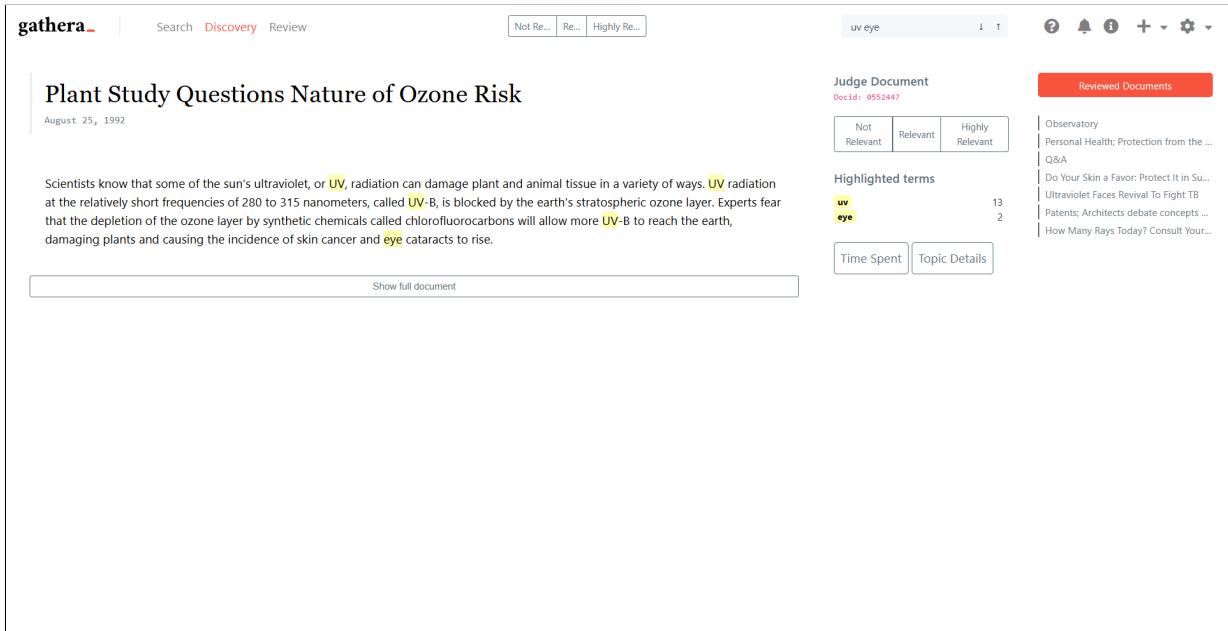


Figure 5.1: User-Interface for CAL.

5.4 User-Interface and configurations

Both “CAL” and “Discovery” were used interchangeably throughout the interface to refer to the BMI version of AutoTAR implemented in C++ by [Abualsaud et al.](#). The system used in our experiment has five different user interface configurations:

CAL Only: This is the base configuration where the search component is disabled, and the users are limited to judging paragraphs returned by CAL. Figure 5.1, shows the user interface for the tool. Users are initially shown a paragraph, however, if it lacks detail, users can also view the full document by pressing the “show full document” button. Once they make their relevance judgment, the system immediately shows the next paragraph. Users can also keep track of their time spent and review topic details again if needed while completing the task. This configuration is similar to the “CAL-D” configuration from the

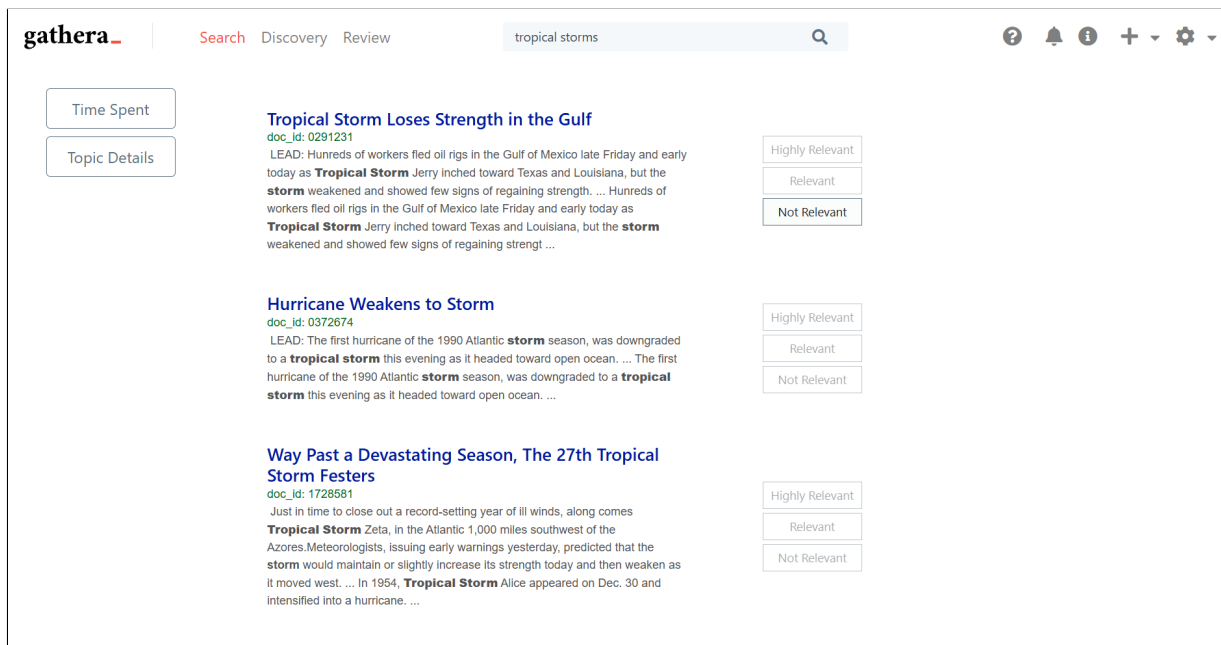


Figure 5.2: User-Interface for search engine.

Zhang et al. experiment (Zhang et al., 2018) and thus serves as a baseline to evaluate the performance of CAL when isolated, and also allows us to compare our results with the prior experiment.

CAL and Search (CAL&Search): In this configuration, both search and CAL components are available for use at all times; the configuration is also similar to the “CAL-D&Search” configuration from Zhang et al. (Zhang et al., 2018). Figure 5.2 depicts the search interface once the users have searched for a query. Alongside each result, a query-biased snippet is also shown with query terms in bold, along with buttons to make relevance judgments on the side, allowing for quick judgments. Users can also click on the search results to view full documents, as shown in figure 5.3. The configuration uses the same interface for CAL as in figure 5.1.

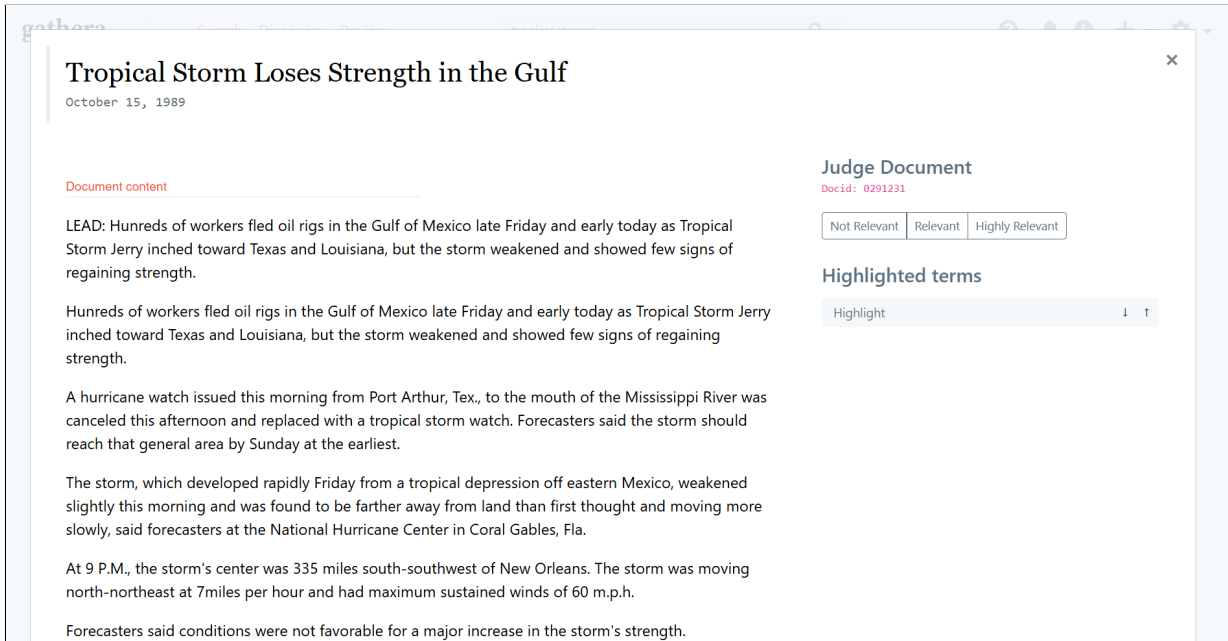


Figure 5.3: Modal to view full document.

Search with a nudge towards CAL (CAL&SearchNudge): The configuration is similar to “CAL&Search”, however, the CAL component is disabled until the users have found at least 5 relevant documents through search. Once found, the system shows a feedback message at the top of the page which encourages users to switch to CAL (figure 5.4). The message also includes a button to navigate users to CAL. The goal of this interface is to speed up CAL’s initial training phase by forcing users to prime the model with relevant information before they start the document review process.

Search with Integrated CAL (iCAL): This variation provides a more streamlined user interface by combining search and CAL on the same page. The goal of the configuration is to reduce the switching costs between the two tools and to allow users to use them simultaneously. As seen in figure 5.5, once users search for something, a CAL wid-

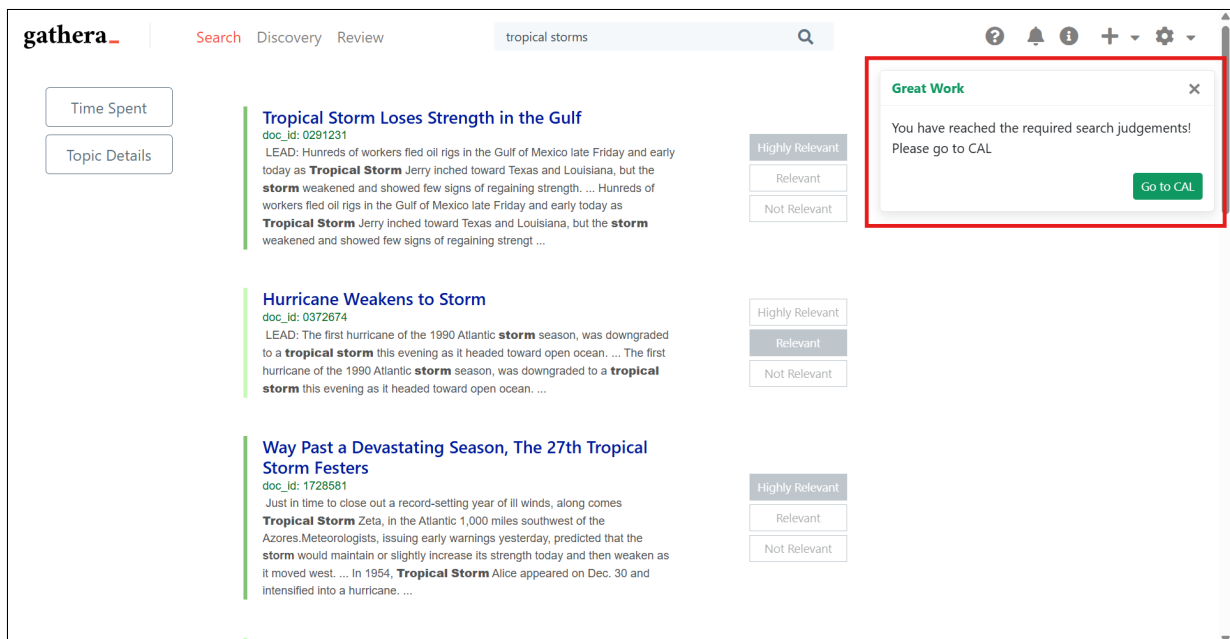


Figure 5.4: User interface for CAL&SearchNudge once 5 relevant documents have been identified. The section highlighted in red shows the ‘Nudge’ message.

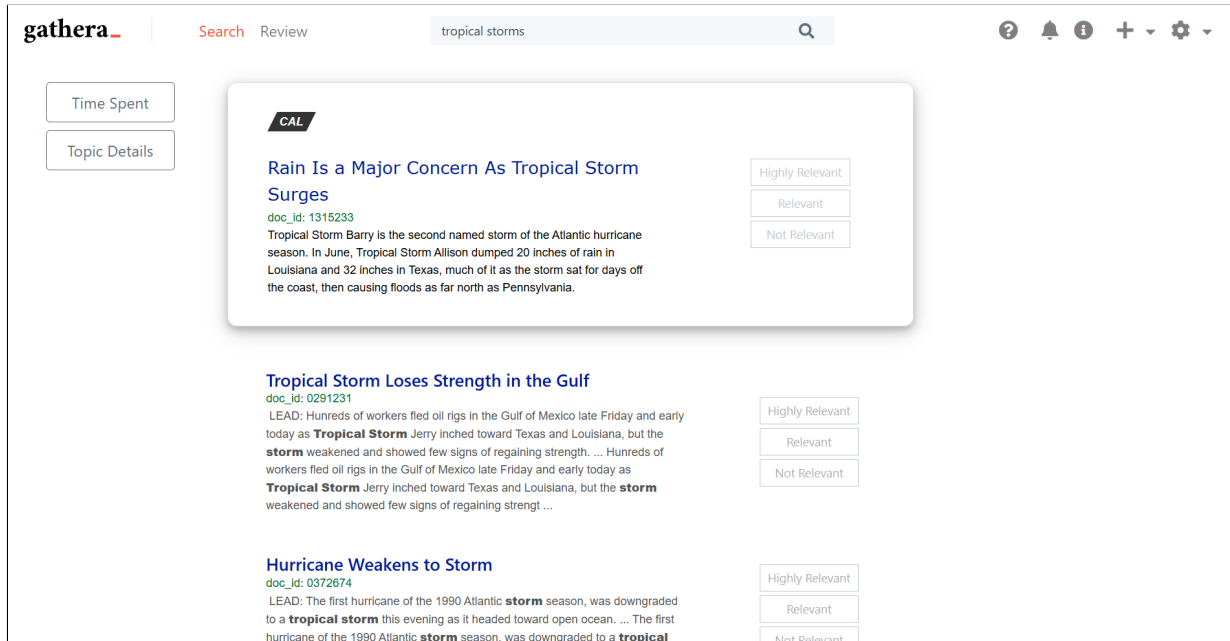


Figure 5.5: User-Interface for Search with Integrated CAL.

get is shown as the first result followed by the normal ranked list. The document shown within the CAL widget is the same document users would get if they were to use CAL on a separate page. Similarly to normal search results, users can also click on it to view the complete document. Moreover, similar to normal CAL, as soon as a judgment is made in iCAL, the system immediately finds the next best document and replaces it with the current one.

Search with a nudge towards Integrated CAL (iCALNudge): This variation is a hybrid configuration of “CAL&SearchNudge” and “iCAL”. The configuration uses the normal search user interface until users have found at least 5 relevant documents. Once found, users are shown a message in the top-right corner of the page and the CAL widget from the “iCAL” configuration appears at the top, allowing for a seamless switch between

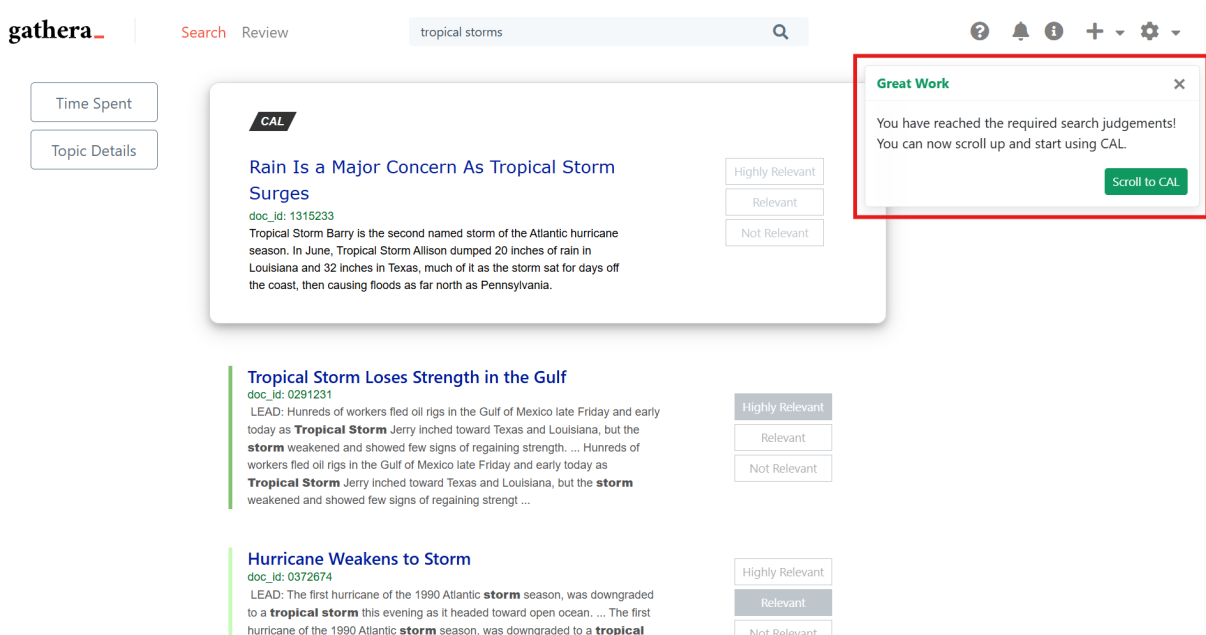


Figure 5.6: User-interface for iCALNudge once five relevant documents are found. The section marked in red shows the ‘Nudge’ message.

both systems (figure 5.6). The message also has a button that scrolls the screen to the top of the page, where the CAL widget is located.

The five different configurations not only allow us to understand how model priming effects the rate of finding relevant information, but also enable us to understand how the newly implemented interface will affect the efficiency of the high-recall system. The system also included a highlight feature, allowing users to quickly navigate to the relevant sections of the documents.

5.5 User study procedure

After receiving approval from the University of Waterloo ethics review board, we started recruiting our 40 study participants by placing posters on various university bulletin boards, creating social media posts on Reddit, sharing the poster in numerous WhatsApp groups, and emailing Management Science and Engineering student email lists at the University of Waterloo (see appendix C for poster). The study required participants to be proficient in reading and understanding English. Participants needed to be able to comfortably read and understand documents in English without assistance. If English was not their first language, they were expected to be fluent enough to claim so on their resume or CV. In addition, participants were required to have access to and be able to use a computer independently. We also ensured that participants were available to attend a one-hour in-person tutorial on the University of Waterloo campus.

We created a balanced study design by using 5x5 Greco Latin squares, with rows as user_ids and columns as the task order number. We randomly selected 40 topics from the 50 NIST-assessed topics in the collection (section 5.1). The selected topics were randomly divided into 8 groups of 5 topics. The five experiment configurations, as highlighted in section 5.4, were used as the treatments. For each of the 8 topic groups, the 5 topics and the 5 treatments were randomly assigned to the cells of the Latin square. The user_ids and their respective tasks were stored in a database, and the 40 experiment users were assigned the stored user_ids.

5.5.1 Tutorial

We conducted tutorials in small groups of less than six participants to ensure they were interactive and engaging (see appendix B for tutorial slides). The tutorials started with

a brief introduction to the experiment and what the participants will be doing. The participants were repeatedly told to treat the study as a scientific research and to not use external AI tools, such as ChatGPT, when making relevance assessments, as doing so could adversely affect the validity and reliability of the collected data.

The various tools available during the task were demonstrated (Search, CAL and iCAL). Although the participants were familiar with using search engines, they did not know what CAL was or how it worked. We explained that CAL is a machine learning-based system that uses user-provided relevance feedback to iteratively refine its understanding of the user’s information need. The system runs in a loop and, upon starting, presents a document for assessment. Once the user provides a judgment, the system updates and selects the next document it estimates to be most relevant, and the process continues. Moreover, the participants were informed that both CAL and search tools are linked together, and any document judged in the search would automatically be sent to CAL for learning, improving its understanding of the topic. Furthermore, we also informed participants that they can use the search feature to train CAL by finding a few relevant documents before switching to CAL. This approach helps CAL learn from useful feedback early on, so it can show better documents right from the start.

We provided participants with a detailed explanation of the task and guided them on properly evaluating documents. We introduced our three-point grading scale, highlighting the differences between relevant, non-relevant, and highly relevant documents. We guided the participants to work as fast as possible while maintaining accuracy. Moreover, they were also informed that their judgment speed would not affect the total task time, as the system will keep showing documents until all 1.8 million documents have been judged. Moreover, participants were instructed to remain consistent in their relevance judgments, to periodically refer back to the topic statement, and to evaluate each document based on

its own merit. They were also reminded that if none of the documents appeared relevant, it was perfectly acceptable to mark all of them as non-relevant, and that they should not feel pressured to label any document as relevant without sufficient justification.

To train people in judging documents, we created a website where users could practice relevance assessments. The website had eight documents for two different topics: four documents from topic 303, “*Hubble Telescope Achievements*”, and 4 documents from topic 301, “*International Organized Crime*”. These topics were from the 200 crowdsourced topics of the test collection not assessed by NIST assessors (section 5.1). During the tutorial, we used the website to demonstrate what the actual topic will look like in each of the tasks of the experiment, noting that each task would have a different topic. The participants were then instructed to read the topic and the first document. For each document, we randomly selected a participant and asked them to share their relevance judgment aloud. This was followed by a brief group discussion in which we reviewed their reasoning and provided detailed feedback on why the document is relevant/non-relevant. This approach ensured that each participant had the opportunity to judge at least one document during the tutorial session.

The website had buttons to judge documents as well, and upon judging, the website also provided feedback to users along with a brief explanation of the correct judgment. As shown in Figure 5.7, the practice website provided immediate feedback after each relevance judgment. When a judgment was incorrect, the system explained why it did not align with the expected answer. Similar explanatory feedback was also given when a correct judgment was made, helping participants better understand the criteria for relevance. We handpicked these documents to demonstrate that for a document to be relevant, it has to completely satisfy the relevance criteria highlighted in the topics. Furthermore, through these documents, we also explained that documents could have all the keywords and still

Practice

! Incorrect Judgment! ×

Title: Hubble Telescope Achievements

Description: Identify positive accomplishments of the Hubble telescope since it was launched in 1991. Documents are relevant that show the Hubble telescope has produced new data, better quality data than previously available, data that has increased human knowledge of the universe, or data that has led to disproving previously existing theories or hypotheses. Documents limited to the shortcomings of the telescope would be irrelevant. Details of repairs or modifications to the telescope without reference to positive achievements would not be relevant.

Document number: 1 / 8

Document Content:

To the Editor:

We were pleased to read your strong endorsement of the Hubble Space Telescope ("Premature Death for the Hubble," editorial, Feb. 29). However, Hubble is merely the canary in the coal mine.

NASA has an extraordinary science program, with planned missions that range from detecting the ripples in space-time produced by merging black holes to probing the mysterious "dark energy" that is accelerating the expansion of the universe. Its smaller missions draw on the best ideas from the scientific community. All of these are slated for delay or cancellation to pay for the Moon-Mars initiative.

Whatever one thinks of the president's ambitious but expensive vision, it is doomed to fail unless it first acquires broad public support and scientific backing through extensive discussion.

The administration's blunderbuss approach to reforming the space program threatens to repeat the failures of the space station, squeezing out an enormous range of exciting science while creating nothing of value.

ANDREW GOULD DAVID WEINBERG Columbus, Ohio, Feb. 29, 2004

The writers are professors of astronomy at Ohio State University.

The document is not relevant because: The document discusses the cancellation of various NASA missions, including the Hubble Space Telescope, in favor of the Moon-Mars initiative. It does not provide specific achievements or positive contributions of the Hubble Telescope.

RELEVANTNOT RELEVANT

PREVIOUS DOCUMENTNEXT DOCUMENT

Figure 5.7: Website used to train participants on relevance judgments.

be non-relevant.

Apart from the five actual experiment tasks, users were also assigned five practice tasks to be completed during the tutorial. For these practice tasks, we used topic 303, “*Hubble Telescope Achievements*”, from the 200 crowdsourced topics of the test collection not assessed by NIST assessors (section 5.1). For these tasks, users used all 5 configurations of the experiment but with reduced time limits (three tasks were 2 minutes each, and the remaining two were 4 minutes). During the practice tasks, we highlighted all the features of our system and ensured that the participants were completely comfortable with the system before the actual experiment started. They were also given time to explore the system themselves while reading and judging documents, similarly to how they would in the experiment tasks.

The participants were allowed to complete the experiment remotely at the time of their choosing, and a time-tracking system was added to the application. Participants were allowed to take breaks at any time except during the tutorial, however, we encouraged them to only take breaks in between the tasks, as during the tasks their time was only counted as long as they were actively judging documents. The task timer started when the participants went to either the search page, or the CAL page, and was incremented on each judgment made. If we detected that the users had not made any judgment, searched for a query, or switched to a different tool for a minute and 45 seconds, we showed them a pop-up asking if they were still working. If the users did not confirm within the first 15 seconds of the popup being shown, we assumed that they stopped working on the task, and only the time until their last judgment was counted.

For each task, the homepage showed an overview of what the task entails, i.e., the tools they will have access to during the task and how the user interface will look like, followed by the topic details and a pre-task questionnaire button. The users were required to complete

the questionnaire before they could start the task, and the task-related tools were disabled until the pre-task questionnaire was completed. The pre-task questionnaire assessed users' prior knowledge about the topics, asking questions about the topic familiarity, perceived difficulty, and general topic feedback. After completing the questionnaires, the participants could navigate to the available tools and start judging documents. For each task, the participants were required to search documents for 1 hour. The 1 hour started as soon as they navigated to search or CAL page.

Once we detected that the participants had searched for the complete 1 hour, another popup was shown, instructing them to go to the home page and complete the post-task questionnaire. The post-task questionnaire asked users how difficult it was to find documents, their confidence in the relevance judgments they made, and their overall mood during the task. We also asked for an interface rating from (1-7), and for the effectiveness of the user-interface in finding relevant documents on a Likert scale. The questionnaire also included a general feedback section.

Once the post-task questionnaire was completed, the next task was automatically activated for the participants. After completing all five tasks, users were given a post-experiment questionnaire, asking about the study difficulty, their overall experience/mood throughout the experiment, what they liked/disliked about the study, and the feature they thought was the most helpful in finding documents. After completing the post-experiment questionnaire, the users were shown a banner saying the experiment had concluded.

We started the experiment with 5 individuals to ensure the system was running properly, then after minor improvements in displaying document content (removing automatic keyword highlighting done by CAL, removing metadata from the top of the documents, removing the duplicate of the "LEAD" paragraph for numerous documents, and showing the article headlines for the documents on the search results page instead of the first 55

characters of the document as the title), the system was rolled out to the remaining 35 participants. The system remained fully identical in terms of retrieval quality, task design, and data collection. Since no changes were made to the core experiment conditions, and all 40 participants completed the experiment under the same conditions, we decided to include everyone in the results for a bigger sample size.

5.6 Evaluation

Throughout the experiment, we collected detailed user analytics data. All user actions were recorded including search queries, search results, search snippets produced, judgments made, documents viewed, and the tool through which judgments were made. To make our results comparable to those of [Zhang et al.](#), we used primarily the evaluation measures they used ([Zhang et al., 2018](#)). The timer for our experiment only updated after a judgment was made, thus, the last judgment was always over the one-hour period. We cleaned our data and removed any judgments made after the one-hour mark.

Since the experiment compared user actions and relevance judgments across five different configurations within a fixed time frame, we report the mean number of relevant documents found per configuration across the 40 participants. A higher mean number of relevant documents indicates a more effective configuration. Furthermore, we also report the mean number of NIST-judged relevant documents, calculated as the mean of the union of relevant documents identified by users and those marked relevant by NIST assessors, as an evaluation measure. Both these measures were also utilized by [Zhang et al.](#)

Since, HRIR systems aim to establish high levels of recall with high precision, precision and recall are extremely important evaluation measures for this experiment. Following

prior work by [Zhang et al.](#), we defined precision as the total number of NIST marked relevant documents found by users over the total number of self reported relevant documents ($Precision = |U_{rel} \cap R| / |U_{rel}|$, where U_{rel} is the documents marked relevant by the users, and R is the total relevant documents in the collection marked by NIST), and recall as the total number of NIST relevant documents found by users divided by the total number of relevant documents in the collection as defined by NIST ($Recall = |U_{rel} \cap R| / |R|$). A good high-recall retrieval system aims to achieve high recall while maintaining high precision. Thus, we use both of these measures to evaluate our system. Moreover, F1 score ($F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$), that combines both precision and recall into one measure, was also utilized.

Furthermore, following the metrics used in [Zhang et al.](#), we also show a plot for the number of NIST-judged relevant documents found, across the number of self-reported relevant documents for the first 50 self-reported relevant documents found ([Zhang et al., 2018](#)).

We also used time to find x relevant documents, where x was 1 and 5, to investigate the effects of priming and search. A lower time would indicate the effectiveness of the user interface in providing more information earlier to the users. We also used *switches/judgments*, that show the number of switches made between search and CAL normalized by the total number of judgments, to test our newly developed interface. A higher number would indicate that users felt more comfortable in using the two tools together.

For statistical significance testing, we used generalized linear mixed-effect models from lme4 package in R ([Bates et al., 2015](#); [R Core Team, 2020](#)). To build the models, we followed the tutorial by [Winter](#) ([Winter, 2013](#)). ‘Users’ and ‘Topics’ were chosen as random effects, and our experiment factors were used as fixed effects. Likelihood ratio test, that reports a p -value, was conducted for each performance measure (dependent variable), by

first building a complete model including all fixed and random effects and then comparing to a model with the factor in consideration removed. To obtain the p -value for search, we conducted the test by comparing a model that includes search with one that does not.

5.6.1 Effort Curves

Recall-effort curves are frequently used to demonstrate changes in recall over the total documents judged. Recall-effort curves were used by Cormack and Grossman (2014, 2015), Zhang et al. (2020), and were also the measure used in TREC Total Recall Tracks of 2015 and 2016 (Cormack and Grossman, 2015, 2014; Zhang et al., 2020; Roegiest et al., 2015; Grossman et al., 2016). Thus, we also utilized recall-effort curves for our analysis. The curves allow us to visualize and understand the effectiveness of different system configurations at varying levels of recall. In our study, all participants used each of the five system configurations for one hour. Therefore, to plot our effort curves, we measured effort as the time spent viewing documents, with a maximum possible effort of one hour. Since participants made relevance judgments at different times, we calculated averages at 30-second intervals to plot the effort curves. After every 30 seconds, the recall level was calculated for each session and then averaged across all participants. Moreover, we used the same effort definition and similar calculations to plot the average total number of relevant documents found-effort and the average number of NIST-marked relevant document found-effort curves to further highlight the performance of the different configurations in presenting relevant documents during the one-hour period.

Chapter 6

Results

6.1 Quantitative Analysis

Our experiment builds on the work of [Zhang et al.](#), to investigate the performance of five different configurations of an information retrieval system. The study uses 2x2 factorial design, with a new user-interface that combines CAL and search on a single page as the first factor, whereby, the first document shown in the search results is from CAL and the remaining are from Search (as seen in figure 5.5). The second factor controls whether CAL is disabled initially, requiring users to find five relevant documents through search, before the system *nudges* them towards CAL. Additionally, we also have a case with search disabled for a comparison with [Zhang et al.](#). The factors can be observed in a tabular format in table 1, and for our results, we will use the following notations from the table to refer to the factors: CAL-Only (CAL-D from [Zhang et al.](#)), CAL&Search (CAL-D&Search from [Zhang et al.](#)), CAL&SearchNudge, iCAL, and iCALNudge. All configurations except ‘CAL-Only’ include search. The goal of the experiment is to investigate the effects of these

factors on the performance measures highlighted earlier.

Table 6.1 recreates a table similar to Zhang et al., and highlights the main key to understand our results. For each of the performance measures, the tables start by highlighting the average of CAL-Only configuration, wherein search is disabled. The cell next to CAL-Only Average shows the p -value for CAL-Only, versus all 4 configurations that include search. This will be followed by a detailed analysis for the 2x2 factors. For each of the configurations, the tables highlight the mean performance. In addition, the table also reports marginal means for each factor. Moreover, the table shows the p -values from likelihood ratio tests to understand whether the effect of each factor was statistically significant. Finally, in the bottom right corner, we can see the mean performance measure for the 2x2 factors without including the average for CAL-Only configuration (the four configurations that include both Search and CAL). In cases we observed statistically significant differences, the means were marked by a ‘*’ sign.

Table 6.1: Key/primer for reading Tables.

CAL without search (base case / CAL-Only)	CAL-Only Average		p value (Search vs. No Search)	
CAL disabled initially	CAL types		Marginal means	
(Nudge)	Normal CAL	Integrated CAL	(Nudge)	
No	CAL&Search Average	iCAL Average	Average without Nudge	p value (Nudge vs. No Nudge)
Yes	CAL&SearchNudge Average	iCAL&Search Average	Average with Nudge	
Marginal means	Average for Normal CAL	Average for Integrated CAL	Overall Mean (without CAL-Only)	
(CAL types)	p value (Normal vs. Integrated CAL)			

6.1.1 Self-reported relevant documents

Table 6.2 shows the analysis for the documents marked relevant by users while searching. Although the averages show improvement in the total number of relevant documents found

CAL without search (base case / CAL-Only)	55.6	p value = 0.873	
CAL disabled initially (Nudge)	CAL types		Marginal means (Nudge)
	Normal CAL	Integrated CAL	
No	53.8	58.0	55.9
Yes	52.3	62.8	57.9
Marginal means (CAL types)	53.4	60.4	56.9
	p value = 0.637		

Table 6.2: Self-reported relevant documents found by users.

by users when using Integrated CAL, or when CAL was disabled initially, these results were not statistically significant. Our values were also close to the results provided by [Zhang et al.](#). For the CAL-Only (CAL-D in [Zhang et al.](#)) configuration, with full document available, [Zhang et al.](#) reported a value of 58.3, whereas, the average value for our CAL-Only configuration was 55.6. Moreover, for Search&CAL (CAL-D&Search in [Zhang et al.](#)) [Zhang et al.](#)'s average was 51.4, whereas our average was 53.8. These values indicate a good replication of [Zhang et al.](#)'s work, and again points to search being slightly slower. However, upon comparing the averages for the all configurations that include search with CAL-Only, we can see that the difference between them is negligible, with the average for the configurations that use search being slightly more. With the p -value being more than 0.05 in this case as well, we failed to reject the null hypothesis. This is in contrast to [Zhang et al.](#)'s work, who found that when full documents were available, users on average could find approximately 6.9 more relevant documents when search was disabled.

Moreover, the results from the user study also contradicted the findings from our simulation experiment in chapter 4, where we found that priming CAL with relevant documents increases the number of relevant documents found. On average, our participants marked

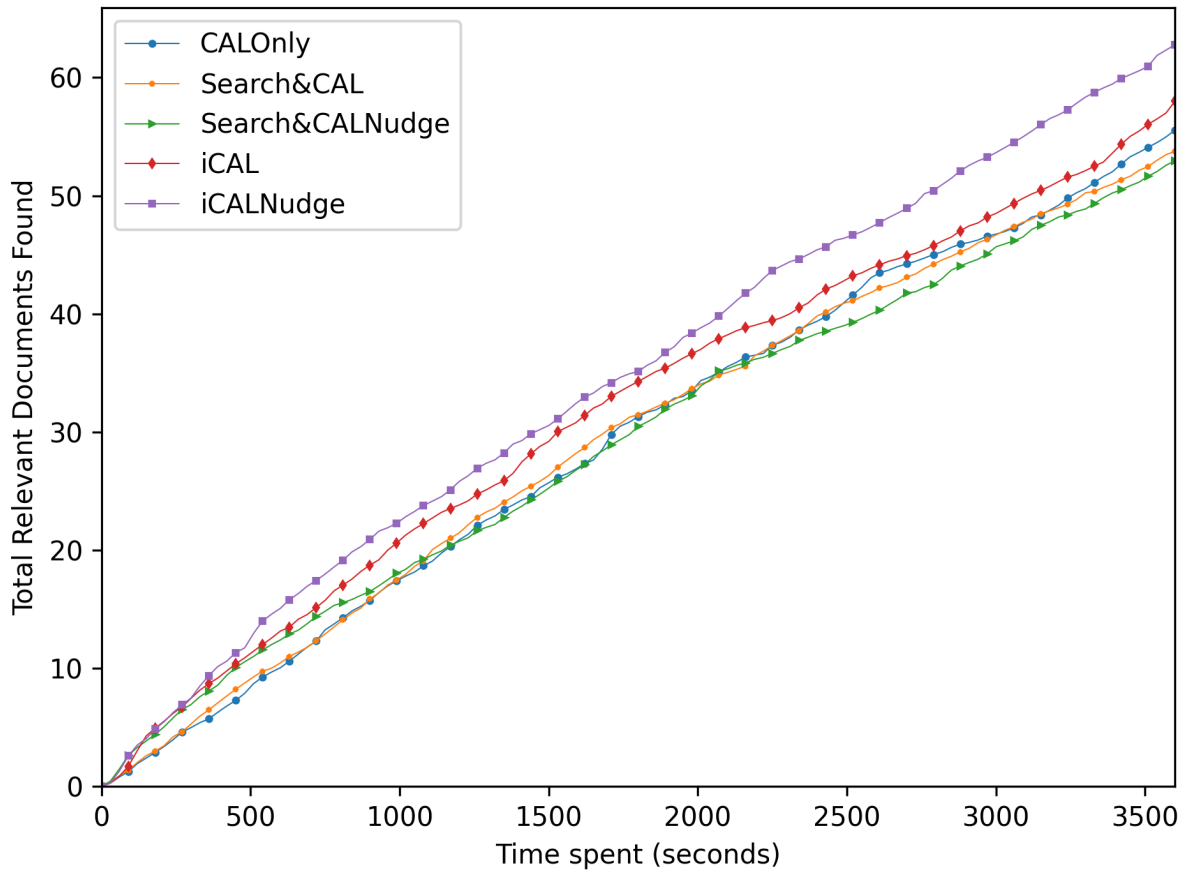


Figure 6.1: Average total number of self-reported relevant documents found by the participants.

140 documents and the plot on figure 4.2 shows a significant difference at that effort level, however, we did not observe similar effects in our user study. Nevertheless, it is important to note that user behavior can be unpredictable and differences can arise from the way systems are used.

Figure 6.1, shows a plot for the average total number of self-reported relevant documents during the one-hour task for each user interface. It can be observed that in early-stages, other than Search&CAL, all configurations that include search perform better than the CAL-Only configuration. However, contrary to our simulation, the difference between the curves for the configurations primed with 5 relevant documents (CAL&SearchNudge and iCALNudge), and the curve for CAL-Only is not as much as we expected. However, the effects of priming can still be observed in the early stages, where the curves for CAL&SearchNudge and iCALNudge are noticeably above CAL-Only.

6.1.2 NIST marked relevant documents

As highlighted by Zhang et al., the metric is important as it is similar to the notion of “*Second pass*” in e-discovery, whereby, once the first review is done, an expert comes in and reviews the selected documents to pick the final set of relevant documents. Moreover, the metric also underscores the importance of “*quality*” over “*quantity*”. Here, for our experiment, we used the judgments from NIST assessors for our expert review and the session by our participants was considered as the first pass. The results can be seen in table 6.3. Although we can see that the inclusion of search increases the number of relevant documents found (from 20.4 to 24.0), the results are not statistically significant (p -value = 0.122). Slight differences can also be observed for the two factors, iCAL and Nudge, but again, the differences are not statistically significant. These results also align with

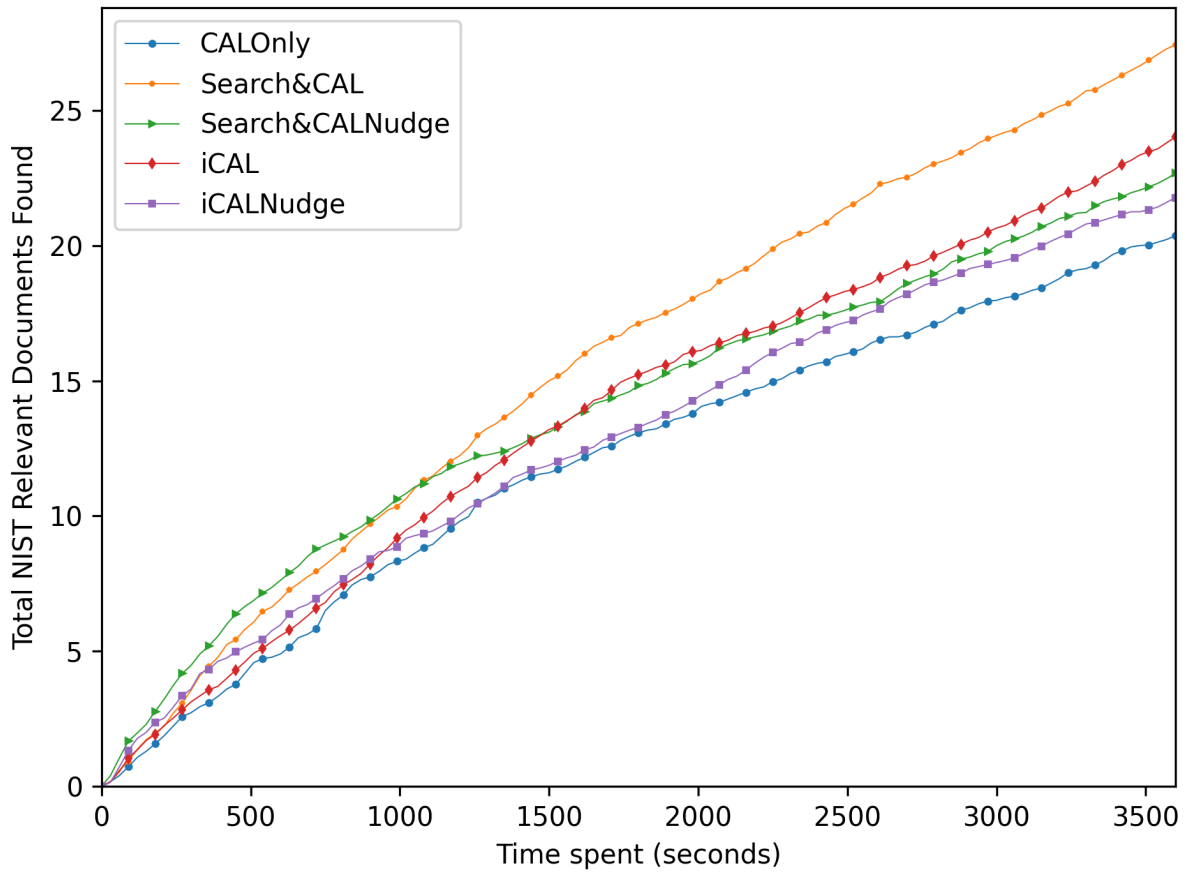


Figure 6.2: Average total number of NIST-marked relevant documents found by the participants.

the findings by [Zhang et al.](#), where the inclusion of search in CAL-D resulted in slight improvement in the total number of NIST marked relevant documents found. Moreover, we can also observe that the performance is similar for both Normal CAL and iCAL based configurations.

CAL without search (base case / CAL-Only)	20.4		p value = 0.122	
CAL disabled initially	CAL types		Marginal means	
(Nudge)	Normal CAL	Integrated CAL	(Nudge)	
No	27.4	24.0	25.7	p value = 0.228
Yes	22.7	21.9	22.2	
Marginal means	25.1	22.9	24.0	
(CAL types)	p value = 0.513			

Table 6.3: NIST marked relevant documents found.

Figure 6.2, shows a plot for the average total number of NIST-marked relevant documents-effort curves for each of the five configurations. We can observe that, on average, all configurations that include search perform better than the CAL-Only configuration. Moreover, apart from “iCALNudge” configuration, we can see a clear difference in the number of NIST-marked relevant document found when search was available vs CAL-only, even at higher efforts.

6.1.3 Recall

Recall is an extremely important metric and, as noted by [Zhang et al.](#), it normalizes the number of NIST-marked relevant documents by the total known relevant documents for a topic, as each topic has a different number of relevant documents. Table 6.4 highlights

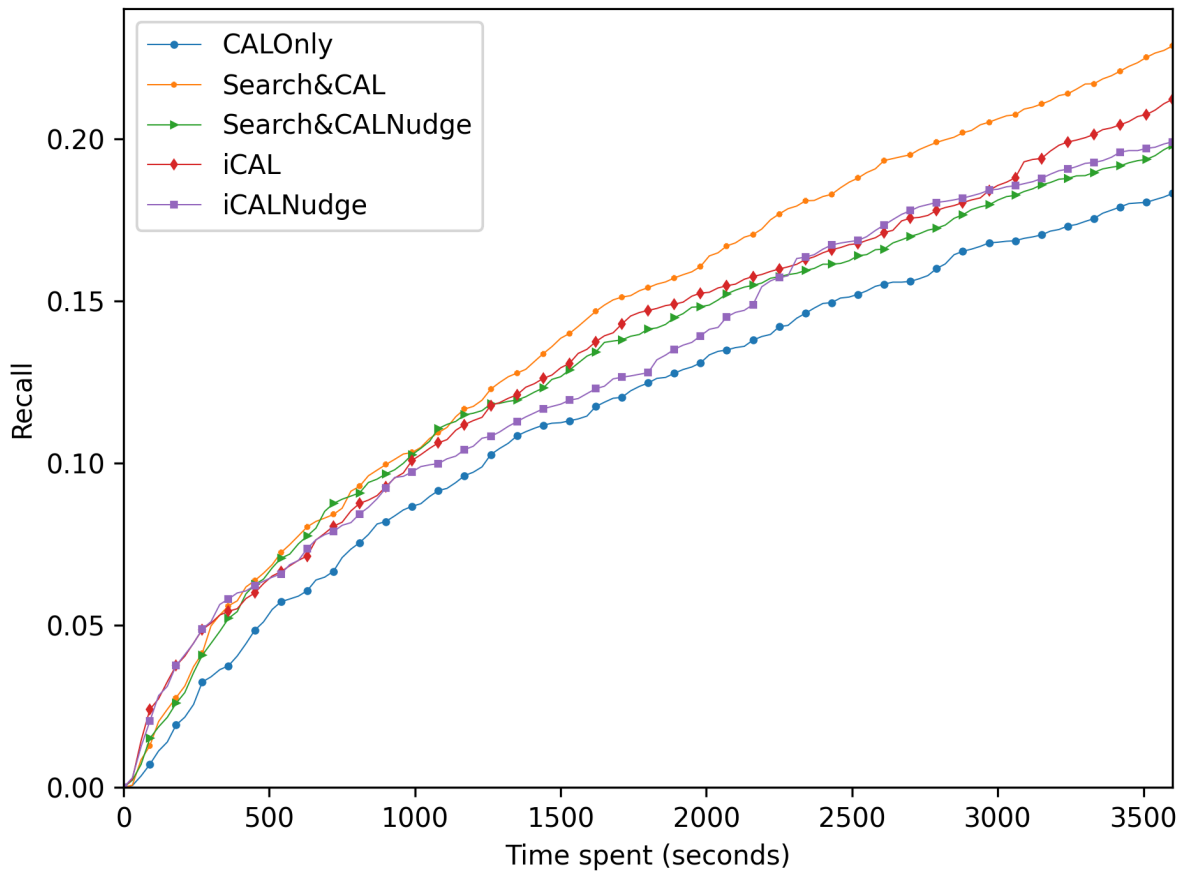


Figure 6.3: Recall-effort plot for all 5 configurations.

CAL without search (base case / CAL-Only)	0.18	p value = 0.134	
CAL disabled initially (Nudge)	CAL types		Marginal means (Nudge)
	Normal CAL	Integrated CAL	
No	0.23	0.21	0.22
Yes	0.20	0.20	0.20
Marginal means (CAL types)	0.21	0.21	0.21
	p value = 0.761		

Table 6.4: Recall.

the summary for the performance measure. Similar to figure 6.3, we can observe that the addition of search does improve recall, but the difference is not statistically significant (p -value = 0.122). Moreover, apart from slight differences, disabling CAL initially, or using Normal vs Integrated CAL did not yield significant improvements. The results are somewhat consistent with Zhang et al.’s experiment, who found that inclusion of search did not impact recall.

Figure 6.3 shows the average recall-effort curves for the five configurations. It can be seen that, on average, the addition of search yields improvements at all levels of recall. Although our statistical significance testing showed no significant differences, a relatively low p -value of 0.134 between search-based configurations and CAL-only, along with the fact that all search-based configurations performed better, suggests a potential effect worth further investigation.

6.1.4 Precision

As explained by Zhang et al., in case of two-pass review scenario, false-positives (non-relevant documents marked as relevant) result in wasted effort and reduced efficiency for the second reviewer. Table 6.5 highlights the precision for all configurations for the study. Similar to Zhang et al. findings, we observe that the inclusion of search improves the mean precision for the system from 0.47 to 0.53. Moreover, despite observable differences in averages between the “Nudge” and “iCAL” conditions, these variations did not result in statistically significant improvements. The results are also consistent with Zhang et al., who found that search does improve precision.

CAL without search (base case / CAL-Only)		0.47	p value = 0.029	
CAL disabled initially (Nudge)	CAL types		Marginal means	
	Normal CAL	Integrated CAL	(Nudge)	
No	0.55	0.51	0.53	p value = 0.926
Yes	0.53	0.51	0.52	
Marginal means (CAL types)	0.54	0.51	0.53*	
	p value = 0.616			

Table 6.5: Precision.

Figure 6.4 shows the curve that plots the total number of NIST marked relevant documents found against the total number of self reported relevant documents. Average of NIST marked relevant documents, for each user reported relevant document point was calculated but only for those participants who had that many relevant documents. The gradient of the curve is precision. A similar plot was used by Zhang et al. to observe whether a reduced number of NIST-reported relevant documents found was due to a slower rate of judgment or due to judgment mistakes, and they concluded that the availability of search

slows down the judgment rate. In our case, we did not observe a decrease in NIST-marked relevant documents with the availability of search. Nevertheless, we believed the results would be still interesting to observe. We can see that early on, precision is very good for all configurations, however, later on, as more documents are marked relevant, the performance under CAL-Only configuration gets worse.

6.1.5 Precision at minimum number of relevant documents

Zhang et al. used mean precision@k (where k is the minimum number of relevant documents found for a given topic across the five configurations) to better compare the different conditions, as each topic has a different number of relevant documents. Contrary to their experiment, where it was reported that addition of search in the user-interface did not significantly impact the performance, we found that search does improve the precision@k (from 0.48 to 0.59), and the results were also statistically significant. This also aligns with our results for precision. However, we did not find any statistically significant differences for when CAL was disabled initially, or when integrated CAL was used. Overall, we can conclude that the addition of search significantly improves precision.

6.1.6 F1 score

F1 score is an important measure that combines both recall and precision into one metric. Figure 6.7 summarizes the F1 scores for all configurations. Although we can observe slight improvements in the F1 score when search is added, or when Normal CAL is used instead of iCAL, or when CAL is not disabled initially, the changes are not statistically significant. Zhang et al. also did not find any statistically significant differences in F1 score when search was added.

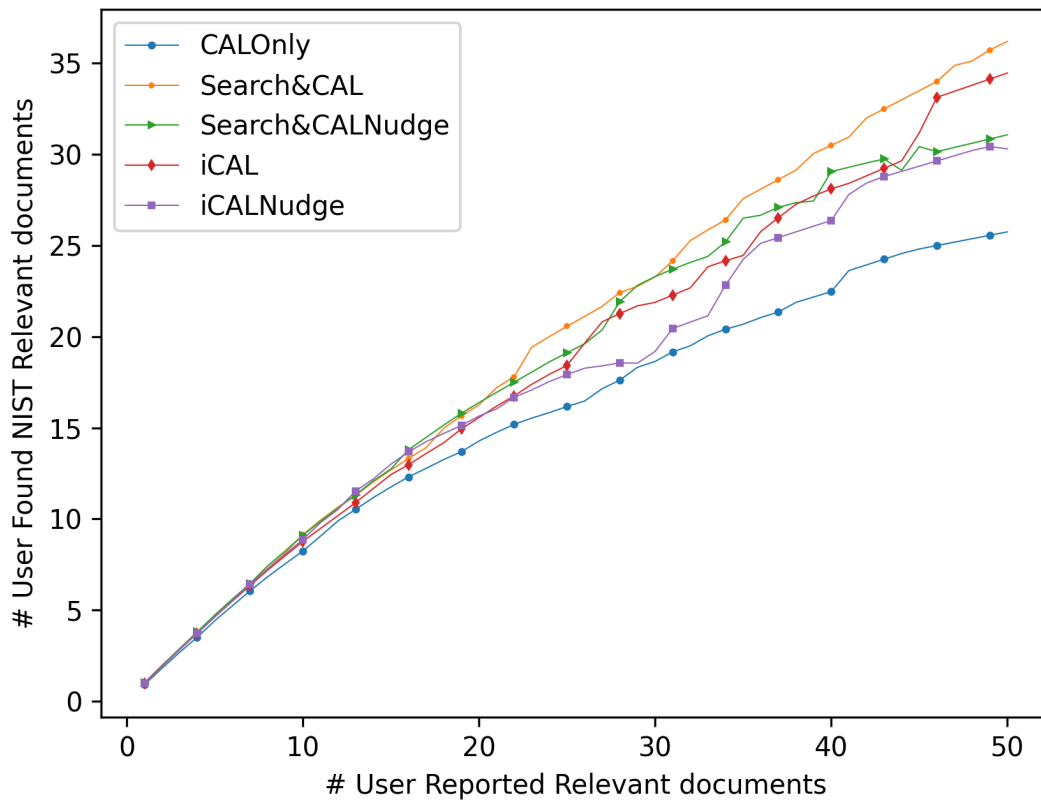


Figure 6.4: User found relevant documents over user found NIST relevant documents for the first 50 self reported relevant documents.

CAL without search (base case / CAL-Only)	0.48	p value < 0.001	
CAL disabled initially (Nudge)	CAL types		Marginal means (Nudge)
	Normal CAL	Integrated CAL	
No	0.61	0.57	0.59
Yes	0.61	0.57	0.59
Marginal means (CAL types)	0.61	0.57	0.59*
	p value = 0.243		

Table 6.6: Precision at minimum number of self reported relevant documents. One user found 0 documents during one of their 1-hour sessions. Their precision was reported as 0 for that session for the calculation.

6.1.7 Time to find relevant documents

To further understand the effects of priming, we investigated the time it took users to find relevant documents. Specifically, we measured the time taken to find the first relevant document, and the time taken to find the first five relevant documents. Table 6.8 highlights the average time taken, in minutes, to find the first relevant document for each of the different user-interface configurations. For this calculation, we removed all values for all tasks where the number of user-reported relevant documents was less than the required number (1 or 5). We can observe that the time taken to find the first relevant document when search is available is 50% less than when working with CAL alone. Moreover, we can also observe small improvements for when CAL was disabled initially, or when iCAL was used, these differences were not statistically significant. From a user-interface perspective, having to wait twice as long to find the first relevant document can be frustrating, and might reduce users' willingness to engage with or continue using the system.

Table 6.9 shows the time taken to find the first five relevant documents. Although

CAL without search (base case / CAL-Only)	0.23		p value = 0.170	
CAL disabled initially (Nudge)	CAL types		Marginal means (Nudge)	
	Normal CAL	Integrated CAL		
No	0.28	0.25	0.27	p value = 0.186
Yes	0.25	0.23	0.24	
Marginal means (CAL types)	0.27	0.24	0.25	
	p value = 0.196			

Table 6.7: F1 score.

CAL without search (base case / CAL-Only)	3.19		p value < 0.001	
CAL disabled initially (Nudge)	CAL types		Marginal means (Nudge)	
	Normal CAL	Integrated CAL		
No	1.93	1.92	1.93	p value = 0.542
Yes	1.73	1.57	1.63	
Marginal means (CAL types)	1.83	1.75	1.79	
	p value = 0.899			

Table 6.8: Time taken to find the first relevant document (in minutes).

we can still see similar trends as in table 6.8, whereby the addition of search reduces the time taken to find the first few relevant documents, the differences are not statistically significant. The results show that although the inclusion of search in the user-interface can improve the user experience during early stages, the effect fades away as time passes on.

CAL without search (base case / CAL-Only)	9.61		p value = 0.220	
CAL disabled initially (Nudge)	CAL types		Marginal means (Nudge)	
	Normal CAL	Integrated CAL		
No	10.2	7.63	8.90	p value = 0.281
Yes	7.68	7.97	7.83	
Marginal means (CAL types)	8.92	7.80	8.31	
	p value = 0.257			

Table 6.9: Time taken to find 5 relevant documents (in minutes).

6.1.8 Average number of search judgments made in the first five judgments

Our results from tables 6.1, 6.3, and 6.4 show that addition of search, or priming CAL (disabling CAL initially) does not significantly affect the performance. This is in contrast to our simulation, where we found that priming CAL would result in improvements. The results were also different from the findings of Zhang et al.’s experiment, who found that the inclusion of search harms performance.

The difference in results compared to the previous study on the addition of search can likely be explained by how we designed and introduced the system to users. During the tutorial, we repeatedly informed users that CAL works on user feedback and the more documents they judge, the better it will get. We also informed users that they could use search as a means to seed CAL, encouraging them to find a few relevant documents before switching to CAL. Our “Nudge” interface (CAL disabled initially) also trained users to use search first, as they had to find five documents before CAL became available, essentially teaching them to prime the system before relying on CAL. During the tutorial,

CAL without search (base case / CAL-Only)	N/A		N/A	
CAL disabled initially (Nudge)	CAL types		Marginal means (Nudge)	
	Normal CAL	Integrated CAL		
No	3.23	2.48	2.85*	p value < 0.001
Yes	5.00	4.95	4.98	
Marginal means (CAL types)	4.11	3.71*	3.91	
	p value = 0.037			

Table 6.10: Average total number of search documents judged for the first five documents marked by users. The average for “*iCALNudge*” is 4.95 instead of 5 because one participant changed their judgment made through review tab (a page that shows past judgments).

participants also went over the complete experiment but with shorter time controls, and thus, observed the effects of priming through the Nudge interface through *iCALNudge* and *Search&CALNudge*. As a result, users in our study may have used search more effectively, which could explain why we did not observe the same negative impact of search reported in the previous study.

To investigate whether users were naturally using search for seeding, even when not explicitly instructed, we examined how many of their first five and ten judgments came from search. By analyzing the proportion of search-based judgments early in the review process, we aimed to investigate whether training users could serve as an effective alternative to limiting the user’s options and restricting the interface.

Table 6.10 shows the summary for the number of documents found through search for the first five documents. We can observe that even when the users are not restricted in terms of the interface, they still use search. Especially in the *CAL&Search* case, where on average, 3.23 documents from the first five are found through search initially. The number

is slightly low for iCAL, but we believe this might be due to the user-interface design, whereby the CAL component is always shown as the first result in search.

Table 6.11 shows similar results. Participants on average marked 5.41 documents through search, even when CAL was not disabled initially. This is approximately on average just two documents less than when CAL was disabled initially, indicating that when users are trained properly, user interface limitations may not be required. Furthermore, we believe this is why we did not observe the promised improvement from priming.

CAL without search (base case / CAL-Only)	N/A		N/A	
CAL disabled initially (Nudge)	CAL types		Marginal means (Nudge)	
	Normal CAL	Integrated CAL		
No	6.10	4.73	5.41*	p value < 0.001
Yes	7.13	8.00	7.56	
Marginal means (CAL types)	6.61	6.36*	6.49	
	p value = 0.037			

Table 6.11: Average total search documents judged in the first 10 judgments.

6.1.9 Switches per judgment

We hypothesized that our new user interface, integrated CAL, would allow users to switch back and forth between the two tools, search and CAL, effortlessly providing a seamless user experience. To test this hypothesis, we calculated the number of switches, normalized by the total number of judgments, whereby a higher number could indicate an ease in switching. Table 6.12 highlights the average number of switches per judgment for each of the configurations. It is interesting to note that using iCAL causes more than a 100%

increase in the number, with the difference being statistically significant. The number confirms our hypothesis that the interface does let users switch effortlessly between the two tools, and considering the results from tables 6.1, 6.3, and 6.4, the increased amount of switching does not cause a loss in overall performance.

CAL without search (base case / CAL-Only)	N/A		N/A	
CAL disabled initially	CAL types		Marginal means	
(Nudge)	Normal CAL	Integrated CAL	(Nudge)	
No	0.07	0.16	0.11	p value = 0.257
Yes	0.06	0.13	0.09	
Marginal means	0.06	0.14*	0.10	
(CAL types)	p value < 0.001			

Table 6.12: Total switches between search and CAL, normalized by total number of judgments ($totalSwitches/totalJudgments$).

6.2 Usability Analysis

Figures 6.6 and 6.5 summarize participant feedback on interface effectiveness and usability, respectively. To capture the usability of the system, we asked participants “*On a scale from 1 to 7, how would you rate the usability of this interface?*”, through the post task questionnaire. For the interface effectiveness evaluation, we asked, “How effective was this user interface in helping you find relevant documents?” on a Likert scale, with options being: “Not at all”, “A little”, “Moderately”, “Very”, and “Extremely”. Each option was mapped into a number, ranging from one to five, and the responses were added to a box plot.

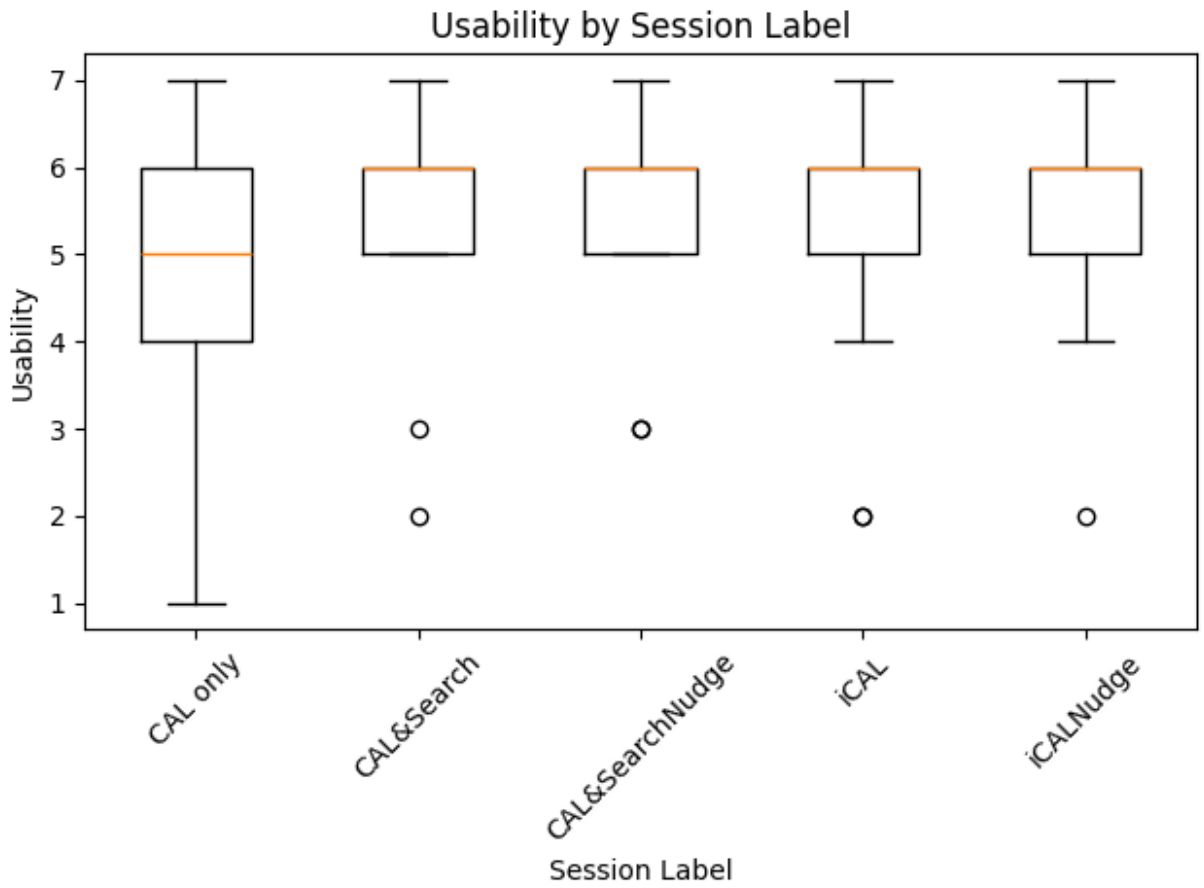


Figure 6.5: On a scale from 1 to 7, how would you rate the usability of this interface?

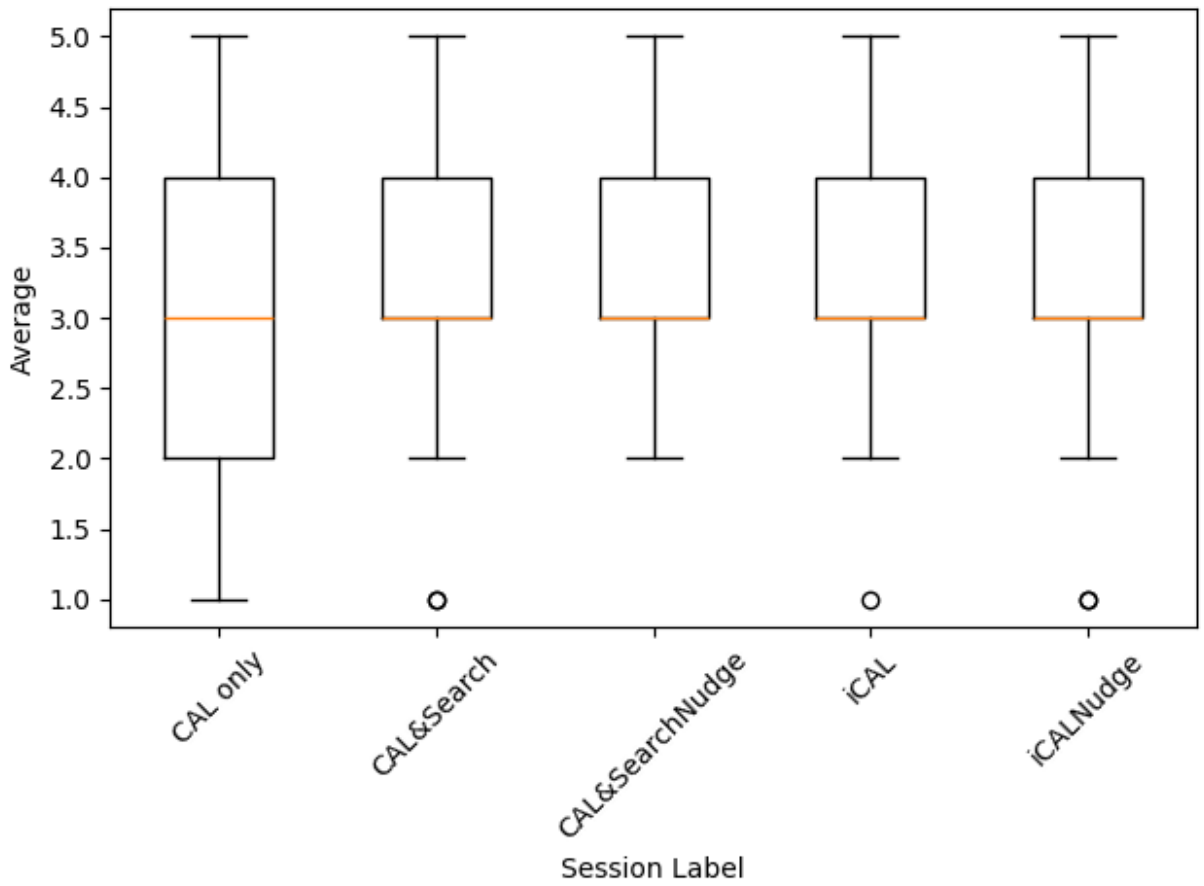


Figure 6.6: How effective was this user interface in helping you find relevant documents?

In terms of usability (Figure 6.5), all configurations involving search (CAL&Search, CAL&SearchNudge, iCAL, iCALNudge) received higher median ratings compared to the CAL-Only interface. Notably, CAL-Only had a wider range of responses and a lower median (5), indicating greater variability in user experience and some clear dissatisfaction. It was also interesting to note that our newly designed interface, the integrated CAL (iCAL, iCALNudge), was consistently rated 6 or above by most participants, similarly to CAL&Search interface, highlighting a similar user perception for both.

A similar trend is seen in interface effectiveness (Figure 6.6), where CAL-Only again shows a lower and more variable distribution of scores. While all other configurations cluster around a median of 3, the CAL-Only setup shows both lower minimums and wider variance. This further supports the notion that while CAL-Only can sometimes be effective, it often leaves users feeling helpless and less confident in the system’s ability to help them find relevant documents.

These results highlight the benefits of using both search and CAL together, no matter which interface is chosen, and underscore the usability limitations of using CAL in isolation.

6.3 Qualitative Analysis

For a qualitative analysis of our study, we used the questionnaire given to the participants, namely the “Post Experiment Questionnaire”. Although the form had numerous questions, for this section we will just go over the three open-ended questions. The questions mainly inquired about what the participants thought the most useful feature was in finding relevant documents, what they liked, and what they disliked. The exact questions were as follows and the answers can be seen in appendix A.

- Which feature was the most useful in finding documents? Explain. (see [A.1](#))
- What did you like about the study? (see [A.2](#))
- What did you dislike about the study? (see [A.3](#))

6.3.1 Most useful feature

CAL, also referred to as “Discovery” in the study, was highlighted by the majority of the participants as the most useful feature that helped them in finding relevant documents. Participants liked how it “became very accurate near the ending”, “yielded more accurate results”, “give very related articles that cannot be found by keyword searching or are buried very deep in search”, and “more personalized” etc. Some of the participants also highlighted that CAL made them “less fatigued” and “easy to focus”.

Several of the participants identified the highlighting feature as being the most useful. This feature allowed users to input words, separated by spaces, and automatically highlighted all instances of those words within a document. The highlights persisted across documents and sessions, eliminating the need for users to re-enter terms repeatedly. The feature helped participants quickly scan for the relevant terms within the document, highlighting the important parts. Users reported that the feature “makes it convenient to identify keywords in documents”, “makes it so much faster to judge documents”, and “useful for quick checks” etc.

Moreover, we could also see some differences in opinions regarding which interface the users preferred. Some liked having CAL and Search on the same page, in the form of Integrated CAL. Others specified their favorite interface as “when the discovery tab was separate from the search tab”. There was also a minority of participants who preferred

the Search feature over CAL, with one person highlighting how they could “see multiple documents and skip uncertain ones.”

6.3.2 Most liked

For this question, participants had diverse inputs, and they covered a variety of things. Many participants liked that the system was “user-friendly”, “simple”, “intuitive”, “easy”, “convenient”, and “easy to use”. For this question as well, a lot of people highlighted how helpful CAL was and how it “simplified decision-making and saved mental effort”. They also liked how CAL adapted to their information need and how it got better with time. Some also highlighted how they liked the freedom to search for anything related to the topic and the ability to switch between CAL and search whenever needed.

Some participants also reported how they liked certain topics, highlighting them as “new”, “interesting” and “thought-provoking”. Others remarked how they read about topics they would not have done otherwise, and learned about “unfamiliar areas”.

Participants also liked the structure of the user-study, and a few participants reported how the tutorial was helpful in understanding the system and getting familiar with it. They also liked features like highlighting and query-biased snippets in search. One participant even remarked how got “insight into how to better phrase search queries”.

6.3.3 Most disliked

Although participants touched on a variety of elements, there were numerous common themes. Most participants reported that the one-hour search task was too long and a smaller session of 30 to 40 minutes would have been more convenient. They reported that

the long sessions were “tiring”, “monotonous”, “tedious”, and how they lost interest for some topics. A participant also reported how the documents they marked towards the end of the task might have been poor quality due to the long task length.

Participants also raised concerns with some of the topics. A participant reported how the topics involving “Cults” and “Deaths” could have come with “content warnings”. Others remarked how the topics were very “specific”, “confusing”, and how it was difficult to judge documents for unfamiliar topics.

A few participants did not like the pop-up that inquired if they were still working, noting its sound as “annoying”, and the pop itself as “too frequent” and “constant even when working”. Participants also disliked seeing redundant articles within the test collection. Some would have liked having a “Not sure” button for judging documents.

Chapter 7

Conclusion

The study tested five different configurations for an information retrieval system to investigate the effects of two main factors: an Integrated CAL interface that combines both search and CAL together into a single page, and the availability of CAL from start versus an initial search and then a nudge towards CAL. We also investigated the effects for when search was completely removed from the interface. We evaluated the system's performance using a variety of both qualitative and quantitative evaluation measures to move towards creating a user-interface that better aligns with users' needs and wants.

Contrary to the findings of [Zhang et al. \(2018\)](#), we found that search does not significantly harm, and can sometimes even improve performance. For all of our observed performance measures including number of user-reported relevant documents, NIST-marked relevant documents found, recall, precision and F1 score etc., we found that the performance of all four interfaces that used search was similar to, if not slightly better than the CAL-only interface. We also found statistically significant differences in precision and precision at minimum number of user reported relevant documents, whereby inclusion of

search significantly helped performance. Moreover, the average time taken to find the first relevant document when search was available was 50% less than the time taken when working with CAL-Only interface, highlighting how helpful search can be in early stages to counter the initial learning phase in CAL. We also saw this same effect in the recall-effort curve, where there were significant differences in performance between the search-based and CAL-only configurations. The study confirms that search might not be as harmful as previously considered, and can even be helpful in many cases.

Our experimental design to test the results of priming (“Nudge” interface) did not lead to significant improvements in performance, like we had hypothesized and observed in simulation (chapter 4). This was highly unexpected, as the simulation showed considerable improvements in recall. We believe this may have been due to the way we designed our experiment. During the tutorial, we explained to users what CAL was and how it worked on relevance feedback, whereby the more relevant documents they find, the better it will get. And that they could use search early on to seed the model with examples of relevant documents. Upon a careful examination of data, we found that even when CAL was not disabled initially, users still went on to search early on, judging approximately three out of the first five documents in search (figure 6.10). This shows a natural tendency of the participants to seed the model, indicating how the tutorial instructions and the experiment design might have contributed to their understanding of using search as a tool to seed CAL, that could have been missing in the earlier experiment. In addition, even when search was enabled, participants marked 70.7% of their documents through CAL, indicating that search was likely used as a tool to support CAL. This also suggests that instead of restricting the interface, we can also educate users on how to use the system properly to achieve similar outcomes. Moreover, an overwhelming majority of participants highlighted how they preferred CAL to search, indicating that even when both tools are

available, users tend to use them strategically. This further supports the notion that a more effective user-experience can be achieved through guidance, rather than restrictive design.

Although, through the post experiment questionnaire, many users indicated CAL as the most important tool, a highly consistent result and feedback that we got was how important search is for the usability and the perceived effectiveness of the system. All four systems that included both search and CAL were rated consistently higher than CAL-only on the usability scale (figure 6.5). All four systems received a median rating of 6, whereas CAL-Only's median rating was 5, with ratings as low as 1 observed. Ratings for systems that included search were higher and much more consistent. A similar trend was observed for the interface effectiveness question (figure 6.6) where, although the median was 3 for all 5 configurations, CAL-Only had much lower lows. We believe that this was caused by the inconsistent nature of CAL-Only interface. Upon analyzing performance for individual tasks, we observed that in the scenarios where CAL-Only worked, it performed exceptionally well, however, in other cases, it often failed to deliver consistent results. In many cases, users had to wait a considerable amount of time before finding their first relevant document. This delay was not present in conditions where search was available, as users could just type a good query to quickly find relevant documents. Additionally, some users experienced long stretches of time when they did not find any relevant documents, during which the option to switch to search could have helped them regain momentum. The notion was also observed in the post experiment questionnaire, where many people mentioned how they liked having access to both tools, and how a user expressed their frustration with CAL by saying "stumps at times caused me to click 'not relevant' for 5+ articles at a time...". All these factors highlight the limitations of relying solely on CAL, and underscore the importance of complementary tools like search. Without such tools,

users could feel frustrated and helpless.

We had hypothesized that using BM25 search to find 5 relevant documents to manually seed the CAL model before starting the document review process would shorten the model building phase and allow users to find more relevant documents quickly. Although we observed the effects of priming in our simulation (chapter 4), we did not find any statistically significant differences in performance for the “CAL disabled initially” (priming) factor. As a result, we were unable to confirm the hypothesis. Nevertheless, the findings suggest a more nuanced outcome: search can improve performance in several cases, as highlighted earlier in this chapter, and is a highly desired feature, contributing to both the perceived user-effectiveness and the usability of the interface.

Through our experiment, we also introduced a new user-interface, Integrated CAL. The interface combined both search and CAL tools together into one page for a more seamless integration. We hypothesized that the iCAL interface would allow for easier switching in between the two tools, by reducing the need to switch pages. The results from the study confirmed our hypothesis, whereby users in integrated CAL interface were more than twice as much likely to switch between CAL and search as compared to the normal interface. Moreover, the increased switching did not result in a loss of performance. The results from table 6.2, 6.3, and 6.4 etc. prove that iCAL based configurations performed similarly to, if not better, other configurations. Moreover, insights from a post experiment questionnaire showed split opinions on the interface. Some people preferred the old traditional interface with both search and CAL tools on different places, others preferred iCAL, because of its integration in regular search results. Overall, our experiment concludes that there is no statistically significant evidence on whether one is better than the other in terms of performance, and it all comes down to what users prefer to use. Hence, for scenarios where both search and CAL tools are needed, the study offers an additional, and just as effective,

user-interface.

Another interesting point we observed through the post task questionnaire was the importance of the “highlighting”. What we considered to be a small quality of life improvement, turned out to be a highly liked and used feature. Many people, through the questionnaire, underscored the importance of highlighting in finding relevant documents quickly, noting how it makes it easier to identify keywords, and faster to judge documents.

7.1 Summary of Contributions

We conducted a controlled user study with 40 participants to evaluate the performance of five different user-interface configurations of an information retrieval system, building on the prior work by [Zhang et al.](#). Contrary to their experiment, our findings offer a more optimistic view on integrating search with CAL. We found that using a hybrid system that includes both search and CAL can perform just as well, if not better, than a CAL-Only system. Moreover, we found evidence that the inclusion of search, when paired with the right training, can significantly improve initial performance and the overall user experience. Furthermore, we also observed how although, CAL-Only systems can sometimes perform very well, they are extremely inconsistent and the lows can be very low, and users prefer systems that include both search and CAL. Lastly, our newly developed user-interface, “Integrated CAL”, allowed for effortless switching between search and CAL without any cost on the performance, and was well received by participants. In conclusion, our findings reinforce the importance of hybrid High Recall Information Retrieval systems, including both search and CAL, that provide maximum control to users.

Despite the strengths of this study, several limitations should be acknowledged. First, our sample size of 40 participants may have limited our ability to detect statistically sig-

nificant differences across conditions, especially for performance metrics like recall and NIST-marked relevant documents found, where the p value was less than 0.15. Additionally, the tutorial emphasized the importance of using search to seed CAL early on, which may have led participants in both the Nudge and non-Nudge conditions to adopt similar priming behaviors. As a result, the intended treatment effect of the Nudge interface may have been effected. Some topics, involving sensitive themes such as cults and murder, were also reported as uncomfortable for participants to work with and may have impacted task performance. Finally, the one-hour time limit per task was frequently described as burdensome, and a shorter task duration, such as 30 minutes, may have helped.

Future research could further investigate the effects of seeding. Large Language Models (LLM) can be employed to form more comprehensive seed queries to start-up CAL, that cover a variety of different words to test improvement in retrieval quality. LLM guided search queries can also be used to help out with search and speed up querying times. Moreover, the impact of using diverse seeding strategy on performance is also worth investigating. Lastly, a between subject experiment with different instructions to the participants can be conducted to further investigate the effects of priming.

References

- Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D Smucker, Gordon V Cormack, and Maura R Grossman. A system for efficient high-recall retrieval. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1317–1320, 2018.
- Mustafa Abualsaud, Kamyar Ghajar, Linh Nhi Phan Minh, Dake Zhang, Irene Xiangyi Chen, Mark D Smucker, and Amir Vakili Tahami. Uwaterloomds at the trec 2021 health misinformation track. In *TREC*, 2021.
- James Allan, Donna Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel, and Ellen M Voorhees. Trec 2017 common core track overview. In *TREC*, 2017.
- John Bace. Cost of e-discovery threatens to skew justice system. *Gartner RAS Core Research Note G*, 148170, 2007.
- Jason R Baron, R Braman, K Withers, T Allman, M Daley, and G Paul. The sedona conference® best practices commentary on the use of search and information retrieval methods in e-discovery. In *The Sedona conference journal*, volume 8, 2007.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67:1–48, 2015.

- David C. Blair and M. E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM*, 28(3):289–299, March 1985. ISSN 0001-0782. doi: 10.1145/3166.3197. URL <https://doi.org/10.1145/3166.3197>.
- William S. Cooper. Exploiting the maximum entropy principle to increase retrieval effectiveness. *Journal of the American Society for Information Science*, 34(1):31–39, 1983. doi: <https://doi.org/10.1002/asi.4630340106>. URL <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630340106>.
- Gordon V. Cormack and Maura R. Grossman. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, page 153–162, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450322577. doi: 10.1145/2600428.2609601. URL <https://doi.org/10.1145/2600428.2609601>.
- Gordon V. Cormack and Maura R. Grossman. Autonomy and reliability of continuous active learning for technology-assisted review, 2015. URL <https://arxiv.org/abs/1504.06868>.
- Gordon V Cormack and Mona Mojdeh. Machine learning for information retrieval: Trec 2009 web, relevance feedback and legal tracks. In *TREC*, 2009.
- Da Silva Moore v. Publicis Groupe. 287 F.R.D. 182, S.D.N.Y., 2012.
- Maura R Grossman and Gordon V Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Rich. JL & Tech.*, 17:1, 2010.

- Maura R Grossman and Gordon V Cormack. Grossman-cormack glossary of technology-assisted review, the. *Fed. Cts. L. Rev.*, 7:85, 2014.
- Maura R. Grossman, Gordon V. Cormack, and Adam Roegiest. Trec 2016 total recall track overview. In *Proceedings of the 25th Text REtrieval Conference (TREC 2016)*, Gaithersburg, Maryland, USA, 2016. National Institute of Standards and Technology (NIST).
- William E Hick. On the rate of gain of information. *Quarterly Journal of experimental psychology*, 4(1):11–26, 1952.
- Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21*, page 2356–2362, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463238. URL <https://doi.org/10.1145/3404835.3463238>.
- Xinyu Mao, Bevan Koopman, and Guido Zuccon. A reproducibility study of goldilocks: Just-right tuning of bert for tar. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval*, pages 132–146, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-56066-8.
- 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: 347–386, 2010.

- Jeremy Pickens, Tom Gricks, Bayu Hardi, and Mark Noel. A constrained approach to manual total recall. In *TREC*, 2015.
- Jeremy Pickens, Tom Gricks, Bayu Hardi, Mark Noel, and John Tredennick. An exploration of total recall with multiple manual seedings. In *TREC*, 2016.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL <https://www.R-project.org/>.
- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Text Retrieval Conference*, 1994.
- Adam Roegiest, Gordon V. Cormack, Maura R. Grossman, and Charles L.A. Clarke. Trec 2015 total recall track overview. In *Proceedings of the 24th Text REtrieval Conference (TREC 2015)*, Gaithersburg, Maryland, USA, 2015. National Institute of Standards and Technology (NIST).
- Herbert L Roitblat, Anne Kershaw, and Patrick Oot. Document categorization in legal electronic discovery: computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1):70–80, 2010.
- Nima Sadri and Gordon V. Cormack. Continuous active learning using pretrained transformers, 2022. URL <https://arxiv.org/abs/2208.06955>.
- G. Salton, C. Buckley, and E. A. Fox. Automatic query formulations in information retrieval. *Journal of the American Society for Information Science*, 34(4):262–280, 1983. doi: <https://doi.org/10.1002/asi.4630340406>. URL <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630340406>.

- Evan Sandhaus. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752, 2008.
- Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the international conference on intelligent analysis*, volume 2, pages 2–6. Washington, DC., 2005.
- Jacob Tingen. Technologies-that-must-not-be-named: Understanding and implementing advanced search technologies in e-discovery. *Richmond Journal of Law Technology*, 19: 1, 2012-2013.
- Andrew Turpin, Yohannes Tsegay, David Hawking, and Hugh E. Williams. Fast generation of result snippets in web search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 127–134, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595935977. doi: 10.1145/1277741.1277766. URL <https://doi.org/10.1145/1277741.1277766>.
- UWaterlooIR. Gathera. <https://github.com/UWaterlooIR/gathera>. URL <https://github.com/UWaterlooIR/gathera>.
- Ellen M. Voorhees. The philosophy of information retrieval evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-45691-9.
- Ellen M Voorhees. Overview of the trec 2004 robust retrieval track. In *Trec*, 2004.
- Bodo Winter. A very basic tutorial for performing linear mixed effects analyses. *arXiv preprint arXiv:1308.5499*, pages 1–22, 2013.

Eugene Yang, Sean MacAvaney, David D. Lewis, and Ophir Frieder. Goldilocks: Just-right tuning of bert for technology-assisted review. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, page 502–517, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-030-99735-9. doi: 10.1007/978-3-030-99736-6_34. URL https://doi.org/10.1007/978-3-030-99736-6_34.

Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Daniel Campos. On the reliability of test collections for evaluating systems of different types. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2101–2104, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401317. URL <https://doi.org/10.1145/3397271.3401317>.

Haotian Zhang, Mustafa Abualsaud, Nimesh Ghelani, Mark D. Smucker, Gordon V. Cormack, and Maura R. Grossman. Effective user interaction for high-recall retrieval: Less is more. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 187–196, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3271796. URL <https://doi.org/10.1145/3269206.3271796>.

Haotian Zhang, Gordon V. Cormack, Maura R. Grossman, and Mark D. Smucker. Evaluating sentence-level relevance feedback for high-recall information retrieval. *Information Retrieval Journal*, 23(1):1–26, February 2020. ISSN 1573-7659. doi: 10.1007/s10791-019-09361-0. URL <https://doi.org/10.1007/s10791-019-09361-0>.

Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research*

and Development in Information Retrieval, SIGIR '98, page 307–314, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130155. doi: 10.1145/290941.291014. URL <https://doi.org/10.1145/290941.291014>.

APPENDICES

Appendix A

Forms

Table A.1: Participant Responses to: *Which feature was the most useful in finding documents? Explain.*

#	Response
1	Discovery
2	The discovery tab, it was large text and easy to focus, with big buttons and overall I felt it yielded more accurate results.
3	The CAL search feature became very accurate near the ending, definitely useful.
4	Discovery, as it led to me finding the most documents and the most relevant information.
5	The highlight and the summary on top was useful. I didn't have to waste much time reading the full document. My throughput was higher because of these two.
6	The discovery feature.
7	Discovery.

#	Response
8	I was less fatigued with the Discover/CAL option, but I don't know if it was more effective.
9	I believe the discovery feature was the most useful given its ability to filter through what I have reviewed and recommend me documents based on my previous reviews.
10	I like how you could switch between CAL and regular search feature. I think having CAL at the top but still having the other search results present is good.
11	The "highlight" feature helped me find keywords quickly.
12	The ML was the most useful because, for the most part (excluding the Cuban sugar exports topic), it did eventually find a pattern of articles that were relevant to the topic.
13	I liked the integrated discovery tool.
14	Discovery feature, and the buttons to categorize the documents.
15	The discovery tool is useful; it does try to find documents with similar data.
16	CAL generally feels better. Occasionally it could give very related articles that cannot be found by keyword searching or are buried very deep in search. A combination of both could certainly improve efficiency. It made me happier not relying on a single tool.
17	The Discovery feature - gives you one article at a time to review and improves with time, is more personalized.
18	Discovery and CAL.
19	Highlight feature was really helpful. Knowing how many times the word I'm looking for appears can speed up the identification process.
20	Discovery.

#	Response
21	Search with CAL and Discovery.
22	Highlight tool and Discovery, I was able to go through the documents faster and look at more articles instead of going back and forth on the search page.
23	Keyword highlight feature was good; search was good, and I think CAL accuracy was really good as well.
24	My favorite interface was the one in which the discovery tab was separate from the search tab, and discovery opened only after 5 relevant articles.
25	CAL was really useful in suggesting related documents.
26	Search.
27	The highlight feature makes it convenient to identify keywords in documents, making it easier to spot relevant topics.
28	CAL – this refreshes the fastest, whereas search keeps showing already ranked documents.
29	CAL was about even with the search feature. CAL would still give a fair amount of irrelevant stuff, so maybe a narrowing feature would help.
30	- In order of usefulness: Integrated CAL, Discovery, Search. Search was rarely helpful. Discovery sometimes tunneled on irrelevant things like Quebec for hydroelectric. Integrated CAL was optimal for switching between strategies.
31	Highlight feature makes it so much faster to judge documents.
32	I prefer the search tool the most.
33	I would suggest using the highlight feature; it's really useful for quick checks.
34	CAL, it's easier to switch between documents.

#	Response
35	CAL was very good when it had a large sample of what I found relevant, but search was also very good for some topics.
36	It was mostly CAL; it seemed more useful overall. Search was hard to use contextually.
37	I found the most useful to be the discovery ML tool that first let me choose a few articles before making suggestions.
38	I liked having the search to understand the topic, but CAL was helpful once accurate. Overall, I'd prefer integrated CAL.
39	I liked the search feature because I could see multiple documents and skip uncertain ones.
40	The discovery feature was the most useful. Once the model picked up relevance, navigation was much easier without having to reopen topics.

Table A.2: Participant Responses to: *What did you like about the study?*

#	Response
1	Interface was very user friendly and the information was easy to grasp
2	I liked the unique platform of this study, and how we are the sole shapers of the results we yield. It was really interesting to see how the platform reacted to my answers.
3	It was nice and simple
4	I liked how original and intuitive everything was. Certainly a tool that I can see myself using.

#	Response
5	I like the setup of the experiment. It encourages me to explore different strategies to find as many documents as I can in an hour.
6	It was pretty easy to differentiate
7	Interesting to see how discovery is able to guide so well.
8	Even though “Search” mode had more options, I started craving “Discovery” mode because it simplified decision-making and saved mental effort.
9	I liked using the discovery feature!
10	The intuitive features of the program.
11	I liked the fact that the topics were new to me and mostly geo-political issues, which I found interesting.
12	The articles were interesting; I learned a few things. The suggestions via ML made it easier to find and judge relevant articles.
13	The study employed different cases for each task, helping to reduce potential biases.
14	The documents were interesting to read.
15	It is fun to read about the different topics.
16	The concept is fun. Personally, I’d like to see advancements in document search and classification systems.
17	The interface was easy to use, and some topics were interesting and fun to read about.
18	It’s very interesting.
19	Very comprehensive in terms of options and the UI was great—easy to use and learn.
20	Fairly easy to understand and complete.
21	I like how finding documents got easier the more detail I put into the search.

#	Response
22	It would be helpful for researchers. I enjoyed reading about different topics.
23	The study was interesting and I understood how we assess, classify, and categorize information.
24	The variety of topics ensured that different backgrounds were reflected. A good interface for both experts and novices.
25	Discovery and CAL were helpful in getting refined searches.
26	Having the freedom to search anything related to the topic.
27	The ability to switch between search and discovery was good depending on relevance.
28	There were a limited number of tasks.
29	It suggests stuff closer and closer to relevance over time.
30	- Topics were diverse and interesting. - I liked being able to choose search method. - The UI had many QOL features (e.g., highlight, top snippets).
31	I like the interface where you can review documents one by one.
32	Working with the interface was easy and convenient.
33	The tutorial helped and the functions were easy to understand.
34	The continuous learning of CAL, which eventually showed more related documents.
35	Gave me insight into how to better phrase search queries.
36	I liked the engagement throughout and using different search methods.
37	I liked being able to take breaks and that the tutorial familiarized me before the test.
38	Going through all the conditions helped me understand how retrieval works.
39	I liked seeing new, thought-provoking topics I'd never considered before.

#	Response
40	Topics matched my interests and helped me learn more about otherwise unfamiliar areas.

Table A.3: Participant Responses to: *What did you dislike about the study?*

#	Response
1	Some topics were harder to judge since I was unfamiliar with it
2	The difficulty and stumps at times caused me to click “not relevant” for 5+ articles at a time...
3	It’s no one’s fault but the fact that the site wouldn’t work properly twice was a little annoying
4	Nothing in particular
5	An hour was a bit long. The quality of the document chosen toward the end might have been compromised on my end.
6	Not much
7	The constant pop-up even if I am working
8	I found the one-hour time chunks too long. Around the 40-minute mark I would always check the remaining time more.
9	N/A
10	The topics themselves could have come with content warnings (e.g., cults or deaths).
11	Nothing in particular. It was a good experience overall.

#	Response
12	I could tell why the ML was suggesting irrelevant articles, but couldn't tell it *why* it was wrong.
13	I expected the ML-based engines to be more robust, but they were sensitive to exact phrasing.
14	The 1-hour duration is a long time to study one single topic.
15	Many articles were near-topic but still not truly relevant.
16	No skip/unsure button. CAL doesn't update after new judgments. Sessions are too long. Lacks advanced search. CAL keeps giving unrelated documents.
17	Some of the topics were unclear or uninteresting and difficult to find relevant documents.
18	Sometimes frustrating to reread documents just to decide relevance.
19	Articles were about CD piracy instead of water piracy.
20	Sessions were very long and it was easy to get distracted.
21	Sometimes documents were repeated.
22	Lengthy. Some articles were repetitive. Pages sometimes loaded slowly.
23	Too much reading; one hour straight was tiring. Timer felt longer than it was.
24	Discovery ran out of relevant articles after a while, even though it tried adapting.
25	Long and monotonous.
26	Task duration was too long. Half an hour would have been enough.
27	Some topics were hard to understand, making relevance judgments difficult.
28	Tasks were too long.
29	Even after using CAL, it still took time to find what's relevant.

#	Response
30	1 hour is tedious. Some topics were confusing (e.g., hydroelectric futures). Cuba sugar topic was especially frustrating.
31	Highlight feature didn't work unless you clicked into the file and re-selected highlight.
32	Topics were too specific. Some articles didn't contain obvious keywords.
33	Alert sound was annoying and too frequent.
34	Absence of visual content made experience less enjoyable.
35	Lengthy, and some articles repeated or didn't make much sense.
36	Topic wording was occasionally confusing.
37	Fonts were too small; lack of dark mode; many redundant articles.
38	Study duration too long. Lost interest during some boring topics.
39	Topic descriptions weren't detailed enough to judge relevance confidently.
40	The study was well conducted. No issues.

Appendix B

Tutorial



Experiment Tutorial

Abdul Manaam
amanaam@uwaterloo.ca

Department of Management Science and Engineering
University of Waterloo

Scientific Research

- This is a scientific research study, and its purpose is to understand human performance using the available tools.
- This experiment is for my Master thesis research.
- We have designed the system to collect high-quality data from you.
- If you're not interested in fully engaging with the experiment, we kindly ask that you excuse yourself instead of continuing.

Your Task in This Study

- For this study, you will work to find and label documents using our specially designed review tools.
- You will use a search system to find relevant documents.
- The system has both a traditional search engine as well as a “discovery” (CAL) machine learning tool.
- For each of five search tasks, your goal will be to find as many relevant documents as possible within one hour.

Search Tool

gathera | Search Discovery Review | explore

Time Spent
Topic Details

Explorer Tires Had to Carry A Heavy Load Firestone tire
doc_id: 1224704
Explorer Tires Had to Carry A Heavy Load Firestone tires used on Ford Motor Co's Explorer have fairly low weight-carrying capacity for a vehicle that size, particularly at low inflation pressure recommended by Ford, officials at Bridgestone/Firestone have been saying since recall that 15-inch tires are losing their treads at high speeds partly because drivers are overloading their vehicles; table (M) The Firestone tires t

Highly Relevant
Relevant
Not Relevant

Under the Hood List of program files loaded by Microsoft
doc_id: 099922
Under the Hood List of program files loaded by Microsoft's Internet Explorer and, for few key files, description of what they do; Microsoft contends that without Explorer, its Windows 95 operating system and features of other software will not work; photo (M) The Justice Department is seeking to prevent Microsoft from making computer manufacturers load Internet Explorer in new machines. ... Microsoft says, however, that w

Highly Relevant
Relevant
Not Relevant

Ads for Ford's New Explorer Sidestep Safety Issue Ford
doc_id: 1282609
Ads for Ford's New Explorer Sidestep Safety Issue Ford Motor Co's new advertising drive for redesigned 2002 Explorer sport utility vehicle will ignore recent safety concerns and instead present vehicle as way for customers to enrich their lives, will not mention design changes made to reduce risks of rolling over or killing other motorists in crashes (M) When it comes to image polishing, the Ford Explorer sport utility ve

Highly Relevant
Relevant
Not Relevant

Explorer Model Raises Doubts About Safety Consumer advo
doc_id: 1284485
Explorer Model Raises Doubts About Safety Consumer advocates and trial lawyers call for Ford Motor Co to do more to warn customers about dangers of driving sport utility, especially stressing high rate of rollovers in its two-door Explorer Sport model; Public Citizen president Joan Claybrook holds Ford's continuing marketing blitz for Explorer, which does not mention safety explicitly, misleads customers into believing th

Highly Relevant
Relevant
Not Relevant

Fatal Explorer Accidents Involving Bad Tires Soared in
doc_id: 1254887

Highly Relevant
Relevant
Not Relevant

Discovery Tool

gathera... Search **Discovery** Review

Not Re... Re... Highly Re...

uv eye 1 1

Plant Study Questions Nature of Ozone Risk
August 25, 1992

Scientists know that some of the sun's ultraviolet, or **UV**, radiation can damage plant and animal tissue in a variety of ways. **UV** radiation at the relatively short frequencies of 280 to 315 nanometers, called **UV-B**, is blocked by the earth's stratospheric ozone layer. Experts fear that the depletion of the ozone layer by synthetic chemicals called chlorofluorocarbons will allow more **UV-B** to reach the earth, damaging plants and causing the incidence of skin cancer and **eye** cataracts to rise.

Show full document

Judge Document
docId: 852247

Not Relevant	Relevant	Highly Relevant
--------------	----------	-----------------

Highlighted terms

uv	13
eye	2

Time Spent Topic Details

Reviewed Documents

- Observatory
- Personal Health: Protection from the ...
- Q&A
- Do Your Skin a Favor: Protect It in Su...
- Ultraviolet Faces Revival to Fight TB
- Patents: Architects debate concepts ...
- How Many Rays Today? Consult Your...

Integrated Discovery Tool

gathera... Search Review tropical storms

Time Spent

Topic Details

CAL

Rain Is a Major Concern As Tropical Storm Surges
docId: 1315233
Tropical Storm Barry is the second named storm of the Atlantic hurricane season. In June, Tropical Storm Allison dumped 20 inches of rain in Louisiana and 32 inches in Texas, much of it as the storm sat for days off the coast, then causing floods as far north as Pennsylvania.

Highly Relevant
Relevant
Not Relevant

Tropical Storm Loses Strength in the Gulf
docId: 0291211
LEAD: Hundreds of workers fled oil rigs in the Gulf of Mexico late Friday and early today as **Tropical Storm Jerry** inched toward Texas and Louisiana, but the **storm** weakened and showed few signs of regaining strength. ... Hundreds of workers fled oil rigs in the Gulf of Mexico late Friday and early today as **Tropical Storm Jerry** inched toward Texas and Louisiana, but the **storm** weakened and showed few signs of regaining strengt...

Highly Relevant
Relevant
Not Relevant

Hurricane Weakens to Storm
docId: 032674
LEAD: The first hurricane of the 1990 Atlantic **storm** season, was downgraded to a **tropical storm** this evening as it headed toward open ocean. ... The first hurricane of the 1990 Atlantic **storm** season, was downgraded to a **tropical**

Highly Relevant
Relevant
Not Relevant

Definition of Relevance

Three point relevance scale:
Highly Relevant/Relevant/Non-Relevant

Assume that you have the information need stated in the topic and that you are at home searching the web for appropriate material.

- If the document contains information that you would find helpful in meeting your information need, mark it **relevant**.
- If the document directly addresses the core issue of the topic, mark it **highly relevant**.
- Otherwise, mark it **non-relevant**.

Source: Voorhees, Ellen M. "Evaluation by highly relevant documents." Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001.

Consistency of Judgments

- Be **consistent** in your judgments. Referring to the topic text from time to time will help you maintain consistency in your judgments.
- The number of relevant or irrelevant documents has no bearing on the "value" or "goodness" of the topic. So, **don't feel that you have to "stretch" your definition of relevant**.
- Document should **not be judged as relevant or irrelevant based only on the title of the document**.
- **Judge duplicate documents the same**. If you judge a document as relevant and find an identical document, judge that document as relevant too.

Source: Voorhees, Ellen M. "Evaluation by highly relevant documents." Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001.

Pitfalls

- Relying too heavily on query/key terms to make your relevance judgments could be misleading and cause your judgments to be inaccurate.
- You should try to work as fast as possible while maintaining your accuracy.
- We expect you as a human to do all the work. Using AI (Chat GPT) will destroy the scientific merit of this research.

Source: Voorhees, Ellen M. "Evaluation by highly relevant documents." Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001.

Procedure of User Study

The study will involve about six hours of work. The first hour is the tutorial, and the next five hours will be dedicated to spending one hour per task.

You can finish the 5 tasks on your computer within one week.

1. As part of today's tutorial:
 - I. You will complete a demographic questionnaire.
 - II. There will be a training task where you will learn how to judge documents.
 - III. Next, there will be a practice task to allow you to get familiar with our search and discovery (CAL) interfaces.
2. There are five 1-hour tasks after the tutorial. Each task consists of:
 - I. A pre-task questionnaire,
 - II. A judging/searching task,
 - III. A post-task questionnaire.
3. After completing the 5 tasks, there will be an End-of-Study questionnaire.
4. After you have completed all five tasks and the tutorial, we will remunerate you for the study.

Proper Behavior of User Study

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, while working on a task:

- If possible, please work on a given task from start to finish. You may take breaks in between if needed. If you take a break while you are completing a task, the timer will automatically pause.
- Minimize distractions as much as possible to stay engaged with the study tasks.
- Activities such as using mobile phones, listening to music, or checking emails can interfere with focus and may affect the study outcomes.
- The system will keep displaying documents until all 1.8 million documents have been judged, meaning the task will not conclude earlier, regardless of how quickly judgments are made.

Demographics Questionnaire



Practice Tasks

mooneye.cs.uwaterloo.ca:3000

System Demo

mooneye.cs.uwaterloo.ca:9000

Questions

Appendix C

Poster

Call for Participants for research in Information Retrieval

- You will be asked to attend a tutorial and complete 5 tasks using a text retrieval system to read and find documents relevant to a topic.
- As a participant in this study, you will also be asked to complete a demographic survey, task-related questionnaires, an exit survey upon the completion of the study, and to review documents on given topics.
- For your contribution to the study, you will receive a \$120 gift card upon attending the tutorial and for completion of all 5 tasks.
- Your participation will take approximately 6 hours. The first hour will be a tutorial and training and the next 5 hours will be dedicated to spending one hour per search task. You will be required to come to the University of Waterloo to take the tutorial/train task. For the 5 tasks, you can complete each task on your own computer or a computer on campus.

For more information regarding the study, please contact:

Abdul Manaam
Department of Management Engineering
at amanaam@uwaterloo.ca

Or visit: <https://forms.gle/NanMV8Lw24vZ1PUx6>

