Learning and Relevance in Information Retrieval: A Study in the Application of Exploration and User Knowledge to Enhance Performance

Harvey Stuart Hyman
University of South Florida, hymanlaw@gmail.com

Follow this and additional works at: http://scholarcommons.usf.edu/etd
Part of the American Studies Commons, and the Library and Information Science Commons
Learning and Relevance in Information Retrieval: A Study in the Application of
Exploration and User Knowledge to Enhance Performance

by

Harvey Hyman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Information Systems and Decision Sciences
College of Business
University of South Florida

Co-Major Professor: Dr. Richard Will
Co-Major Professor: Dr. Manish Agrawal
Dr. Terry Sincich
Dr. Balaji Padmanabhan

Date of Approval:
July 5, 2012

Key Words: eDiscovery, Context, Iterative Feedback, Text Mining, Document Search

Copyright © 2012, Harvey Hyman
# Table of Contents

List of Tables .................................................................................................................. v

List of Figures .................................................................................................................. vii

Abstract ............................................................................................................................. ix

Chapter 1  Introduction ..................................................................................................... 1
  1.1 What is eDiscovery and Why Study it? ................................................................. 3
  1.2 What Makes eDiscovery Unique as a Form of IR? ............................................. 5
  1.3 Three Related Studies ........................................................................................... 7
    1.3.1 Study One: The Relationship of Exploration with Knowledge Acquisition ................................................................. 9
    1.3.2 Study Two Elimination Terms: A Technique for IR Document Filtering ........................................................................ 11
    1.3.3 Study Three: A New System for eDiscovery IR Implementing Context Learning with Relevance Feedback ..................... 11
    1.3.4 Evaluating User and System Performance ..................................................... 12

Chapter 2  Research Methodology .................................................................................. 13
  2.1 Information Systems Framework for Research .................................................... 13
    2.1.1 Design Science Paradigm as Framework ..................................................... 14
    2.1.2 IT Artifact and Prototypes ........................................................................... 16
    2.1.3 Research Model for the Dissertation Studies .............................................. 17

Chapter 3  Data Collection ............................................................................................. 22
  3.1 Population Frame and Sample Selection .............................................................. 22
  3.2 Document Corpus ................................................................................................. 23
  3.3 Data Collection Method ....................................................................................... 24
  3.4 Method of Analysis and Measurement .................................................................. 26
    3.4.1 Document Seeding ....................................................................................... 27
    3.4.1.1 Data Runs ................................................................................................. 28
    3.4.2 Pre-Task IR Behavioral Questionnaires ....................................................... 29
    3.4.3 Verbal Protocols, Interviews, Post-Task Questionnaires and Usability Study ........................................................................ 30

Chapter 4  Information Retrieval (IR) ............................................................................. 32
  4.1 Relevance .............................................................................................................. 33
  4.2 Electronically Stored Information (ESI) ............................................................... 34
    4.2.1 Approaches for Reducing the Search Space .................................................. 35
  4.3 The Problem of Uncertainty and Unstructured Documents ................................ 35
Appendix E Exploration Study Computer Print Screens ........................................... 178
Appendix F Learning/Relevance Feedback Computer Interface Screens ................. 184
Appendix G Pre-Task Questionnaire .................................................................... 186
Appendix H Post-Task Exploration Questionnaire ................................................. 187
Appendix I System Usability Study ...................................................................... 190
Appendix J Table of Covariates, Levels, and Descriptions ................................. 191
Appendix K Table of Independent and Dependent Variables .............................. 192
Appendix L Table of Pearson Coefficient Correlations ....................................... 193
List of Tables

Table 1 Dissertation IT Artifacts ........................................................................................................15
Table 2 Examples of Traditional Legal IR ..........................................................................................46
Table 3 List of Variables Tracked in Exploration Study .....................................................................63
Table 4 List of Behavior Scales ............................................................................................................75
Table 5 SAS 9.2 Printout for Recall Variables ...................................................................................92
Table 6 SAS 9.2 Printout for Precision Variables ...............................................................................92
Table 7 Summary of Exploration Model Results ...............................................................................93
Table 8 SAS 9.2 Printout for Interactive Effect Upon Recall .............................................................94
Table 9 SAS 9.2 Printout for Interactive Effect Upon Precision .........................................................95
Table 10 Summary of Behavioral Model Results ..............................................................................96
Table 11 SAS 9.2 Printout for Recall Variables .................................................................................97
Table 12 SAS 9.2 Printout for Precision Variables .............................................................................98
Table 13 SAS 9.2 Printout for Recall Variables (Interaction) .............................................................99
Table 13.1 SAS 9.2 Printout Precision Variables (Interaction) .........................................................100
Table 14 SAS 9.2 Printout Multi-Collinearity Analysis .....................................................................100
Table 15 Summary of Behavioral-Exploration Model Results .........................................................101
Table 16 SAS 9.2 Printout for Number of Documents Viewed ..........................................................102
Table 17 SAS 9.2 Printout for Time per Document ..........................................................................102
Table 18 SAS 9.2 Printout for Interaction (Time per Document) .......................................................103
Table 19 SAS 9.2 Printout for Interaction (Total Documents Viewed) .............................................104
Table 20 Recall-Precision Correlation ...............................................................................................105
Table 21 List of Hypotheses Supported and Not ..............................................................................108
Table 22 List of Variables and Descriptions in Learning/Relevance Study ..................117
Table 23 Reduction of Non-Relevant Documents ..........................................................120
Table 24 Paired Difference/RBD Results .........................................................................122
Table 25 Reduction in Non-Relevant Documents .............................................................123
Table 26 SAS 9.2 Report for Precision for eDiscovery Learning Experiment ..............150
Table 27 SAS 9.2 Report for Recall for eDiscovery Learning Experiment ....................150
Table 28 Covariates, Levels, and Descriptions .................................................................191
Table 29 Table of Independent and Dependent Variables ..............................................192
Table 30 Table of Pearson Coefficient Correlations .......................................................193
List of Figures

Figure 1 Three Studies ........................................................................................................9
Figure 2 Information Systems Research Framework ..............................................................14
Figure 3 Architectural Design for Study Prototypes ...............................................................17
Figure 4 Executives’ Information Behaviors Research Model ...............................................19
Figure 5 Adapted Information Retrieval Behavior Model .....................................................20
Figure 6 Relevance Confusion Matrix ..................................................................................47
Figure 7 Job Run Process ....................................................................................................66
Figure 8 Model of Artifact IVs ............................................................................................69
Figure 9 Information Retrieval Behavior Model ..................................................................70
Figure 10 TOA Effect upon Recall and Precision ...............................................................76
Figure 11 LOC Effect upon Recall and Precision ...............................................................78
Figure 12 DISPO Effect upon Recall and Precision ............................................................79
Figure 13 PIIT Effect upon Recall and Precision ..................................................................81
Figure 14 Recall over Time .................................................................................................87
Figure 15 Precision over Time ............................................................................................88
Figure 16 System Design Model .........................................................................................118
Figure 17 Schweighofer and Geist Relevance Feedback Model .........................................139
Figure 18 Study Three: Legal Intelligence® Model for eDiscovery IR ................................144
Figure 19 Average Recall and Precision by Iteration .........................................................151
Figure 20 eDiscovery Study User Informed Consent Screen .............................................167
Figure 21 eDiscovery Study User Group Selection Screen ................................................167
Figure 22 eDiscovery Study User Group 1 Instruction Screen ............................................168
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>eDiscovery Study User Group 2 Instruction Screen</td>
<td>168</td>
</tr>
<tr>
<td>24</td>
<td>eDiscovery Study User Group 3 Instruction Screen</td>
<td>169</td>
</tr>
<tr>
<td>25</td>
<td>eDiscovery Study User Group 4 Instruction Screen</td>
<td>169</td>
</tr>
<tr>
<td>26</td>
<td>eDiscovery Study User Demographics Screen</td>
<td>170</td>
</tr>
<tr>
<td>27</td>
<td>eDiscovery Study Retrieval Task Screen</td>
<td>170</td>
</tr>
<tr>
<td>28</td>
<td>Exploration Screen with Countdown Timer</td>
<td>171</td>
</tr>
<tr>
<td>29</td>
<td>Recall Terms User Input Screen</td>
<td>171</td>
</tr>
<tr>
<td>30</td>
<td>Elimination Terms User Input Screen</td>
<td>172</td>
</tr>
<tr>
<td>31</td>
<td>eDiscovery Learning Tool Search Term Selection Screen</td>
<td>173</td>
</tr>
<tr>
<td>32</td>
<td>eDiscovery Learning Document Retrieval Interface Screen</td>
<td>174</td>
</tr>
</tbody>
</table>
Abstract

This dissertation examines the impact of exploration and learning upon eDiscovery information retrieval; it is written in three parts. Part I contains foundational concepts and background on the topics of information retrieval and eDiscovery. This part informs the reader about the research frameworks, methodologies, data collection, and instruments that guide this dissertation.

Part II contains the foundation, development and detailed findings of Study One, “The Relationship of Exploration with Knowledge Acquisition.” This part of the dissertation reports on experiments designed to measure user exploration of a randomly selected subset of a corpus and its relationship with performance in the information retrieval (IR) result. The IR results are evaluated against a set of scales designed to measure behavioral IR factors and individual innovativeness. The findings reported in Study One suggest a new explanation for the relationship between recall and precision, and provide insight into behavioral measures that can be used to predict user IR performance.

Part II also reports on a secondary set of experiments performed on a technique for filtering IR results by using “elimination terms.” These experiments have been designed to develop and evaluate the elimination term method as a means to improve precision without loss of recall in the IR result.
Part III contains the foundation, and development of Study Three, “A New System for eDiscovery IR Based on Context Learning and Relevance.” This section reports on a set of experiments performed on an IT artifact, Legal Intelligence\textsuperscript{®}, developed during this dissertation.

The artifact developed for Study Three uses a learning tool for context and relevance to improve the IR extraction process by allowing the user to adjust the IR search structure based on iterative document extraction samples. The artifact has been developed based on the needs of the business community of practitioners in the domain of eDiscovery; it has been instantiated and tested during Study Three and has produced significant results supporting its feasibility for use. Part III contains conclusions and steps for future research extending beyond this dissertation.
Chapter 1

Introduction

Information retrieval (IR) is the process of determining the presence or absence of a relevant document (or documents) that satisfy an information need (Vanrijsbergen, 1979). It is important because of the increasing reliance upon digital documentation used to record everyday information such as business transactions, agreements, medical records, and other information stored electronically. This increased reliance has led to large volume collections from which relevant documents must be extracted.

Prior research has found that certain IR domains are highly context and content dependent which can lead to under inclusion of relevant documents and over inclusion of non-relevant documents resulting in poor performance when using automated methods alone (Grossman and Cormack, 2011; Oard et al., 2010).

Two problems that have been identified in information retrieval (IR) are high volume of documents in a collection making manual processing infeasible (Baron, 2005) and uncertainty associated with the relevancy of a document (Bates, 1986; Chowdhury et al., 2011). The first problem is an emerging challenge in the field of IR: Finding solutions for handling large amounts of unstructured information which accounts for approximately 85% of electronically stored information (ESI). Examples of unstructured ESI are emails existing as PST files, documents of various formats such as Word, Excel, PDF, PPT and
JPEG, and paper documents scanned into ESI. As the cost of storage continues to go down, the amount of information stored continues to go up and the need to develop effective methods to handle large ESI collections will continue to become more acute.

Ray Lane, Chairman of Hewlett Packard, recently commented about the potential for developing tools and methods to address the organizing and searching of unstructured ESI (September 23, 2011 press conference announcing appointment of CEO Meg Whitman). When addressing the size of a collection to be searched, a manual approach can be used for review when the collection is small and the documents are easily managed by human retrieval. However, as the size of a collection grows to thousands of documents and beyond, a manual approach is less practical and automated approaches are used to reduce the size of the search space.

The second problem associated with IR is uncertainty (Bates, 1986; Chowdhury, 2011). Uncertainty describes the difficulty of matching the query to a document and is one of the central limitations associated with using automated tools. An automated tool can reduce the search space by taking a user query and matching it to the desired documents to be retrieved from the collection. However, the effectiveness of an automated retrieval tool will only be as good as the search terms it is given to match. If the terms are too narrow relevant documents may be missed; if the terms are too broad then too many non-relevant documents will be returned. Two concepts associated with uncertainty are polysemy and synonymy (Deerwester et al., 1990). Synonymy describes the situation of multiple words with similar meanings; polysemy describes the situation of a single word having multiple meanings.
This dissertation focuses on developing and evaluating methods for IR that harness user knowledge for the legal domain, specifically electronic discovery (eDiscovery) – explained in the next section. This research draws on the traditional technique of using search terms and coefficient weighting. This technique is enhanced by combining it with three different methods for leveraging the knowledge of the user: (1) Exploration, (2) Elimination terms, and (3) Learning via Relevance Feedback. The goal is to enhance performance in the IR result. The methods are evaluated by studying user groups in a series of controlled experiments designed to measure IR performance in the legal domain of eDiscovery.

1.1 What is eDiscovery and why study it?

Electronic discovery (eDiscovery) is the process of retrieving ESI documents for the purpose of review for anticipated or actual litigation (Oard et al., 2010). The purpose of eDiscovery is to find relevant ESI documents for use in litigation. The law takes a broad view of relevance (Oard et al., 2010). Given this fact, emphasis tends to be placed more upon recall than precision when evaluating the quality of IR. eDiscovery calls for solutions provided from the two separate disciplines of law and information systems (Hyman and Fridy, 2010). Traditionally, discovery — which is the production of documentation during litigation — was not electronic. It mainly consisted of manual review of paper documents to meet adverse party requests.

In the case of eDiscovery, information retrieval of ESI involves text mining of unstructured documents in high volume collections where there is uncertainty about which parameters will most likely produce the most relevant documents. Documents
vary in format (PDF, GIF, DOC, PPT, EML, PST) and in structure (headers, footers, columnar reports). Prior to eDiscovery, legal IR was structured and predictable. Search and retrieval episodes in litigation cases occurred in two categories: (1) Structured search of legal cases with tools such as Westlaw and Lexis, and (2) Manual search of paper documents in client physical files prior to the proliferation of electronically stored information (Oard et al., 2010). A typical search in the legal domain would be to find a case or series of cases that answer a specifically defined question.

eDiscovery has been identified as an area of importance in the legal domain and draws from the discipline of information systems for solutions. “In an era where vast amounts of electronic information is [sic] available for review, discovery in certain cases has become increasingly complex and expensive” (Judge Shira Scheindlin, Montreal Pension Plan V. BOA, 2010 WL 184312, S.D.N.Y. 2003). The Philip Morris litigation had as many as 32 million emails in the overall corpus; TREC’s version of the Enron corpus is based on the EDRM version 2 and has between 650,000 and 680,000 email objects depending on how one counts attachments; Illinois Institute of Technology’s version of the Tobacco corpus contains between 1 and 2 million objects. In the case of U.S. v Philip Morris, 25 individuals spent 6 months reviewing 200,000 emails one at a time (Baron, 2005).

1.2 What makes eDiscovery unique and distinct as a form of IR?

Four characteristics describe eDiscovery as a form of IR. First, eDiscovery is defined by a typical user who has domain knowledge about the nature of context, content
and concepts associated with the documents being sought (Grossman and Cormack, 2011). The second characteristic of eDiscovery is the premium placed on recall over precision. The reason for this is that the law takes a harsher view of failure to disclose versus disclosing too much. The third characteristic of eDiscovery lies in the sheer volume of information to be reviewed by a human inspector prior to release (Baron, 2005). The fourth characteristic is shared with other forms of high volume IR — uncertainty (Bates 1986). Time and cost associated with searching high volume collections result in the need to reduce the search space.

As computing power has increased and memory costs have reduced, more and more documentation is collected and stored in electronic form. These collections become the subject of retrieval requests when a party engages in litigation — eDiscovery. In most IR episodes, the failure to retrieve a relevant document results in a loss of potential information and nothing more. What makes the retrieval problem particularly acute in eDiscovery is that failure to retrieve a relevant document opens the door to a range of penalties including the possible loss of the legal case. An example of this occurred in Zubulake v UBS Warburg, 217 F.R.D. 309, 322 (S.D.N.Y. 2003). The Zubulake case is regarded as the seminal landmark case for eDiscovery. It is actually a series of cases and decisions dealing with what data a litigant has a duty to preserve, and under what circumstances the party seeking, versus the party providing, must pay for the search and production costs. Zubulake represented a shift in the way courts determine what must be provided and who must assume the cost for search and production.
In eDiscovery, an automated tool can address the problem of volume by producing a smaller set of selected documents to be inspected by humans and thus reduce costs of review (Grossman and Cormack, 2011). Automated tools rely on search terms provided by the user. Effectiveness of retrieval is measured by recall (percentage of relevant documents retrieved from the total available), and precision (percentage of relevant versus non-relevant documents retrieved).

The limitation associated with IR, whether probabilistic or natural language processing (NLP) is the fact that the IR system relies on search terms alone (Oussalah et al., 2008). This limitation is particularly germane to the task of eDiscovery IR where documents being sought have particular meaning and context (Oard et al., 2010; Grossman and Cormack, 2011). This limitation can affect performance of IR tools which can vary widely for a given set of search terms and weighting methods, sometimes resulting in low percentages of recall and precision in retrieval results (TREC Proceedings Legal Track 2010). This dissertation addresses the problem of meaning and context by developing and evaluating methods to harness user knowledge to improve IR performance.
1.3 Three Related Studies

This dissertation investigates some specific challenges unique to eDiscovery. The central problem addressed in the study of IR generally is: How can we retrieve what we are looking for and leave the rest behind? This is typically measured in recall\(^1\) (relevant documents retrieved) and precision (non-relevant documents left behind). The problem addressed in eDiscovery IR is: How can we harness user knowledge about context and contents of the subject matter as well as the nature of the documents contained within the corpus/collection to improve retrieval performance in eDiscovery?

This problem is similar to large scale information search such as web search insofar as numerous documents of various types (mainly web pages and links) exist, but only a small percentage of documents answer the query posed. This often results in retrieval sets containing numerous items that are irrelevant, and the user has to sift through many non-responsive items to find the few that are responsive.

It is distinct from large scale search insofar as the collection being queried is bounded -- meaning it is a finite database of items, usually email files or documents in the form of PDF, Excel, or Word Docs. The items comprising the collection share a commonality in that they were created in association with a transaction or event that has given rise to the information need.

An additional distinction of eDiscovery information retrieval (IR) has to do with the goal of the retrieval. In typical scale free search, such as a web search, the typical

---

\(^1\) In this paper we are focused on the domain of Information Retrieval. In this domain, the term *recall* is defined as the set of documents extracted or “recalled” from a larger collection. This definition is starkly different from what many behavioral researchers are used to such as human memory experiment therefore, we have included this explanation to avoid confusion about the meaning of the word for readers outside this domain.
user’s information need is satisfied with ‘a document’ that meets their information need. In eDiscovery, the user’s information need is to seek ‘all documents’ associated with information need. Unlike scale free search, eDiscovery IR is highly content and context oriented – meaning that unique terms such as colloquial references and local, controlled vocabularies have a direct impact on successful retrieval of relevant documents and successful avoidance of non-relevant documents.

Figure 1 depicts the three studies comprising this dissertation. The first study in exploration informs the third study on learning through relevance feedback. The first study assesses approach of exploration. Study Two investigates the effects of using elimination terms upon the problem of uncertainty. The goal of the first study is to develop and evaluate a method of corpus/collection exploration for knowledge acquisition to improve IR. Study Two evaluates the effect of using elimination terms as filters to reduce the number of non-relevant documents in the retrieval set. The results from the two studies lead to the third study which evaluates a method to harness the user’s knowledge by translating it to a system built to support user learning. Based on methods developed in Study One and Study Two, Study Three applies a learning tool for context and relevance by having a system select documents based on iterative user input, present those documents to the user via a prototype IR tool, and have the user tune the system through identification of relevant and non-relevant documents to improve the IR result.

Context learning and relevance are implemented using an iterative algorithm to select documents based on the user evolving and refining his/her definition of relevance with each new pass, until a threshold level of improvement in performance has been
reached. In this dissertation the number of iterations has been fixed in order to measure the IR result. Both studies involve the use of a prototype IT architecture developed for this dissertation research. Screen captures from the two prototypes developed to support the methods are located in the appendix section.

**Figure 1:** Three Studies

1.3.1 **Study One: The Relationship of Exploration with Knowledge Acquisition**

In eDiscovery as well as other forms of IR there is a trade-off between recall and precision. The wider the retrieval net is cast, the greater the amount of both relevant and non-relevant documents produced. eDiscovery IR is litigation driven; it begins with a request to produce ALL electronic documents associated with a specific topic or event. This feature of eDiscovery lies in the nature of the task — the corpus is a static and defined collection rather than being open ended, such as a web search. It is a targeted collection of information objects.

The IR starts with a corpus of a certain size of ‘X’ objects and is always settled with human inspection (Grossman and Cormack, 2011). As the size of the corpus increases, time and cost associated with inspection (review of documents) becomes more
expensive. In the case of *California v. Phillip Morris*, there were about 10 million hard
copy documents (Oard et al., 2010). This motivates a significant problem in eDiscovery
IR: How to balance the leverage achieved through automated methods against the final
review stage of human inspection. The first experiment of this dissertation is designed to
study this issue. It is behavioral in so far as it seeks to explain how the user’s
performance in IR can be improved.

This study examines the relationship between exploration of a sample selection of
a corpus and IR performance. Intuitively one would think that the relationship between
exploration and performance is positive and direct – the more a user explores items from
the corpus, the greater the performance in the IR result. But what is it about exploration
that affects IR? Are certain factors more significant than others? This study evaluates the
following factors: Aggregate time spent exploring, Number of documents explored, and
Time spent per document.

The fundamental questions this study seeks to answer are: How does exploration
impact IR performance? How can exploration be manipulated to achieve improvement in
IR results? For instance, how much time does a user need to spend, or how many
documents need to be reviewed, in order to observe a measureable or meaningful
difference in performance? At what point does exploration fail to improve performance?

The contribution of this study will be the development of insights and benchmarks
to identify effective methods for exploration that produce improved results in retrieval.
These results translate into savings in time and cost during the human review process —
the most expensive portion of eDiscovery given that the most expert and highly
compensated are assigned to the final review – of great concern to the practitioner.
1.3.2 Study Two, Elimination Terms: A Technique for IR Document Filtering

This study evaluates a technique for filtering as a method to reduce non-relevant documents. Due to domain and subject matter expertise, eDiscovery users typically have a significant idea of what documents are not relevant. The technique is a modified version of traditional filtering. We evaluate the use of this technique to answer the questions: Can the use of elimination terms reduce the number of non-relevant documents in the retrieval? (2) How significant is the reduction in the number of non-relevant documents? (3) Will the use of elimination terms have a negative effect on relevant documents by reducing the number retrieved? (4) How significant is the reduction of relevant documents? The contribution of this study is the demonstration of the impact of adding a separate method for improving precision (reduction of non-relevant documents) versus relying on present methods such as variations to the query through trial and error methods and Boolean operators (Interviews of practitioners). This study focuses on the construct Elimination Terms and how this construct impacts the IR result.

1.3.3 Study Three: A New System for eDiscovery IR: Implementing Context Learning and Relevance Feedback

Research Study Three focuses on the application of context learning and relevance feedback upon the domain of eDiscovery IR. This domain is defined by subject matter expert users and constrained collections, or non-subject matter experts who acquire context and content knowledge through a method such as exploration. This study seeks to answer the questions: How can we harness user knowledge to improve results in eDiscovery IR? What is the relationship between acquired or a priori knowledge and IR performance? How can we design a system to support knowledge translation in the IR process? The contribution of this study is a new and novel approach to eDiscovery
created through a prototype algorithm and computer interface that applies principles and methods for learning context and relevance to the IR domain. The proposition developed is: An approach to implement user generated descriptions and rules with an iterative format, is likely to produce an improved IR result. This study combines a behavioral approach with a design science approach by developing a method to improve user selections and instantiate an IT artifact to support user interaction with the IR system.

1.3.4 Evaluating User and System Performance

The unit of analysis in both studies is the individual. Study One evaluates differences in IR performance of individuals as measured by dependent variables Recall and Precision and independent variables Total Time Exploring, Per Document Time, and Number of Documents.

Study One evaluates performance against two base-line measured results: (1) A random extraction of documents from the corpus equal to the average number of documents extracted by the participants, thereby validating whether human analysis should be used at all versus a simple random extraction; and (2) An extraction of documents using the specific words from the eDiscovery request itself (verbatim) after removing all non-function (stop) words, thereby validating whether human analysis is any better than use of the plain language of the request itself.

Study Three compares individual performance results over a set of ten iterations. The study measures the difference in incremental performance between iterations and the aggregate difference in performance between the first and the final iteration. The results are measured on two dimensions: iteration instance and Precision and analyzed using a repeated measures/random block design (RBD).
Chapter 2

Research Methodology

2.1 Information Systems Framework for Research

Hevner, March, and Park in their 2004 MISQ article “Design Science in Information Systems Research,” proposed a conceptual framework for performing information systems research. This approach is depicted in the Information Systems Research Framework illustrated in Figure 2. The framework guides this dissertation. The goal of information systems research is to produce knowledge that enables the application of information technology. (Hevner and March, 2003).

The framework models two paradigms, behavioral science and design science. The behavioral paradigm represents the rigor in research. The Rigor Cycle provides established theories, methods and expertise from the foundation knowledge base. The research contributes new knowledge generated by the research to the knowledge base. This is represented on the right side of the framework model. The design science paradigm represents the relevance in research. The Relevance Cycle translates requirements from the environment into the research and introduces artifacts into the environment for field testing, ultimately solving a problem in the domain. This is represented on the left side of the model.

“Design science research is motivated by the desire to improve the environment by the introduction of new and innovative artifacts, and the processes for building these
artifacts,” (Hevner 2004, citing Simon, 1996). This dissertation is motivated by the desire to improve the domain of eDiscovery IR by introducing a new and innovative artifact – an architecture supporting two prototypes that harness user knowledge through the implementation of learning context and relevance as a process for eDiscovery IR.

“Good design science research often begins by identifying and representing opportunities and problems in an actual application environment” (Hevner 2004). This dissertation begins by identifying the problems of uncertainty and volume in the eDiscovery IR environment.

2.1.1 Design Science Paradigm as Framework for this Dissertation

The research performed for this dissertation is conducted using the Design Science paradigm. The artifacts created in this dissertation are listed in Table 1. There are two prototype systems that have been built for the research studies evaluating methods for support of user knowledge process in eDiscovery IR. The first system is designed to

Figure 2: Information Systems Research Framework
support user exploration of a collection and measure the impact of that exploration on the IR result. The goal is to instantiate a new method for knowledge acquisition. The second system is designed to collect user defined attributes of relevant and non-relevant documents and display test run samples via a user interactive interface. The system produces document retrieval results based on user generated input in an iterative format until a threshold level of performance has been achieved. The goal is to instantiate a new method for knowledge translation. In this instance, a pre-determined number of iterations are set in order to measure system performance.

The methods define the solution — they are the algorithms and textual descriptions of the approaches (Hevner and March, 2003). The instantiation shows how to implement the constructs, models and methods in a working system — demonstrating “feasibility, enabling concrete assessment of an artifact’s suitability to its intended purpose” (Hevner and March, 2003).

<table>
<thead>
<tr>
<th>ARTIFACT</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construct</td>
<td>Exploration – Knowledge Acquisition</td>
</tr>
<tr>
<td>Construct</td>
<td>Filtering - Elimination Terms</td>
</tr>
<tr>
<td>Construct</td>
<td>Context Learning – Knowledge Translation</td>
</tr>
<tr>
<td>Construct</td>
<td>Relevance</td>
</tr>
<tr>
<td>Model</td>
<td>Executives’ Information Behaviors Research Model</td>
</tr>
<tr>
<td>Model*</td>
<td>Adapted Information Retrieval Behavior Model</td>
</tr>
<tr>
<td>Model*</td>
<td>Iterative Relevance Feedback using Learning</td>
</tr>
<tr>
<td>Method*</td>
<td>System to support Knowledge Acquisition</td>
</tr>
<tr>
<td>Method*</td>
<td>System to support Knowledge Translation</td>
</tr>
<tr>
<td>Instantiation*</td>
<td>Implementation of learning-relevance feedback IR system.</td>
</tr>
</tbody>
</table>

* Artifacts Created in this Dissertation
2.1.2 IT Artifact and Prototypes

The central IT artifact in this case is a system architecture that is used as a foundation for developing the two prototypes evaluated in this dissertation. A diagram of the artifact architecture is depicted in Figure 3. The architecture supports two separate functions. The first function is dedicated for handling data preparation by engaging a file parser. The prepared files are moved to a job process queue. The system performs operations by separate calls. For example, each relevance term or non-relevance term is handled by a separate operator. This allows the system to store the results for each operation.

The second function of the system is dedicated to managing the user-system interactions. The system is designed to accept user input of relevance feedback and iteratively process the collection until a designated retrieval threshold has been reached.

The first prototype is a system designed for users to explore a small sample of a large collection and submit selection criteria based on the conclusions drawn from their exploration. The system randomly selects samples from the full collection of documents. The proposition here is that user exploration produces better understanding of the nature and context of the collection and therefore, better decisions can be made for selecting search criteria. The system supports user chosen terms to process a retrieval result.

The second prototype is a system designed to apply learning to relevance feedback to IR domains where users have gained contextual insight about the collection from exploration or whereby expert users have a priori concept and/or context knowledge and the collection is finite. This system retrieves documents on an iterative cycle based upon user feedback of relevance of the documents presented. The purpose of
the system is to provide the user with a means to improve retrieval performance through relevance feedback. The system provides radio buttons for the user to indicate relevant or not and a text box to record terms that make the document relevant or not. This allows the system to collect correctly identified documents and apply user tuning for documents falsely identified by the system as relevant (false positives).

![Diagram of Software Design and Job Queue](image)

**Figure 3:** Architectural Design for Study Prototypes

### 2.1.3 Research Model for the Dissertation Studies

The research conducted in this dissertation seeks to describe and explain factors impacting the process of eDiscovery IR. The process can be distinguished into two separate phases. The first phase is *knowledge acquisition*. During this phase the user is seeking to understand the nature of the subject matter of interest, context of the corpus
containing the documents of interest, and structure of the relevant documents sought for extraction. The goal of the first phase is to learn about the context and organization of subject documents contained within the collection. The second phase is knowledge translation. During this phase the user is seeking to apply knowledge acquired about the subject matter, corpus and documents to extract the greatest number of relevant documents (recall) and fewest non-relevant documents (precision).

Prototype One is an IT artifact developed during this dissertation to study user knowledge acquisition. It is an application built to support user exploration of a randomly selected set of documents from the corpus.

Prototype Two is an IT artifact developed during this dissertation to study user knowledge translation. It is an application built to support context learning through relevance feedback. The objective of Prototype Two is to apply context knowledge to improve the IR result. It is implemented using an iterative process that allows the user to tune the IR extraction process to increase performance.

The research model used to guide these two studies is adapted from the Executives’ Information Behaviors Research Model (Vandenbosch and Huff, 1997). The model is depicted in Figure 3.1. Vandenbosch and Huff use their model to describe and explain factors affecting executives’ information retrieval behaviors. They propose two distinct behaviors, focused search and scanning search. These two behaviors impact efficiency and effectiveness in performance.

An executive information system model is a close approximation of an eDiscovery system. EIS and eDiscovery are similar in that both circumstances assume users are domain and/or subject matter experts and knowledge of context has significant
impact upon the performance result. EIS users seek solutions to problems in uncertain environments (Vandenbosch and Huff, 1997); similarly, eDiscovery users seek solutions in an uncertain environment – extracting relevant documents from a corpus of uncertainty.

Figure 4: Executives’ Information Behaviors Research Model

The studies conducted within this dissertation seek to: (1) Measure factors that impact recall and precision, and (2) Explain factors affecting eDiscovery IR behavior of individuals. Therefore, the model in Figure 3.1 has been adapted to fit the factors of interest in this research. The adapted model is depicted in Figure 3.2.

The construct of Focused Search is adapted to approximate the focused search of context learning in Study Three. The construct is operationalized using the measured variables of per document time and total number of documents viewed. This study is designed to evaluate a method for IR when a user has acquired knowledge through an exploration method, or when a user possesses a priori contextual or content knowledge.
This construct represents the user who formulates a specific question to solve a well-defined problem (Huber, 1991; Vandenbosch and Huff, 1997).

The construct of Scanning is adapted to approximate the *scanning* behavior of Exploration. This construct represents the user who browses data looking for trends or patterns, seeking a broad, general understanding of the issue in question (Aguilar, 1967; Vandenbosch and Huff, 1997). It is operationalized in Study One using the measured variable *total time spent exploring*.

Efficiency—doing things better according to Huber, 1991-- is adapted in this study for Precision (efficiency in the extraction by avoiding non-relevant documents) and Effectiveness -- being more productive is adapted in this study for Recall (effectiveness in retrieving the maximum number of relevant documents).

---

**Figure 5:** Adapted Information Retrieval Behavior Model
Four scales representing individual differences impacting the latent factors of Learning and Exploration are adapted for use in this dissertation. The scales Tolerance for Ambiguity (TOA), Locus of Control (LOC), Dispositional Innovativeness (DISPO), and Personal Innovativeness toward Information Technology (PIIT), are operationalized using previously validated instruments (Rydell and Rosen, 1966; Levenson, 1974; Steenkamp and Gielens, 2003; Agarwal and Prasad, 1998). The instruments are explained in further detail in the Data Collection and Exploration Study chapters.
Chapter 3  
Data Collection/Experimental Procedure

3.1 Population Frame and Sample Selection

The population of interest in this research dissertation is eDiscovery users. The research presented here explores methods for improving the IR results of eDiscovery users. Two populations representing eDiscovery users have been identified.

Study One focuses on legal professionals and litigation support personnel as scanning/exploration IR users. This study approximates the IR user who does not have an \textit{a priori} mental model for relevance; he/she seeks a broad scanning/exploring of the corpus/collection to gain insight into context and meaning to develop a relevance model. This study uses law students to approximate legal professionals and litigation support personnel — 60 third year law students representing three universities have volunteered to participate in the exploration task. These students are well suited for the study because they have been exposed to eDiscovery concepts in the classroom or have experience through summer clerkships, yet they are relatively less experienced than eDiscovery professionals such as lawyers and paralegals. This allows the study to control for legal experience and litigation expertise. Our goal is to measure the differences between the groups and avoid the expertise biases that legal professionals develop during their litigation experience.
Study Three focuses on legal professionals as focused IR users. This study also uses law students to approximate the legal professionals. Third year law students who have graduated from Study One are used to approximate the focused IR legal professional. This study assumes a significant relationship exists between exploration and IR results produced in Study One. An additional assumption is that the user acquired knowledge in Study One. The main hypothesis in Study Three is that knowledge acquired in Study One will successfully be translated and result in improved IR performance as measured by *recall* and *precision* through a random block design experiment using repeated measures. This study approximates the expertise and biases of domain experts seeking specific documents to meet a constrained mental model of relevance.

The task is designed to have the user provide an initial set of search terms for document retrieval followed by relevance judgments upon the documents retrieved by the system using an iterative process. The iterations provide the user with selections of documents based on the search terms. The user declares relevance and non-relevance to the system by use of radio buttons along with additional search and elimination terms using text boxes. The terms are absorbed by the system and a new selection of documents is presented to the user for judgment.

### 3.2 Document Corpus

The document collection used in this case is the ENRON collection, version 2. This collection has been made available to researchers from The Text Retrieval Conference (TREC) and the Electronic Discovery Reference Model (EDRM). The collection contains between 650,000 and 680,000 email objects depending on how one counts attachments. The collection has been validated in the literature (TREC
Proceedings 2010, Vorhees and Buckland, editors). The Enron collection is a good representation for a corpus of documents sought during litigation. The collection is a corpus of emails formatted in PST file type. The collection is a reasonable approximation of the problem of uncertainty because the emails in the collection contain a variety of instances of unstructured documents, in varying formats (Word, Excel, PPT, JPEG) making retrieval particularly challenging for an automated process. With over 600,000 objects, the collection is also large enough to be a good representation for the problem of volume.

3.3 Data Collection Method

There are five different data collection methods used in this dissertation: (1) IT artifacts to support the task and treatment, (2) Notes taken during physical observations of the users, (3) Pen and paper questionnaires to record the behavioral scales used, (4) Post-task interviews, and (5) Verbal protocols whereby the users are asked to “think out loud” during the experiment.

The two IT artifacts developed and built as prototype systems support the studies in this dissertation.

The first prototype supports Study One and Study Two. It is a computer interface application designed to present a series of screens to support the following actions:

(1) Informed consent protocol which must be agreed to by the participant,

(2) Description of the study,

(3) Task and treatment,

(4) Collection of user selection of search terms,

(5) Collection of user elimination terms.
The second prototype supports Study Three. It is a computer interface application designed to present a selection of documents based on user submitted criteria using an iterative process. The system accepts user relevance feedback to create the next round of selections. The system supports the following behaviors and functions:

(1) The user is given radio buttons to indicate whether a document is relevant or not relevant;

(2) The user is able to give the system hints in the form of identified terms within the document as rules for relevance or non-relevance;

(3) The system performs multiple iterations of document selection based on user feedback until a pre-determined threshold is reached, measured by recall and precision. In this study the number of iterations is fixed at 10, the unit of analysis is the individual, and the design is a repeated measures format.

Data collected from the pen and paper questionnaires were transferred to a spreadsheet and inputted into SAS 9.2 for statistical analysis. This data is used to triangulate the results of the experiments to explain relationships among IR behaviors, eDiscovery user chosen techniques, and IR results produced.

Data collected from observations, verbal protocols, and pre and post-task interviews are used to develop quotes for useful descriptions to the reader for insight into the experiment and to formulate future research questions.
3.4 Method of Analysis and Measurement

SAS 9.2 is the statistical package used for the analysis in this dissertation. Study One investigates how user knowledge is acquired through exploration. Knowledge is measured by IR results produced using dependent variables (DVs): Recall and Precision with three separate linear regression models. The first model is comprised of the artifact independent variables (IVs) Total time Explored, Time per Document, and Number of Documents Viewed to operationalize the exploration construct. The second model is comprised of the behavioral scales Tolerance of Ambiguity (TOA), Locus of Control (LOC), Personal Innovativeness in Information Technology (PIIT), and Disposition Toward Innovativeness (DISPO). The third model is comprised of the artifact exploration variables as dependent variables with the behavioral scales as independent variables.

Study Three investigates how learning combined with relevance feedback can be implemented to support acquired knowledge in user selections for the IR process – knowledge translation. Knowledge translation is measured by differences in the IR result using a repeated measures (random block design) analysis.

The artifacts developed for this dissertation serve two roles. The first role is data collection; the artifacts collect user selections and user demographic information tracked in the studies. The second role the artifacts serve is for the process to support the system. There are two supports. The first support is the presentation of the treatment and task to the user. The second support is submission of user input (data collected) to the job run process. The job run process calls each method serving the constructs operationalized as the user selections – recall terms, and elimination terms. The results produced by each method called are stored in a system result bin and exported as an Excel spreadsheet. For
analysis in this dissertation the results are reported by term occurrence and frequency, and evaluated for recall and precision using analysis of variance between the runs.

Performance in Study One is measured by the dependent variables recall and precision. Independent variables Total Time, Time per Document, and Number of Documents are analyzed for significance as main effects and for interactive effects upon the dependent variables.

Performance in Study Three is measured by dependent variables recall and precision. This study analyzes differences in performance between iterations and for the total differential from the first iteration to the last iteration using a repeated measures design/random block design (RBD).

Both studies control for litigation experience and subject matter experience.

Data collected to measure Locus of Control, Tolerance for Ambiguity, Dispositional Innovativeness, and Personal Innovativeness are analyzed for significance as main effects upon Recall and Precision. Interactive effects are also considered. All four scales are analyzed for reliability using alpha.

3.4.1 Document Seeding

Both studies are concerned with results produced from human choices resulting from acquisition and translation of contextual and subject matter knowledge. We measure the differences in Recall and Precision in the retrieval result. Study One is designed to access how well users are able to identify relevant documents when exploration is offered and when time to explore is manipulated. We use “seeding” to establish a base-line number of relevant documents within our data set. Seeding is a technique that has been used in research studies to improve initial quality for developing algorithms, evaluating
performance and testing software (Burke, et al., 1998; Brown, 2000; Fraser and Zeller, 2010). We accomplish this seeding by randomly selecting 9,000 previously identified non-relevant documents from the 680,000 item collection. A selection of 1,000 documents, previously identified by TREC 2011 as relevant to the eDiscovery task, are added to the 9,000 random items to create a 10,000 document set. The analysis in this case is concerned with the number of relevant documents retrieved and the percentage of relevant documents within the retrievals.

Study Three does not use seeding. This study is concerned with the user’s ability to translate his/her perception of relevance to the system, and the system’s ability to return relevant documents reflecting the user’s criteria. The user is the judge in this study. We report precision per iteration based on how well the system returns documents meeting the user’s criteria for relevance and compare the user’s assessments to the previously judged documents. We also report the total improvement from the first iteration to the final iteration.

3.4.1.1 Data Runs

In Study One we perform three different types of runs. All runs are performed on the 10,000 document set. The goal here is to measure the effects of time exploring the corpus, number of documents viewed, and time per document upon performance as measured by the dependent variables Recall and Precision. The first run is based on the participant selected terms. The second run is a random extraction of documents from the set, equal to the average number of documents produced by the participants. The purpose of this run is to benchmark the participant results against a simple random selection to
determine if “human in the loop” is any better than chance. The third run is an extraction from the set using the verbatim language from the task after deleting the non-function words, such as the, and, or, etc. The purpose for this run is to provide an alternative benchmark, in this case a non-random but equally non “human in the loop” method to compare against the participants’ exploration performance.

3.4.2 Pre-Task IR Behavioral Questionnaires

In this dissertation we use known scales previously validated in the literature to anchor our findings about individuals’ exploration attitudes and techniques. The scales are administered using pre-task questionnaires. We have chosen two scales known to be associated with user IR behavior and two scales known to be associated with innovativeness. The questionnaires are adapted from previously validated item inventories. Two scales associated with user IR behavior in this dissertation are: (1) Tolerance for Ambiguity and (2) Locus of Control (Vandenbosch and Huff, 1997). The two scales associated with innovativeness are: (1) Dispositional Innovativeness (Steenkamp and Gielens, 2003) and (2) Personal Innovativeness toward Information Technology (Agarwal and Prasad, 1998).

This technique was used as a means to verify how well the participant understood the task requested by the study. After review of the eDiscovery task, the participant was asked to complete a short pen and paper questionnaire designed to validate that the participant had a threshold understanding of the problem they were being asked to solve. The rationale was to control for a participant’s poor performance resulting from a failure to understand the task.
3.4.3 Verbal Protocols, Interviews, Post-Task Questionnaires, and Usability Study

The data collected from the verbal protocols, interviews, and questionnaires have been analyzed to find illustrative quotes to support the relationships observed among the variables and to develop future research questions. The purpose for using verbal protocols, post-task questionnaires, and interviews is to gain greater insight into what users focus upon when exploring a collection, how users determine and formulate their search strategies (Bates, 1979), and how user IR behavior impacts the eDiscovery IR process. Users are encouraged to “think out loud” during the IR task so that their thinking process and physical action can be recorded and subsequently transcribed (Vandenbosch and Huff, 1997; Todd and Benbaset, 1987).

Semi-structured interviews have been developed with questions adapted from Vandenbosch and Huff (1997). The interviews are designed to gain insight into the differences between IR behaviors that favor Recall (effectiveness) versus Precision (efficiency). Questions were asked post-task to determine how users’ IR behaviors had been impacted by the system. The post-task questions asked during the interviews are listed in Appendix - H.

Post-Task paper and pen questionnaires were used during the exploration study to gain insight into what specific techniques participants used to complete the task, how the participants characterized their chosen technique as a form of exploration solution, and the participants’ attitudes toward solving exploration problems for development of future research questions.
A paper and pen usability questionnaire was implemented at the end of both studies for the purpose of feedback regarding the utility of the system, how it can be improved, and the likelihood that the participant would use the artifact in a real world IR instance. Data collected from the usability study has been analyzed for triangulation with the PIIT and DISPO scales to determine whether a significant relationship exists between an individual’s innovativeness and attitude toward use of the systems developed and evaluated within this dissertation.
Chapter 4

Information Retrieval (IR)

An information system does not inform on the subject matter being queried; it informs about the existence of documents containing the subject matter being queried (Vanrijsbergen, 1979). Information Retrieval is concerned with determining the presence or absence of documents meeting certain criteria (relevance) within a corpus and a method for extracting those documents from the larger collection. Retrieval can be manual or automated. In this dissertation we are concerned with automated processes for IR. An assumption at work here is that criteria are expressed as terms and have been selected by the user to be processed by the automated tool because they have certain meanings that correspond to relevancy (Giger, 1988).

The limitation of an IR automated tool lies in the flat nature of search terms. The tool can only count up the occurrences and distributions of the terms in the query; it does not know the meaning behind the words or what may be the greater concept of interest. Users assume dependencies between concepts and expected document structures, whereas tools use process of statistical and probabilistic measures of terms in a document to determine a match to a query – relevance (Giger, 1988). If the measure meets a predetermined threshold level, the document is collected as relevant. However, the meaning behind the terms is lost and can result in the correct documents being missed or the wrong documents being retrieved. We see this occurring with instances of polysemy
and synonymy (Giger, 1988; Deerwester and Dumais, 1990). An example of this would be a user searching for documents related to an “oil spill” and not retrieving documents describing a “petroleum incident,” or a user searching for incidents of a person suffering a “fall” and the search engine returns documents describing an autumn day in September (Hyman and Fridy, 2010).

One way to address the disconnect between a set of search terms and a user’s meaning is to model the strategy behind the search tactic (Bates, 1979). One tactic is file structure. This tactic describes the means a user applies to search the “structure” of the desired source or file (Bates, 1979). Another tactic is identified as term; it describes the “selection and revision of specific terms within the search” (Bates, 1979). A user develops a strategy for retrieval based on their concepts. These concepts are translated into the terms for the query (Giger, 1988). The IR system is based on relevancy which is the matching of the document to the user query (Salton, 1989; Oussalah et al., 2008).

4.1 Relevance

“Relevance is a subjective notion” (Vanrijsbergen, 1979). In eDiscovery relevance as a concept refers to the match of a document to the particular subject matter requested. The ad hoc nature of this definition results in an ambiguity in the identification of a relevant document from a collection. As a result of this circumstance, relevance judgments are made by users who are experienced or specially informed on the subject matter (Interviews: Bill Hamilton and Aaron Laliberte, November 2011). Quite often judgments fall to a panel of experts in the domain for determining relevance of a document (Vanrijsbergen, 1979; TREC Proceedings 2009, 2010, 2011). This is quite
different from general document IR – where a single user seeks documents on a subject matter, and the user determines whether the retrieval satisfies his/her information need.

4.2 **Electronically Stored Information (ESI)**


Some practitioners in recent years have estimated that “more than 30% of corporate communications are in electronic form, and as much as 97% of information is created electronically,” (Russell T. Burke, Esq. and Robert D. Rowe, Esq., Nexsen Pruet Adams Kleemeier, LLC, 2004; M. Arkfeld, *Electronic Information in Litigation* §1.01, Law Partner Publishing, 2003). A particular difficulty in the domain of ESI retrieval is that prior to the Federal Rules requiring preservation of ESI in anticipated or actual litigation, little or no strategic planning had been done to organize this data – largely unstructured — for the purpose of a third party’s later review and extraction (Hyman and Fridy, 2010). A significant volume of ESI exists as unstructured information such as emails, texts and scanned documents, taking on various forms such as PDF, DOC, EXCEL, PPT, PST, JPEG and others. This motivates the problem of volume (large
collections of ESI) and uncertainty (unstructured and unorganized information) in document retrieval.

4.2.1 Approaches for Reducing the Search Space

As early as the 1950s, methods for reducing the search space were being explored (Singhal, 2001). Harris published a study about the distributional structure of documents (Harris, 1954). H.P Luhn proposed indexing words in documents and using their occurrences as criteria for relevance (Luhn, 1957). Probabilistic methods rank relevancy of documents based on estimations (Maron and Kuhns, 1960; Singhal, 2001).

Probabilistic based IR began to increase in use in the 1980s to allow for conditional candidacy of documents. Cluster methods have been proposed as a means of sorting large data sets into smaller, more cohesive subsets to address the polysemy problem – multiple meanings of words. Rooney et al., 2006, propose a contextual method to cluster documents semantically related to each other. The clusters are organized as minimum spanning trees with the similarities between adjacent documents compared. (Kostoff and Block, 2004) propose an approach called contextual dependency using “trivial word” filtering. The underlying assumption is that if trivial words can be removed, analysis can concentrate on salient terms of the target document. This is described in greater detail in section 4.5.3, discussing techniques for modeling.

4.3 The Problem of Uncertainty and Unstructured Documents

Uncertainty refers to the semi-structured or unstructured nature of the data. (Bates, 1986) proposes a design model identifying the three (3) principals: Uncertainty, Variety and Complexity, associated with the search of unstructured documents.
Uncertainty is defined as the indeterminate and probabilistic subject index. Variety refers to the document index. Complexity refers to the search process. One of the features of her proposed model included an emphasis on semantics.

Latent semantic indexing and Latent semantic analytics (LSA) have been proposed methods to deal with the poor fit of limited operators such as synonyms, polymorphisms and dictionaries, as applied to IR of semi-structured and unstructured documents (Huang and Kuo, 2003). The seminal work in LSA can be traced back to Deerwester, Dumais, Furnas, and Landauer (1990). They conducted some early experiments with semantic structure in text search. The foundation behind latent semantic is the creation of a text by document matrix. The matrix is decomposed into a set of orthogonal factors that represent a linear function. Weights are assigned to the terms, and documents that meet a determined threshold are retrieved (Deerwester et al., 1990).

Asymmetry is another description for uncertainty and refers to the wide variety of sources of data production (Huang and Kuo 2003). Some reports are event driven while others contain reflections or impressions never intended to be viewed by others. These conditions make retrieval difficult given that we are seeking items never intended to be found. Some research supports the contention that unstructured search is necessarily iterative and exploratory (Uren et al., 2007). Uren et al. reviewed four types of search methods: keyword, form-based, view-based, and NLP. Their conclusion was multimodal approaches should be used given that methods and modes are still emerging to handle uncertainty associated with unstructured search.
4.4 Performance Measures

Performance evaluation for IR systems can be measured in terms of effectiveness and efficiency (Grossman and Frieder, 1998). Effectiveness measures the accuracy of the system. Efficiency measures cost and time such as computer resources and human effort.

IR performance results in terms of effectiveness and efficiency are typically measured in terms of recall and precision. The methods assessed in this dissertation are evaluated for performance based on recall and precision. A functional value of recall and precision is known as F. Recall, precision and the F-measure have all been in existence for many years and have been validated in the literature (Van Rijsberen, 1979; Moffat and Zobel, 2008; Text Retrieval (TREC) Conference 2010 Proceedings).

Recall is calculated as the proportion of relevant documents retrieved out of the total amount of relevant documents available. It represents the number of documents correctly predicted as relevant. For example, if 15 documents are predicted to be relevant out of 25 total relevant documents available, then recall is 15/25 or .60. Precision is calculated as the proportion of relevant documents retrieved out of the total amount retrieved. For example, if there are 20 documents retrieved and 15 of them are relevant, then precision is 15/20 or .75. Predicting a document is non-relevant, when in fact it is, results in a false negative. Predicting a document is relevant, when in fact it is not, results in a false positive. Figure 4 is a confusion matrix depicting recall and precision as they relate to predicted and actual relevancy.
Document cut-off levels can be applied to measure system efficiency and effectiveness. This is done by calculating recall and precision at specific document cut-off values. For instance, we may ask the question: How many documents need to be generated by the system to achieve a certain level of recall or precision? This can be measured by calculating the number of relevant documents the system generates for the first 5, 10, 20, 30, 100, 200, etc. of documents (Baeza-Yates and Ribeiro-Neto, 1999). This approach is also helpful when comparing systems. For instance, let us suppose two systems each produce 70% recall, but one system generates 2000 documents to achieve that level of performance, and the other system does so by generating 1000 documents. Cut-off values will be an important part of evaluating system performance in the studies reported in this dissertation.

4.5 Approaches to IR

Document representation has been identified as a key component in IR (Vanrijsbergen, 1979). There is a need to represent the content of a document in terms of its meaning. Clustering techniques attempt to focus on concepts rather than terms alone. The assumption here is that documents grouped together tend to share a similar concept (Runkler and Bezdek, 1999, 2003) based on the description of the cluster’s characteristics. This assumption has been supported in the research through findings that less frequent terms tend to correlate higher with relevance than more frequent terms. This has been described as less frequent terms carrying the most meaning (Grossman and Frieder, 1998) and more frequent terms revealing noise (Interviews: Bill Hamilton and Michael Berman, November 2011).
Another method that has been proposed to achieve concept based criteria is the use of fuzzy logic to convey meaning beyond search terms alone (Ousallah et al., 2008). Ousallah et al. proposed the use of content characteristics. Their approach applies rules for locations of term occurrences as well as statistical occurrences. For example, a document may be assessed differently if a search term occurs in the title, keyword list, section title, or body of the document. This approach is different than most current methods that limit their assessment to over-all frequency and distribution of terms by the use of indexing and weighting.

Limitations associated with text-based queries have been identified in situations where the search is highly user and context dependent (Grossman and Cormack, 2011; Chi-Ren et al., 2007). Methods have been proposed to bridge the gap of text-based. (Brisboa et al., 2009) proposed using an index structure based on ontology and text references to solve queries in geographical IR systems. (Chi-Ren et al., 2007) used content-based modeling to support a geospatial IR system. The use of ontology based methods has also been proposed in medical IR (Trembley et al., 2009; Jarman, 2011).

Guo, Thompson and Bailin proposed using knowledge-enhanced, KE-LSA (Guo et al., 2003). Their research was in the medical domain. Their experiment made use of “original term-by-document matrix, augmented with additional concept-based vectors constructed from the semantic structures” (Guo et al., at page 226). They applied these vectors during query-matching. The results supported that their method was an improvement over basic LSA, in their case LSI (indexing).
An alternative method to KE-LSA has been proposed by (Rishel et al., 2007). In their article, they propose combining part-of-speech (POS) tagging along with an NLP software called “Infomap” to create an enhancement to LS indexing. POS tagging was developed by Eric Brill in 1991, and proposed in his dissertation in 1993. The concept behind POS is that a tag is assigned to each word and changed using a set of predefined rules. The significance of using POS as proposed in the above article is its attempt to combine the features of LSA, with an NLP based technique.

Probabilistic models have been proposed for query expansion. These models are based upon the Probability Ranking Principal (Robertson, 1977). Using this method, a document is ranked by the probability of its relevancy (Crestiani, 1998). Examples include: Binary Independence, Darmstadt Indexing, Probabilistic Inference, Staged Logistic Regression, and Uncertainty Inference.

4.5.1 Using Classifiers and Learning in IR

Classifiers are a common learning technique to probabilistically model a corpus. They address the issue of uncertainty in the search domain resulting from polysemy and synonymy. A classifier sets a coefficient weight to the occurrence of a search term in a document and a probability associated for the document within a data set (Blei, et. al, 2003). A classifier as core component of a search algorithm is a collection of rules representing the structure of a prototypical document a user seeks to retrieve (Perols et al., 2009). A classifier is constructed as a vector of a document matrix (Blei et al., 2003). The vector represents a set of rules “learned” by the classifier based on input criteria.
There have been numerous approaches to developing classifiers. Over time classifiers have evolved to include additional dimensions. In 1990, Deerwester, et al. published an article on the method Latent Semantic Analytics (LSA), also called Latent Semantic Indexing (LSI). The significance of the LSA/LSI method lies in its approach for dealing with the issues of synonyms and polymorphisms. Since its publication in 1990, there have been numerous variants to the LSA/LSI approach (Blei, et. al, 2003). Huang and Kuo, 2003 pursued indexing as a method for semi-structured and unstructured documents. Joachims, 1998 and Crestiani et al., 2000 have taken an approach using Support Vector Machine learning (SVM). (Hofmann, 1999) proposed Probabilistic Latent Semantic (pLSI) as a variant to LSA/LSI. (Blei et al., 2003) proposed Latent Dirichlet Allocation (LDA) as an improvement over LSA/LSI and pLSI methods.

Using a classifier developed from a vector matrix representing an exemplar of the target document addresses the issues of polysemy and synonymy as explained above. The approach has been applied using indexing (Sebastiani, 2002) and clustering (Rooney et al., 2006). The vector classifier approach has been proven effective in a variety of situations and data sets (Rooney et al., 2006 using RCV1; Kostoff and Block, 2005 using Swanson; Tang et al., 2009 using Caltech-101).

4.5.2 Limitations with classifiers

Noise can impact the learned structure of a classifier resulting in over-fitting the data (Nanopoulos et al., 2007). (Wang and Oard, 2008) observed that query expansion is subject to noise. Robustness techniques have been proposed to address this problem (Nanopoulos et al., 2007). It has been suggested that item-based paradigms can address the noise sensitivity of user-based paradigms and improve performance against noise
Outliers can also impact the performance of a classifier (Nanopoulos et al., 2007). A robustness technique using robust minimum distance and minimax estimate of location has been offered by (Shevlyakova et al., 2008) to address this issue. Support Vector Machines (SVM) has been used as a method for learning. One example is the IR task being defined as a “constrained quadratic programming problem” (Shalev-Shwartz et al., 2011). (Trafalis and Gilbert, 2006) explored SVM to address uncertainty and (Joachims et al., 2009) described how SVM can address the complexity of many dimensions of document criteria.

The classifier, by statistical method, learns probabilities associated with the occurrence of terms contained within the data sub-set. It is in fact this property that led to Blei et al.’s criticism of LSA/LSI being no better than using maximum likelihood (Blei et al., 2003 at 994). A limitation with the use of classifiers is the tendency to be domain dependent (Perols et al., 2010), meaning performance of a classifier is related to the domain upon which it is implemented.

An attempt at addressing domain dependency was proposed by (Majid Mojirsheibani, 1999) using a combiner method. He suggested that combining different classifiers would develop more effective classifier rules. This approach was improved upon by using a market-based fusion approach to combiners (Perols et al., 2009). (Li et al., 2011), proposes a classifier combination to “reduce dependency and improve overall performance.” Their work is similar to Perols et al., in that they “train single domain classifiers separately with domain specific data.” Whereas Perols et al., use market-based fusion for combining classifiers, Li et al., combine individually trained classifiers to
produce a final result. The study done by Li et al., is significant insofar as they seek to address “domain-dependent” and “domain-restricted” environments. Their work focused on sentiment classification as a special case of text categorization. They looked at attitude instead of specific facts.

4.5.3 Techniques for Modeling

Research suggests that using fewer terms will produce better recall and precision (Grossman and Frieder, 1998). The use of stop words can reduce noise by removing non-functional words from the search. Stop words are terms within a document that are irrelevant to the context and structure of the documents. For example, the preceding sentence would be written like this if stop words are removed: “Stop words terms document irrelevant context structure.” Very little meaning is lost through stop word removal. However, computational complexity is reduced significantly. Some earlier researchers found that up to 40% of document text may be comprised of stop words (Francis and Kucera, 1982; Grossman and Frieder, 1998).

Stemming is a method to normalize key terms down to their roots. For instance the word ‘run’ is a stem for running, runs, and runner, but not for ran. The goal of stemming is to reduce the complexity of the word to a root that will allow the engine to pick up various forms of the word. However, in the run example, if tense is important, then run cannot be used to find ran, but a “wildcard” such as r*n may serve the purpose.
Sometimes a user is searching for a document that contains terms within proximity of each other. For instance, we could be searching for articles on New York City. In this case we want to limit documents that contain the terms ‘New’ and ‘York’ within proximity of each other. Another example would be Vice President. Using a window span technique can also have a stemming effect. For instance “Vice President” may be alternatively spelled as: ‘Vice-President,’ ‘Vice Pres,’ or ‘VP.’
Chapter 5

Legal Domain and Legal Informatics

5.1 Legal IR

Traditional IR in the legal field was structured and predictable. Search and retrieval episodes in litigation cases occurred in two categories: (1) Structured search of legal cases with tools such as Westlaw and Lexis, and (2) Manual search of paper documents in client physical files prior to the proliferation of electronically stored information (Oard et al., 2010). A typical search in the legal domain would be to find a case or series of cases that answer a specifically defined question.

For example, if a present day attorney wants to know the law associated with a person arrested by the police as a result of finding illegal drugs in his car, a simple keyword search — automobile, drugs, seizure — will find the result with an expected 100% confidence level. The reason for this is that there have been decades of dedicated keyword coding for every single released legal case. The lawyer knows that the semantics of car, truck, motorcycle, or auto, will all be accounted for by the keyword automobile. The nature of the search is structured. The documents have been hand sorted and coded to produce the required result. Hence, the bag-of-words approach has served legal IR for many decades. An attorney wanting to know the law determining the fitness of a parent in a divorce custody battle need only to enter the keywords divorce, custody, fitness, and a
reliable result will be produced. A list of examples of traditional legal IR, based on keywords is provided in Table 2.

**Table 2: Examples of Traditional Legal IR Tasks**

<table>
<thead>
<tr>
<th>Issue</th>
<th>Keyword Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>A woman has been arrested for possession of illegal drugs, found in the glove box of the truck she was driving, as a result of a traffic stop for speeding</td>
<td>Automobile, Traffic Stop, Seizure, Probable Cause.</td>
</tr>
<tr>
<td>A man wants to fight his ex-wife for custody of the children because he believes she is an unfit mother. The divorce took place last year.</td>
<td>Divorce, Custody, Fitness, Change Circumstances.</td>
</tr>
<tr>
<td>A renter wants to know if he can hold back his rent because the landlord has not fixed the air conditioning.</td>
<td>Landlord, Tenant, Habitability, Rent, Set-off.</td>
</tr>
</tbody>
</table>

5.2 **eDiscovery Background**

Discovery is the process whereby opposing parties within litigation exchange documentary and testimonial evidence in preparation for trial. In this dissertation we are concerned with the documentary evidence alone. eDiscovery in essence involves two episodes of retrieval. The first retrieval is done by the responding party with two competing goals: (1) Meet the legal requirement of turning over documents that fit the requesting party’s criteria (recall), and (2) Avoid as best as possible turning over private or privileged documents—responder claims these documents are not relevant (precision). The second retrieval occurs when the requester receives the production from the responder. The requester performs a secondary search to reduce the number of documents for manual review to achieve the goal of reducing cost and time. The requestor also seeks to maximize recall — number of relevant documents extracted, and precision—fewest
non-relevant documents in the extraction set. Recall and Precision are represented in a confusion matrix—False negative/False positive table in Figure 6.

**Figure 6: Relevance Confusion Matrix**

An assumption in eDiscovery IR is that the user seeks to accomplish one or both of the following goals: (1) Retrieval of documents pursuant to a request from an adverse party, or (2) Tactical crafting of key words for an adverse party to use as search terms for the production of documents from the party’s data set. In either case, we assume the user has a significant idea about the structure and characteristics of the nature of the targeted documents. Another assumption about eDiscovery is that the user has a significant idea about the nature of the data set (Oard et al., 2010; Grossman and Cormack, 2011; Voorhees, 2000).

Phrases can be more meaningful than words alone (Singhal, 2001). Modeling phrases beyond terms alone can provide the richness that eDiscovery users need. An example of this is the legal concept of privileged communication. Such a communication is not subject to discovery; the law treats the document as if it is non-relevant. There are numerous ways in which a communication can be privileged. However, the use of a search term “privileged” will no doubt fail to capture the eDiscovery user’s need.
What makes eDiscovery an important problem to information systems research is the cost of liability and litigation to companies resulting from the duty to preserve and produce. The example mentioned in this dissertation is the legal case of *Zubalake v UBS Warburg*, 217 FRD 309 (S.D.N.Y. 2003). The background of the case is simple. Linda Zubulake was an employee of UBS. In 2002, she filed a federal law suit against her employer alleging sex discrimination. What made her case different was the part of her claim alleging that the Defendant, UBS, possessed data in electronic form that was discoverable as evidence in the suit. Part of the discovery litigation was routine in its complaints; the defendant claimed that they produced all documents requested and the plaintiff disputed that claim.

What began to turn the case into icon status was the fact that Ms. Zubulake produced more emails than UBS. She also managed to prove that UBS did not search their backup files. The Defendant claimed that a search of those files would be costly and that the Plaintiff should assume the burden of cost. The Court ordered UBS to produce all responsive emails from its active servers and optical disks at its own expense. It further ordered UBS to bear 75% of the costs to restore its back-up tapes and the entirety of the cost required to search and produce data from the restored tapes.

### 5.2.1 Litigation and “Go Fish.”

Jason Baron and Michael Berman comment in their book that most contemporary approaches to electronic searches “resemble a game of go fish.” (Baron and Berman, 2011). There are several legal cases that illustrate this point quite plainly.
The case of *In Re Fannie Mae Securities Litigation* is an example of how voluminous and out of control an electronic search can become; over 400 search terms were submitted by individual defendants. The search scope reached a span of 660,000 documents. The cost of the discovery requests reached $6 million dollars which represented over 9% of the government’s entire budget for the agency for that year (Baron and Berman, 2011).

Then there is the counter-example of *United States v. O’Keefe*. This case involved the criminal prosecution of a Department of State employee engaged in improper conduct in processing visas. The defendant in the case claimed that the government’s electronic search methods were inadequate. What makes this case interesting is the rationale contained in the court’s opinion. The court stated in its opinion that there needed to be some level of expertise offered to support the criticism of the other party’s search method. The defendant provided no such expert testimony to support a claim of inadequate search and therefore the claim was denied. This case, decided in 2008, suggests that there needs to be a credible rigor of expertise to support electronic discovery in litigation.

The case of *Victor Stanley I*, (suggesting there are more versions to follow), involved the inadvertent disclosure of privileged documents. The court acknowledged that “all keyword searches are not created equal,” and that searches are capable of being over-inclusive or under-inclusive (250 FRD at 256). The court found that privilege was waived due to ineffective search design and method. This legal opinion serves as an example of the motivating factor supporting the need for research in this domain. This
case does not go as far as *O'keefe* in requiring expert support, but it does suggest that electronic search needs to be supported by some degree of sophistication in design and method. A party may be required to defend its chosen techniques in court, and if they fail to do so, they may be sanctioned or waive certain rights.

The case of *Qualcomm v. Broadcom Corp.* serves as an example of sanctions being found against a party for failing to disclose 200,000 emails prior to trial (539 F. Supp. 2d 1214, S.D. Cal. 2007).

The case of *Kipperman v. Onex Corp.* serves as an example of a party not taking the potential volume problem in electronic search seriously; as a result, the court stated in its opinion that, “the defendants did not take advantage of these opportunities [offered previously during the case],” and they “must now lie in the bed that they have made.” (2008 WL 4372005, N.D. Ga.).

Another case involving sheer volume is *American International Specialty Lines*. This case involved approximately 19,000 boxes of documents. In this case the court found that based upon proportionality, review of the documents would be an undue burden (240 FRD 401, N.D. Ill. 2007).

These above mentioned cases suggest that simple keyword searching, as done by attorney and legal professionals in the past when faced with an electronic search, is no longer an adequate method for litigation in the newly emerging domain of eDiscovery. Clearly, more sophisticated methods and strategies need to be developed to meet the gaps exemplified in the above court cases. Another issue that is addressed in the domain is how to measure the performance of an electronic search in terms of recall and precision.
5.3 Performance Measures: F1, and Weighting Recall versus Precision

Some eDiscovery researchers and practitioners argue that the F-Measure treats all retrievals the same and therefore a distinction should be made when a recall or precision component performs significantly better or worse than another component (Grossman and Cormack 2011). This concern approximates the limitation with using an equal treatment of recall and precision in discovery litigation – the reader is reminded that the penalty for failure to recall is greater than the penalty for recalling too much, meaning non-relevant documents. F1 addresses the first of the two concerns. F1 is a function of: \[ \frac{2}{1/\text{Recall}} + \frac{1}{\text{Precision}} \]. The argument here is that F1 is a more accurate assessment of how a system is performing — whether it is favoring recall versus precision (Grossman and Cormack, 2011). The F-Measure operates as a “global F” whereas the F1 reports on the trade-off between recall and precision when using one particular method versus another.

In this dissertation the three methods are evaluated by recall and precision separately; no combined F-Measure is reported. The reason for this is to keep the research evaluated in this paper consistent with the goals of eDiscovery IR as indicated in the next section.

5.4 Ultimate Goal of eDiscovery IR

The goal of eDiscovery is to reduce the search space as greatly as possible so that the collection of documents reviewed by humans contains the fewest non-relevant documents without sacrificing too many relevant documents. The current method for accomplishing this goal is to extract sample sets from retrievals and inspect for false
positives, and extract samples from the non-retrieved corpus (remainder) and inspect for false negatives (Bill Hamilton, November 2011). The two prototype systems developed by this dissertation improve on this method by: (1) allowing the user to explore a small search space to identify better choices for search parameters, and (2) presenting the user with iterative small test sets enabling the user to refine their search and produce better IR results.
Chapter 6

Study One: Exploration and IR

6.1 Abstract

Law firms are under constant pressure to reduce the billing charged to clients. Due to the fact that all information retrievals in eDiscovery must be reviewed by humans, there is a significant interest in reducing the number of non-relevant documents to be reviewed. Reducing this number will result in significant cost savings to the firm and their clients. This study addresses the objective of cost savings in human review by developing and evaluating a process for user exploration to reduce the search space. The goal is to improve the retrieval result developed from user exploration of a small representative sample of the full corpus. The desired effect is to reduce the search space such that fewer documents are needed for human review.

This study uses four behavioral measures to predict an individual’s performance and three exploration measures to predict IR results. The data collected are recorded using an IT artifact developed for this study and evaluated for its utility based on the common performance measures of recall and precision.

The research conducted by this study provides insight into the relationship between recall and precision previously validated in the literature but never explained; how behavioral preferences impact IR performance; and how exploration variables measuring documents and time can be used to predict productivity in IR results.
6.2 Introduction

eDiscovery is an instance of document retrieval of a bounded collection/corpus. In such an instance the collection is domain specific, the search is ad hoc, and the typical user is highly educated in the domain – either through direct prior experience or emersion during an investigative process. (Bill Hamilton, November 2011). Study One: Exploration focuses on these two conditions. These conditions of eDiscovery IR go largely unexploited by users.

As previously stated, eDiscovery IR is highly context dependent, meaning that relevant search terms are often linked to subject matter knowledge or a controlled vocabulary of the corpus items. For example, if a user enters the word “the” as a search term, he/she will no doubt return 100% recall of documents but there will be almost zero (0) precision, resulting in the return of the entire corpus — not very useful.

A more useful result would be if a user could surgically identify terms with the goal of reducing the search space that produces a retrieval set containing a high percentage of relevant documents and a low percentage of non-relevant documents. In this study we investigate whether exploration is an effective method to learn context in order to achieve this goal.

This study addresses the gap between brute force, trial and error techniques, and test collection reviews presently employed by eDiscovery practitioners, by designing an artifact to support user exploration of a bounded corpus. The study seeks to explain how users determine their IR strategies and how exploration can be used most effectively to improve IR performance.
In 1985, Blair and Maron conducted a series of experiments designed to address the problem of finding relevant documents from a collection. The collection contained 40,000 objects. They found that on average, recall was in the 20% range – quite unacceptable. Twenty-five years later, instances of low recall are still common in large corpus IR (TREC Proceedings 2010). One explanation for this phenomenon lies in the limitation of a user to predict the matching of terms to relevant documents – recall terms producing relevant documents, and also not producing non-relevant documents (Blair and Maron, 1985). This study addresses this limitation by developing an artifact to support exploration as a method to better predict matching of terms to relevant documents. One of the objectives of this study is to provide insight into how users develop their search strategies (Bates, 1979). We begin our research in this domain by researching the phenomenon of exploration.

Exploration is a natural and intuitive method to use when probing a large collection of documents in an attempt to reduce the scope of the search space. We see common and frequent examples everyday when a person searches the web for information on a subject matter or topic. In such instances the user chooses terms, and sometimes operators, as an initial predictive approximation for the information being requested, and then adjusts the query criteria as results appear. This approach makes conventional sense when conducting a search of scale free collections with no preconceived definition of document(s) satisfying the information need, and where the information need is the presence or the absence of a document containing the information requested rather than a specific answer to a specific factual question.
6.2.1 Exploration

The concept of exploration has been associated with learning (Berlyne, 1963; March, 1991); familiarization (Barnett, 1963), and information search (Debowski et al., 2001). In fact work done by Berlyne in the 1960s classifies exploration as a “fundamental human activity” (Demangeot and Broderick, 2010).

Exploration is seen as a behavior motivated by curiosity. Exploration that is goal directed is classified as extrinsic (Berlyne, 1960). Extrinsic exploration typically has a specific task purpose, whereas intrinsic exploration is motivated by learning (Berlyne, 1960; Demangeot and Broderick, 2010). (Kaplan and Kaplan, 1982) argue that exploration arises from our need to make sense of our environment. (March, 1991) writes about exploration and exploitation. He views exploration and exploitation as competing tensions in organizational learning.

(Berlyne, 1963) suggests that specific exploration is a means of satisfying curiosity. The goals of exploration as a means for making sense of our environment and satisfying curiosity are represented in the problem domain of information retrieval and eDiscovery. (Debowski et al., 2001) view exploratory search as a “screening process,” and state that exploration identifies items “to become the focus of attention.” They suggest that exploration leads to learning through the examining and scrutinizing of items.

The first artifact developed for this study is designed to support the user in exploring a corpus of items and facilitating examining and scrutinizing, so that the user may obtain contextual knowledge about the collection.
6.2.2 Prior Exploration/Search Research

The study of exploration and search behavior and the mental models users create to execute them is not entirely new. In the early 2000s much research focused upon strategies that users formulate for web searches. (Holschler and Strube, 2000) examined the types of knowledge and strategies involved in web-based information seeking. They found that users with higher levels of knowledge were more flexible in their approach and were better able to tackle search problems than those who were less knowledgeable. They characterize the information space as “diverse and often poorly organized content.” This contrasts with the bounded space of eDiscovery which is typically organized around the subject matter in question. Their finding that experts can outperform less experienced users will be extended by this study by evaluating whether knowledge acquired by exploration can improve a user’s ability to tackle the search problem of eDiscovery.

(Muramatsu and Pratt, 2001) evaluated a system designed to provide users with “light weight feedback” about their queries. They found that transparency is “helpful and valuable.” Their conclusion was that interfaces “allowing direct control of query transformation may be helpful to users.” The exploration study and learning study in this dissertation extend their work by designing two separate artifact tools to provide just such control. We evaluate the efficacy of both artifacts in chapter 6 and chapter 8 of this dissertation.

Other research has focused upon browsing behavior and categorizing search behaviors into types. Bates (1989) coined the phrase “berry-picking” to refer to individuals’ search strategy being in constant evolution. A study done by Catledge and Pitkow at Georgia Institute of Technology captured client-side user events to study
browsing and search behavior (Catledge and Pitkow, 1995). Their study evaluated frequency and depth and found support for three different types of searcher characterizations based on Cove and Walsh’s original work in 1988: Serendipitous browser, General purpose browser, and Searcher (Cove and Walsh, 1988).

Broder (2002) proposed a taxonomy of web search to include transactional—a web mediated activity, navigational—seeking a specific site, and informational—a page containing a particular need. This dissertation studies the user’s informational need, and also seeks to explain his/her navigational behavior that may affect the IR result produced.

(Muylle et al., 1999) undertook a study to better understand web search behaviors and motivations of consumers and business people. The study found three constructs describing search behavior: (1) exploratory – title scanning, (2) window – document scanning, and (3) evolved – document scrutinizing. The research conducted in this dissertation adapts these constructs to measure scanning, skimming and scrutinizing behavior in the users conducting exploration of the corpus.

(Navarro-Prieto et al., 1999) studied how people search for information and focused on the “cognitive strategies” followed by the user. Not surprisingly, they found three prevailing strategies: (1) Top-down—broad based followed by narrowing down, (2) Bottom-up—specific terms for specific fact finding, and (3) Mixed—employing both strategies in parallel. Also not surprising, they found that experience mattered. The users who were more experienced developed more complex rules for their searches and followed a top-down approach.
The research and findings mentioned above are consistent with the other research found in this domain: (1) that experience matters, and (2) that experience affects the complexity of search strategy and choice of search terms.

This dissertation seeks to extend the research done on search by evaluating how users can improve their knowledge through exploration, and leverage that knowledge through an automated tool to improve IR results.

6.3 Motivation

eDiscovery extraction is modeled differently than an open ended IR search such as “scanning the web for general information on a topic,” or a prior art search for say, a patent. What makes eDiscovery unique is the manner in which the user frames the universe to be searched; the corpus/collection is bounded — it is defined in a way that those who understand the context of the documents to be sought tend to produce better IR results. The reason for this is that the user in eDiscovery IR is a high compensated individual, an educated professional or team, highly focused on the topic of interest. The topic arises out of a specific series of transactions or related events that are defined by time, population, location, and other ad hoc circumstances making the IR corpus bounded in a particular way that the IR result (relevance) is highly dependent on content and context. This unique set of IR circumstances leads to our main research question in this chapter of whether a user can acquire knowledge of context and content through exploration of the bounded corpus. The assumption is that more knowledge on a topic will produce better IR results than less knowledge on a topic.

Why is this question important? The elements that make IR of a bounded corpus unique also make it an important phenomenon to study. Consider the fact that bounded
collections represent the recorded actions of parties to everyday transactions. As our society becomes more and more dependent on digital storage of recorded transactions the ability to effectively extract relevant documents from large collections of similar items will continue to be a value proposition in terms of time and cost. Whether it is a consumer and merchant, a dispute between commercial actors, or in the case of Ms Zubalake, an employee against her employer, bounded collections are becoming increasingly more frequent in information retrieval.

Most bounded searches in eDiscovery follow a standard pattern of: interviewing personnel expert in the domain, ascertaining the standards of storage and organization, developing potential search terms, and producing test reports indicating the frequency of terms within documents (occurrences). “Hit reports” are reviewed and terms are refined based on the frequencies observed. Quite often, this is done prior to reviewing any document extractions. Patterns are determined and more test runs are commenced; only then are documents reviewed from the extracted sets. As one interviewee describes, IR users engage in a “brute force” trial and error approach to arrive at a focal point of key terms to use for extraction (Aaron Laliberte, November 2011).

Search terms are proxies for relevance descriptors of a document (Salton and Buckley, 1988). Weighting of terms enhances the effectiveness of document description by using statistical occurrences and term frequencies. This technique is fundamental to indexing methods (Luhn, 1957; Spark-Jones, 1971). A major limitation associated with term frequency is the difficulty of distinguishing between the frequency of occurrence in the relevant documents and the frequency of occurrence in the entire collection. There is an assumption here that search terms may be known a priori or may surface as a result of
patterns discovered during exploration of the collection. One hypothesis in this study is that *exploration of the collection will provide the eDiscovery user with the ability to describe the document he/she is seeking, and therefore select better search terms*. A second hypothesis here is that *exploration of the corpus will provide a greater understanding about the nature of the documents (relevant and not), and lead to better decisions for selection of search terms, resulting in improved recall and precision*.

### 6.3.1 Background

eDiscovery is a domain where the nature of the IR is highly user dependent and highly context oriented (Oard et al., 2010; Baron, 2005; Grossman and Cormack, 2011). This nature exploits the weaknesses of term based search alone. Term based search is well suited when a query is narrow in focus and particularity. eDiscovery users have found that term search alone is inadequate when context is important, resulting in (over-inclusion) precision loss, or (under-inclusion) recall loss, (Paul and Baron 2007; *Sendona Conference*, 2007; Oard et al., 2010). This study evaluates whether a user can learn the context of a collection and make better decisions about the selection of search terms to improve the IR result. The prototype developed for this study allows the user to select the “level of importance” of a search term for the method of weighting.

### 6.4 Methods

The method used in this study is a controlled experiment. The purpose of the experiment is to measure the affect upon IR performance of user exploration of a small sample of a large corpus. Performance is measured by the dependent variables *Recall* and *Precision* as previously defined. There are two sets of explanatory variables used. The first set is comprised of behavioral scales known to be associated with preferences that
are predictive in the use of technology and innovativeness. The second set is comprised of operational measures to represent the constructs of scanning and scrutinizing behaviors associated with exploration of digital collections.

The document sample consists of 300 randomly selected documents from the overall collection of 680,000 objects. The task, treatment and data collection are conducted via the prototype developed for this study. The prototype application built for this experiment is housed on a server and accessed by the participants using a URL link from their self provided laptop computers. The computer screens from the application are displayed in Appendix - E.

Participants are assigned an eDiscovery task. Informed consent, task instruction and data collection instrument are displayed as computer screens – graphically depicted in Appendices – B, C and D.

All participants are given the same task. The task is to provide recall (search) terms and elimination terms (filters) in response to an eDiscovery request. The task has been adapted from the TREC Legal Track 2011 Conference Problem Set #401. A list of the exploration independent and dependent variables and covariates are displayed in Table 3.
<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Dependent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exploration/Artifact Variables:</strong></td>
<td><strong>Performance Measures:</strong></td>
</tr>
<tr>
<td>Number of documents viewed</td>
<td>Recall</td>
</tr>
<tr>
<td>Total Viewing Time</td>
<td>Precision</td>
</tr>
<tr>
<td>Viewing time per document</td>
<td></td>
</tr>
<tr>
<td>Order of documents viewed*</td>
<td></td>
</tr>
<tr>
<td><strong>Covariates:</strong></td>
<td></td>
</tr>
<tr>
<td>Age, Gender, Education Level</td>
<td></td>
</tr>
<tr>
<td>Legal, Discovery, eDiscovery experience</td>
<td></td>
</tr>
<tr>
<td>Experience with ENRON collection</td>
<td></td>
</tr>
<tr>
<td>Experience with financial terms, concepts and transactions.</td>
<td></td>
</tr>
</tbody>
</table>

*Collected for Future Study

The independent exploration variables tracked in this study are: Total Number of Documents Viewed, Total Amount of Viewing Time, and Time Spent per Document.

Linear regression analysis is used to measure the following relationships: (1) Correlations of independent variables with dependent variable Recall, (2) Correlations of independent variables with dependent variable Precision, (3) Interactive effect of independent variables upon the dependent variables, and (4) Significance of covariates.

Demographic information has been collected to track co-variables. Specific *a priori* co-variables are legal experience, discovery experience, eDiscovery experience, familiarity with the Enron data set, experience in financial transactions, financial concepts and financial terminology. Age, gender, and educational level have also been collected. The above variables have been identified by our panel of experts and surfaced during our pilots for this study. Not all co-variables appeared during our data collection. The co-variables that have been identified as most relevant in this study are: Litigation experience scaled as 0 – 3 to represent no experience, less than 1 year, 1 to 2 years, and greater than 2 years; Knowledge of subject matter (Financial Terminology) scaled as 0 –
3 to represent no knowledge, some familiarity with terms, amateur investing, and professional training or experience.

6.4.1 Dataset

The dataset in this study is a corpus of electronic documents known as the Enron Collection, Version 2. The full corpus of this version contains approximately 650,000 to 680,000 email objects depending on the counting of attachments. This data set has been previously validated in the literature (TREC Legal Track Proceedings 2010, 2011).

The subset of data we use for the exploration artifact is a collection of 10,000 randomly selected documents from the full corpus. 1,000 of the documents have been selected from the validated set marked relevant, and 9,000 documents have been selected from the validated set marked not relevant. This allows us to make certain assumptions. The first assumption is that a random extraction from the subset should yield a recall of .10. Any level of recall above this number indicates an improvement over chance – a result better than no human input at all. The second assumption is that the set of documents retrieved by the user selections can be measured for precision based on a common base line for relevance from the validated relevant documents.

6.4.1.1 Participants

The participants in this study are 60 third year law students from three different United States, East Coast law schools. Law students have been chosen for this study because they are familiar with fundamental principles of legal procedure and discovery process and techniques. Participants have been given an economic incentive to motivate them to try their best. The participants were told that the best performer in each group would be awarded a cash prize.
6.4.2 Process

An IT artifact developed for this dissertation is a job run procedure used to pre-process the data from the Enron collection. The objects in the collection which consist of emails and attachments need to be prepared in such a way that the text can be read by the engines of the system. Considerable time went into this phase. Some problems encountered with files included: password protected files, power point files that did not translate well with the chosen OCR tool, emails with URL links no longer valid, emails with attachments only and no text in the body, and emails with files pasted into the body instead of native text. The job run process is depicted in Figure 7.

The participants interact with the user-interface screens made available through an URL link to the server from their personal laptops. They are instructed by random assignment to select Group 1, Group 2, Group 3, from a list of radio buttons on the computer screen (depicted in the Appendix - E). The radio button chosen corresponds to amount of exploration time allowed. Group 1 receives up to 15 minutes, Group 2 receives up to 30 minutes, and Group 3 receives 45 minutes to explore the 300 document sample.

The time allotments are maximums, meaning each participant may terminate their individual session at any point during the study. For example, a participant in Group 3 may choose to terminate his/her exploration after 10 minutes. The actual application has 4 group buttons. The additional button is used for testing of the system. This allows us to segregate our system tests from the participant data.

The participants may conclude their exploration at any time by selecting the next button on the screen. The exploration behaviors are logged by the system as sessions, and
tracked as the independent variables (IVs) Total Time, Time per Document, Number of Documents.

All three groups receive the same task. The purpose of using three groups is to spread out the time line. We found during the pilots that if participants are not given an anchor time they all cluster too close together and create a narrow variance in time measure. Therefore, the study uses artificial groups to spread out the time line thereby avoiding a tight cluster and increase the explanatory power of the variables.

The participants supply their recall terms and elimination terms through the interface screen. The user selections are logged per user and submitted to the job queue for processing as depicted in Figure 7.
6.5 Initial Pilot Studies

The first pilot conducted in this study consisted of 7 lay personnel. The purpose of the pilot was to receive feedback regarding presentation, clarity and ease of use. The current version of the system reflects feedback received from the pilot.

The second pilot was conducted with 10, second year law students. Five (5) students were placed in a control group which was given no time to explore, and 5 were given up to 60 minutes to explore the sample collection. The average time exploring the collection clustered around 43 minutes, with a single high of 60 minutes and a single low of 23 minutes. The average number of documents reviewed was 70, with a single high of 90 and a single low of 15. The average time per document was 45 seconds, with a single high of 2 minutes and a single low of 10 seconds.

Data was inconclusive on the issue of significant difference in recall or precision between the two groups. The reason for this probably has to do with the small number in the groups in order to detect a difference from zero. The small number of participants also makes it difficult to draw conclusions about how: total time viewed, number of documents viewed and per document time, may be predictive of recall and precision due to the concentrated clustering of the total time explored.

The most useful information provided from the second pilot was in the form of user feedback. The participants provided feedback consistent with the previous pilot. This increased our confidence in the design of presentation and environment for the full experiment. The main limitation in the second pilot indicted above is the small sample group making it difficult to draw conclusions about relationships of the variables.
However, the pilot has confirmed the design, ease of use and quality issues of the system being tested.

6.6 Design of Full Study

Sixty (60) third year law students are the subjects of this study. They have been randomly assigned to three groups to spread out time performance. The individuals within each of the groups are allotted maximum time allowances to complete their exploration. The participants may terminate their exploration at any time. For example, an individual who is assigned to Group 4 is given up to 45 minutes to explore the corpus however the participant may choose to terminate the exploration at the 10 minute mark; there is no forced time range or minimum amount for the participants, just a maximum allowance depending on the Group assigned.

The participants are administered the behavioral questionnaires at the beginning of the study so that their responses are not affected by the task. The behavioral questionnaires are designed to collect data on the four scales measuring user IR behavioral attitudes: Tolerance for Ambiguity (TOA), Locus of Control (LOC), Disposition Toward Innovation (DISPO), and Personal Innovation Toward Information Technology (PIIT). Three subjects from each group have been selected for verbal protocols and are encouraged to “think out loud.” Post-task interviews are conducted with three (3) additional subjects from each group. The purpose for choosing subjects from each group is to select users from different levels of time exploration.

All subjects are administered the post-task questionnaire and usability study at the end, followed by a hearty thank you for their time and good-bye. The total time for
participation ranged from 45 minutes to 110 minutes. The participants’ sessions have been recorded by a server hosting the artifact/application.

Independent variables (IVs) representing Total Time Exploring, Total Documents Viewed, and Time Spent Per Document have been assigned to track user interaction with the artifact. A model that depicts the artifact IVs and their relationship to dependent variables (DVs) Recall and Precision is illustrated in Figure 8. The model for exploration artifact assumes an input, an output, and a process in the middle. The three IVs represent input, the two DVs represent output, and the exploration construct is in the middle representing the human cognitive process.

![Figure 8: Model of Artifact IVs](image)

Independent variables representing tolerance for ambiguity (TOA), locus of control (LOC), dispositional innovativeness (DISPO), and personal innovativeness toward information technology (PIIT) have been assigned to track user behavioral factors associated with information retrieval technology and innovation. This study focuses on the portion of the Information Retrieval Behavior Model from Vandenbosch and Huff in Figure 3.2, representing the impact of behavioral measures upon exploration and their relationship to the dependent variables (DVs) Recall and Precision. The adapted model is depicted below in Figure 9.
Covariates assigned in this study control for *litigation experience*, *familiarity with the corpus* and *knowledge of financial terms* (subject matter experience) associated with the task. Levels have been assigned to correspond to years of litigation experience, depth of financial knowledge, and exposure to the corpus. A table listing the co-variates, assigned levels, and descriptions is displayed in the Appendix - J.

### 6.6.1 Exploration Artifact Variables

Information seeking can be divided into broad exploration and precise specificity (Heinstrom, 2005). Broad exploration is a possible indicator of a wide overview strategy and knowledge building, whereas precise information seeking may be an indicator of a focused, pinpointed search (Heinstrom, 2005). In the case of precision search, the user has a specific frame of reference from which to investigate and probe a collection.

Research has found that a “common approach” to large collection search is for the user to begin with “an already known term” (Lehman et al., 2010). The use of the known term typically leads to an item that informs the user with additional terms to improve the search for the next iteration. When more than one item is returned the user has the option of reviewing each item one at a time. But when a large volume of items is contained in the retrieval set, the user must apply some method to select items for further inspection from among the set. (Lehman et al., 2010) developed a visualization method for user exploration of large document collections. The visualization approach was employed by...
them to study user information seeking in Wikipedia. The results of their study found that, “visual navigation can be easily used and understood” (Lehman et al., 2010).

Browsing as an information seeking process has been established as a method when the information need is ill-defined (Kuhlthau, 1991; McKay et al., 2004). Browsing has been described as a fundamental information seeking function (Bates, 1979, 1989; Kuhlthau, 1991; McKay et al., 2004). Exploration is an underlying construct representing the human search behavior (Holschler and Strube, 2000; Muylle et al., 1999); it is operationalized in electronic search as browsing. This study operationalizes exploration by use of an artifact built as an interactive tool to support user exploration of a corpus by exploitation of selected items in order to learn context and content.

When a user finds multiple documents they will tend to switch back and forth between items; this activity describes the iterative approach to information seeking (McKay et al., 2004). (Meuess et al., 2005) developed an XML retrieval system combining structure with text references. (McKay et al., 2004) developed three approaches to browsing using the Greenstone digital library database. Ignat et al., 2006 developed an automated tool designed to support exploration of large document collections by use of clustering; it is implemented using a standard web browser. (Chowdhury et al., 2011) focused on uncertainty as an underlying construct in Human Information Behavior (HIB).

The above referenced research has focused on investigating and describing users’ information seeking behavior through exploration and browsing activities. The research in this dissertation is focused upon benchmarking user productivity in the search process.
In this study we have selected variables to measure user productivity in the exploration information seeking process. We have chosen *Total Time Explored* and *Time Spent Per Document* to measure the effort expended by exploration and to account for the exploration/exploitation trade-off (Holschler and Strube, 2000; Muylle et al., 1999; Hills, 2010; March, 1991). We have chosen *Total Number of Documents* explored to account for the iterative nature of information seeking (Bates, 1989; McKay, 2004).

The general proposition for this study is that an exploration method will outperform both random extraction and verbatim/non-function word extraction. The hypotheses representing this proposition are as follows:

H1a Random: Exploration outperforms random extraction measured in units of recall.

H1b Random: Exploration outperforms random extraction measured in units of precision.

H2a Verbatim: Exploration outperforms verbatim extraction measured in units of recall.

H2b Verbatim: Exploration outperforms verbatim extraction measured in units of precision.

The hypotheses for the exploration variables are as follows:

H1a: *Recall* is directly and positively correlated with *Total time exploring a corpus*.

H1b: *Precision* is directly and positively correlated with *Total time exploring a corpus*.
H2a: *Recall* is directly and positively correlated with the *Number of documents viewed in a corpus*.

H2b: *Precision* is directly and positively correlated with the *Number of documents viewed in a corpus*.

H3a: *Recall* is directly and positively correlated with *Time spent per document*.

H3b: *Precision* is directly and positively correlated with *Time spent per document*.

We did not have any prior theory about whether some of the variables might interact to produce effects upon recall and precision. Therefore, we used a null and alternative hypothesis for each:

H₀: *Total time exploring a corpus* affects *Recall* and *Precision* independent of *Number of documents viewed* and *Time per document*.

Hₐ: *Total time exploring a corpus* affects *Recall* and *Precision* depending on *Number of documents viewed* or *Time per document*.

H₀: *Number of documents viewed* affects *Recall* and *Precision* independent of *Total time exploring a corpus* and *Time per document*.

Hₐ: *Number of documents viewed* affects *Recall* and *Precision* depending on *Total time exploring a corpus* or *Time per document*. 

73
H₀: Time per document affects Recall and Precision independent of Total time exploring a corpus and Number of documents viewed.

H₃: Time per document affects Recall and Precision depending on Total time exploring a corpus or Number of documents viewed.

6.6.2 Behavior Scales

The experiment in this chapter seeks to explain factors impacting IR results and uses an innovative tool to do so. Four behavioral scales have been chosen to measure preferences known to be associated with information retrieval and innovation. The goal is to determine which scales are significant in ability to predict IR performance of individuals, measured by the variables Recall and Precision. Personality traits have been associated with information seeking patterns and differences in search approaches and strategies (Heinstrom, 2005). The four behavioral scales used in this study are listed in Table 4. They are further explained in the next sections.
**Table 4:** List of Behavior Scales

<table>
<thead>
<tr>
<th>Variable</th>
<th>Name</th>
<th>Description</th>
<th>Number of Items</th>
<th>Cronbach’s Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOA</td>
<td>Tolerance for Ambiguity</td>
<td>The degree to which an individual is willing to accept ambiguity is “related to an individual’s desire to create uncertainty and tend toward scanning behavior because they are not fearful of the ambiguity that often results.” (Vandenbosch and Huff, 1997)</td>
<td>8</td>
<td>.80</td>
</tr>
<tr>
<td>LOC</td>
<td>Locus of Control</td>
<td>A person who has a higher LOC believes he/she has greater control over what happens to them rather than external factors. This individual is more likely to explore broadly due to greater confidence to produce results.</td>
<td>5</td>
<td>.85</td>
</tr>
<tr>
<td>DISPO</td>
<td>Dispositional Innovativeness</td>
<td>The measure of an individual’s likeliness to try a new product, or think tangentially when solving a problem.</td>
<td>8</td>
<td>.85</td>
</tr>
<tr>
<td>PIIT</td>
<td>Personal Innovativeness in the Domain of Information Technology</td>
<td>The degree to which an individual has a preference for technology use.</td>
<td>4</td>
<td>.97</td>
</tr>
</tbody>
</table>

**Tolerance for Ambiguity**

Tolerance for Ambiguity (TOA) has been found to be associated with uncertainty in tasks intended to replace ambiguity with order (Vandenbosch and Huff, 1997; Rydell and Rosen, 1966; McCasky, 1976). The hypotheses are illustrated in Figure 10, below and in written form as follows;

H4a: *TOA is positively related to Recall.*

H4b: *TOA is negatively related to Precision.*
Figure 10: TOA effect upon Recall and Precision

Given that we know from previous studies that recall and precision are inversely related (Oard et al., 2010; Grossman and Cormack, 2011), we believe in this study that individuals seeking less ambiguity will prefer greater precision, whereas individuals willing to accept more ambiguity will prefer greater recall. The person more comfortable with ambiguity is more likely to seek broader exploration because he/she is not concerned with the additional non-relevant documents that may result. This is especially applicable to eDiscovery where lawyers often go on “fishing expeditions” as mentioned by Oard et al., 2010. The pre-task questionnaire designed to measure this construct has been adapted from the Rydell-Rosen Scale (1966). The original form contained 20 items which proved too unwieldy for our subjects. A confirmatory factor analysis was used to reduce the number of items. The final form contains 8 items and produced a Cronbach alpha of .80.

Locus of Control (LOC) is a measure of the degree to which individuals believe they control their own fate (Levenson, 1974). The LOC inventory developed by Levenson measures three factors: (1) Internal, the extent to which the person believes he or she is in control; (2) External, the extent to which a person believes his or her fate is controlled by others; (3) Chance, the extent to which the person believes their fate is determined by chance events.
Prior MIS research has found that individuals who believe they control their own fate are more likely to engage in scanning techniques for their IR (Vandenbosch and Huff, 1997; Levenson, 1974). Prior analysis of the Levenson three factor scale has shown it to be more reliable than similar scales measuring only two factors (Presson et al., 1997). For these reasons the Levenson three factor scale has been adapted for use in this study. The original form had 24 items. A confirmatory factor analysis was used to reduce the number of items to 5 with a Cronbach alpha of .85.

The proposition in this study is that scanning should be expected to be associated with broader search exploration and therefore, would favor recall over precision. The rationale is that individuals who believe they are in control of their performance results, rather than chance or others being in control, are more likely to conduct broader searches, leading to greater relevant documents returned. Broader searches are associated with return of greater non-relevant documents. We therefore believe that individuals with a higher preference on the LOC scale will explore with greater confidence, search broader, and produce higher recall, but lower precision. The hypotheses are illustrated in Figure 11, and presented in written form as follows;

H5a: LOC is positively related to Recall.

H5b: LOC is negatively related to Precision.
Innovativeness can be described in several ways. It has been used in consumer research to predict an individual’s predisposition to purchase new products (Roehrich, 2004; Steenkamp and Gielens, 2003). It has been shown to predict an individual’s willingness to try a new technology (Agarwal and Prasad, 1998). It has been used to explain an individual’s tendency to engage in thinking exercises such as puzzle solving and pondering (Pearson, 1970). When describing “cognitive innovation” Pearson describes the concept as “thinking for its own sake” (Venkatraman and Price, 1990, citing Pearson, 1970).

In this dissertation we are interested in how an individual’s exploration attitudes and techniques can be explained through known and validated measures. In this case we have settled on two scales for measuring innovativeness. The first scale is designed to measure a user’s disposition toward innovativeness. The second scale is designed to measure a user’s personal innovativeness.

“Dispositional Innovativeness” (DISPO) has been shown to be significant in predicting consumers who are more likely to try a new product (Steenkamp and Gielens, 2003). In this dissertation participants are being asked to use a new method for eDiscovery IR. One of the hypotheses of this dissertation is that participants measuring higher on the scale of dispositional innovativeness will produce a higher IR result. The
administered questionnaire contains eight (8) items measured on a 1 to 5 scored scale, ranging from completely disagree = 1 to completely agree = 5. Cronbach alpha for this inventory is .85.

The proposition here is that individuals with a higher level of dispositional innovativeness are more likely to embrace a new system resulting in greater IR results. It is likely that such individuals are broader thinking and are willing to randomly jump around in their exploration due to their preference for the new and novel. These types of individuals are more tangential in their thinking and approach problem solving from unconventional points of view (Kirton, 1976; Vandenbosch and Huff, 1997). The hypotheses derived from the proposition are depicted in Figure 12 and in written form as follows:

H6a: DISPO is positively related to Recall.

H6b: DISPO is negatively related to Precision.

![Figure 12: DISPO effect upon Recall and Precision](image-url)
These hypotheses are measured using two different methods. The first method analyzes whether DISPO is significant and if the relationship is in fact positively correlated with Recall and negatively correlated with Precision. The second method utilizes a post-task usability study. This study asks the users to rate the system on how well it helped them perform the task and how likely they are to use this system in a real life eDiscovery scenario. The results from the usability study are discussed in Chapter 9.

“Personal innovativeness in the domain of information technology” (PIIT) is associated with early adopters and individuals who are more comfortable with uncertainty (Agarwal and Prasad, 1998 citing Rodgers, 1995). Given that the eDiscovery user specifically operates in the domain of uncertainty, a measure of a user’s PIIT may be helpful in predicting the same user’s exploration preferences and resulting IR performance. The questionnaire contains 4 items and produced a Cronbach alpha of .97.

Agarwal and Prasad argue that individuals with higher PIIT levels are more likely to have positive attitudes toward an innovative technology. These attitudes translate to our experiment in terms of higher values in Precision. We believe that individuals with a preference toward technology will be more surgical in their exploratory behavior and produce higher precision. Given the documented inverse relationship between recall and precision, we believe the higher performance in Precision will result in a lower performance in Recall. The hypotheses are depicted in Figure 13 and in written form below:

H7a: PIIT is negatively related to Recall.

H7b: PIIT is positively related to Precision.
6.7 Data Analysis

SAS 9.2 was the statistical package chosen to support the analysis in this study. Collected data has been analyzed in several steps. The method of analysis in this case is a multiple linear regression. We are analyzing whether the independent (explanatory) variables are significant and whether interactive effects are present. We are also concerned with controlling for the listed covariates. A global F-test was used to evaluate the overall model and partial F-tests were used for testing interactive effects.

The behavioral scales have been analyzed using Cronbach’s alpha. Two of the behavioral scales were extremely long (TOA and LOC); the original version of TOA had 20 items and the original version of LOC had 24 items. In order to reduce these scales to a manageable number of items for participants, a factor analysis was conducted for each scale. The scales were reduced to 8 items and 5 items respectively. Confirmatory Factor Analysis was used with varimax rotation. Cronbach alphas were calculated for the scales and are listed in Table - 4.

The first step was to transfer the pen and paper questionnaires to a spreadsheet for input into SAS. These questionnaires covered the four scales of TOA, LOC, DISPO, and PIIT. These behavioral scales were then analyzed to determine significance in a main

![Figure 13: PIIT effect upon Recall and Precision](image)
effects and full model. The models reflect the underlying theories represented by the hypotheses being tested. The initial theory of the behavioral scales is that individuals’ IR performance can be predicted from their scores on the behavioral scales. The theory is represented by the hypotheses in the previous section and reduced to equations forming the behavioral models indicated below.

**Main Effects Model:**  
\[ DV_{\text{Recall}}, DV_{\text{Precision}} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + e \]

**Full Model:**  
\[ DV_{\text{Recall}}, DV_{\text{Precision}} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + \]
\[ B_5X_1X_2 + B_6X_1X_3 + B_7X_1X_4 + \]
\[ B_8X_2X_3 + B_9X_2X_4 + B_{10}X_3X_4 + \]
\[ CV_1 + CV_2 + CV_3 + e \]

Where:

- \( X_1 = \text{TOA} \),
- \( X_2 = \text{LOC} \),
- \( X_3 = \text{DISPO} \),
- \( X_4 = \text{PIIT} \),
- \( CV_1 = \text{Litigation Experience} \),
- \( CV_2 = \text{Enron Set Familiarity} \),
- \( CV_3 = \text{Subject Matter Familiarity (Financial Knowledge)} \).

After the behavioral models were analyzed, we then conducted analysis upon the exploration models comprised of the independent variables *Total Time Explored* (TTE), *Time per Document* (PER), and *Number Of Documents Explored* (NUM). The initial theory is that individuals’ IR performance can be predicted based on their exploration behavior measured by the independent variables. The theory is represented by the
hypotheses in the previous section and reduced to the equations for the exploration models indicated below:

**Main Effects Model:** \(DV_{\text{Recall}}, DV_{\text{Precision}} = B_0 + B_0 + B_1X_1 + B_2X_2 + B_3X_3 + e\)

**Full Model:**
\[
DV_{\text{Recall}}, DV_{\text{Precision}} = B_0 + B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \\
B_4X_1X_2 + B_5X_1X_3 + B_6X_2X_3 + \\
B_7CV_1 + B_8CV_2 + B_9CV_3 + e
\]

Where:

- \(X_1 = \text{Total Time Explored},\)
- \(X_2 = \text{Time per Document},\)
- \(X_3 = \text{Number of Documents Explored},\)
- \(CV_1 = \text{Litigation Experience},\)
- \(CV_2 = \text{Enron Set Familiarity},\)
- \(CV_3 = \text{Subject Matter Familiarity (Financial Knowledge)}.\)

In addition to analyzing the behavioral models and the exploration models we also investigated relationships between the behavioral variables and the exploration variables. To test for this we set up an equation with the behavioral scales as independent variables and the exploration variables as dependent.

We had no prior theory about whether the behavioral variables would affect the exploration variables. Therefore, we used null and alternative hypotheses to test whether a significant relationship exists along with the linear equations for this model indicated below. In this case the null hypotheses represent the results that there is no significant difference from zero between the independent variables and the dependent variables. The
alternative hypotheses represent the results that at least one of the independent variables is significantly different from zero.

**Main Effects Model**: \( DV_1, DV_2, DV_3 = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + e \)

- **H_0**: \( B_1 = B_2 = B_3 = B_4 = 0 \)
- **Ha**: At least one Beta \( \neq 0 \)

**Full Model**: \( DV_1, DV_2, DV_3 = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_1X_2 + B_6X_1X_3 + B_7X_1X_4 + B_8X_2X_3 + B_9X_2X_4 + B_{10}X_3X_4 + e \)

- **H_0**: \( B_1 = B_2 = B_3 = B_4 = B_5 = B_6 = B_7 = B_8 = B_9 = B_{10} = 0 \)
- **Ha**: At least one Beta \( \neq 0 \)

**Where**:

- \( DV_1 = \) Total Time Explored,
- \( DV_2 = \) Time per Document,
- \( DV_3 = \) Number of Documents Explored,
- \( X_1 = \) TOA,
- \( X_2 = \) LOC,
- \( X_3 = \) DISPO,
- \( X_4 = \) PIIT.

The covariates were coded on a continuous scale based on levels of experience as described previously. The plan was to analyze the covariates for significance. When we collected the data it turned out that none of the 60 students had experience with the data set nor did any of them have financial experience, so those two co-variables dropped out.
of the equation. Also, only three students answered that they had litigation experience and none had eDiscovery experience. The analysis of CV for litigation experience indicated that there was no significant difference in the result. This could be due to the fact that there were only three out of 60 students that answered in the affirmative on this question.

6.8 Results

The average number of documents reviewed was 43, with an average of 27.5 minutes total time and 58 seconds – just under one minute – spent per document. The average number of documents produced was 503 with an average recall of .50 and an average precision of .61.

IR performance results have been compared across three alternative methods for IR extraction: (1) The exploration approach measured by average Recall and Precision based upon incremental units of the Total Time Explored variable, (2) A random extraction of the 503 documents, representing the average number of documents produced by the participants’ selections, (3) An extraction of documents based on the eDiscovery task verbatim after removing non-function words.

The graph of Recall against time appears in Figure 14. Recall performance for participants exploring the corpus for less than 15 minutes produced results in the .3 to .5 range, but outliers at the 14 and 15 minute data points make conclusions about a general trend within this time frame difficult to draw.

Participants exploring the corpus in the 23 - 30 minute time frame produced results in the .4 to .5 range and follow a mostly flat trend line. There is a gap up from .2
at the 15 minute mark to .5 at the 23 minute mark; but with no data points between 16 and 22 minutes, it is difficult to draw a conclusion about why this trend occurs.

There is also a significant trend upward between 30 minutes and 42 minutes, with a gap up between the 40 and 42 minute marks and no data point at 41 minutes. The recall performance results are fairly flat after an initial jump in performance between 30 and 42 minutes. This warrants further study to determine if there is a diminishing return beyond 42 minutes. A future experiment is planned to include a time frame up to 60 minutes to investigate this relationship.

Exploration outperformed random extraction at every data point. However, the verbatim non-function method was a better choice for users who spent 15 minutes or less and was competitive in several data points in the 15 – 30 minute range. Exploration outperformed verbatim non-function at all data points over 30 minutes; the effect also seems to flatten after an initial jump in this time range. This too will need to be studied further to determine if the effect remains flat, meaning that no further exploration yields improvement or if a time range beyond 45 minutes may continue to improve Recall. The study will be performed over a larger sample of users to make the results more generalizeable.
The graph of precision against time appears in Figure 15. Precision performance follows a different trend than recall.

Participants exploring the corpus on the shorter end of the timeline, 23 minutes or less, produced results in a range of .6 to .8, with an outlier at the 14 minute mark. Participants exploring the corpus on the higher end of the timeline, greater than 40 minutes, produced results that were consistently above .6.

Participants exploring the corpus in the middle of the timeline, greater than 23 minutes but less than 40 minutes produced the worst results; they were in the .4 to .6 range. This is a strange result given that we expected participants to improve generally as time increased, and in this time range performance dropped. We have no explanation for why this may be so. We plan to further study this effect to investigate whether in fact a middle time frame exists that should be avoided by eDiscovery users.

The main results indicate that exploration can be an effective method for producing better precision than random extraction or verbatim /non-function word, and
perhaps most effective (results above .60) when a user spends over 40 minutes or less than 13. However, it is difficult to justify such a conclusion with a small sample and with data point gaps within the ranges analyzed. Therefore, a future study has been planned to investigate this effect with a larger sample of users.

![Precision/Time](image)

**Figure 15**: Precision over Time

A random extraction of documents was produced equal to the average number of documents produced by the participants in the study to determine if the exploration method would outperform chance. Given that the subset of documents contained 1,000 relevant out of 10,000 documents, a random extraction should produce 10% relevant documents. The average number of documents extracted based on participant performance was 503. If a random selection of 500 documents from the corpus was performed, the expectation would be approximately 50 documents out of 500 should be relevant. This would yield a precision of .10 and a recall of .05. Given that there were 60 participants, we performed 60 random extractions and averaged the results.
When we performed the random extractions our average result was actually in line with expected chance performance. The average number of relevant documents extracted was 51, with a high of 68 and a low of 38.

Given that the worst performance using the exploration method was .20 for recall and .43 for precision, exploration outperformed random extraction.

In situations when the eDiscovery user has no a priori guidance for what search structure or terms that might produce relevant documents, sometimes the specific words from the request itself can be used as a good starting point to probe for initial trial and error results. The theory is that the terms in the request may in fact be significant indicators of relevant context.

When we performed this type of extraction we produced 2120 documents from the 10,000 item corpus, with 455 relevant. This extraction represents a recall of 455/1000 (.455) and a precision of 455/2120 (.215) – a pretty good starting point if the user has no prior knowledge. The exploration approach produced an average recall of .50 and an average precision of .61, outperforming Method 3 in both measures.

The results show that hypotheses H1aRandom and H1bRandom are both supported. The exploration participants outperformed random extraction at all data points in the study.

Hypothesis H2aVerbatim is partially supported. The exploration participants outperformed verbatim extraction in all data points greater than 30 minutes. Exploration did not outperform verbatim extraction in data points under 15 minutes and produced mixed results in the 15 to 30 minute range.
Hypothesis H2bVerbatim is supported. The exploration participants outperformed verbatim extraction for precision for all data points in the study.

There are several possibilities that may explain the results reported above. The most obvious explanation could be that a certain minimum amount of time must be given to a user to produce any improvement over verbatim extraction. This study suggests that, unless a user is prepared to spend more than 23 minutes on exploration, don’t bother, simply use an automated approach such as verbatim.

Another explanation may be that, after a certain amount of time is spent exploring, there is a significant leap in knowledge acquired about the corpus. This study suggests that the number may be as little as 40 minutes or more to achieve this leap.

The flatter results produced in the 23 to 40 minute range are a mystery. There are several speculative explanations we could suggest. One possibility is; there may be a range of time spent in exploration that produces no increased effect, meaning if a user is going to spend less than 42 minutes, then the user might as well reduce that time to 23, because the additional 19 minutes will not produce any more productivity in recall.

6.8.1 Statistical Analysis of Three Models

Global F-tests were performed on the three models: Exploration, Behavioral, and Behavioral-Exploration. Each model was analyzed separately for Recall and for Precision using null and alternative hypotheses. A table summarizing the results for the behavioral and exploration hypotheses is contained in the discussion section.

A global F-test has been performed for Recall and for Precision. A summary of results appear in Table 5 and Table 6 on the next page.

The null and alternative hypotheses are as follows:
Recall

H₀: B₁ = B₂ = B₃ = 0
Ha: At least one Beta ≠ 0

Precision

H₀: B₁ = B₂ = B₃ = 0
Ha: At least one Beta ≠ 0

Where:

B₁ = Slope for Total time explored,
B₂ = Slope for Number of documents viewed,
B₃ = Slope for Time per document.

Table 5: SAS 9.2 Printout for Recall Variables

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>1.08166</td>
<td>0.36055</td>
<td>105.21</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>56</td>
<td>0.19191</td>
<td>0.00343</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>59</td>
<td>1.27357</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE = 0.05854
R-Square = 0.8493
Dependent Mean = 0.50733
Adj R-Sq = 0.8412
Coeff Var = 11.53878

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>TOTALTIM</td>
</tr>
<tr>
<td>PERDOCTI</td>
</tr>
<tr>
<td>TOTALDOC</td>
</tr>
</tbody>
</table>
Table 6: SAS 9.2 Printout for Precision Variables

The REG Procedure
Model: MODEL1
Dependent Variable: PRECISION

Number of Observations Read 60
Number of Observations Used 60

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>0.26712</td>
<td>0.08904</td>
<td>12.68</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>56</td>
<td>0.39312</td>
<td>0.00702</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>59</td>
<td>0.66024</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.08379 R-Square 0.4046
Dependent Mean 0.61600 Adj R-Sq 0.3727
Coef Var 13.60160

Parameter Estimates

| Variable    | Label               | DF | Estimate | Std Error | t Value | Pr > |t| |
|-------------|---------------------|----|----------|-----------|---------|------|---|
| Intercept   | Intercept           | 1  | 0.66597  | 0.04392   | 15.16   | <.0001|
| TOTALTIM    | TOTAL TIME          | 1  | -0.00975 | 0.00203   | -4.81   | <.0001|
| PERDOCTI    | PER DOC TIME        | 1  | -0.00128 | 0.05229   | -0.02   | 0.9803|
| TOTALDOC    | TOTAL DOCS VIEWED   | 1  | 0.00582  | 0.00102   | 4.94    | <.0001|

The global F-test for the Recall exploration model and the Precision exploration model are both significant at alpha .01. However, Recall and Precision differ in which IVs are significant predictors. Total Time Explored is significant at alpha .01 for recall and precision. However, Number of Documents Viewed is significant at alpha .01 for Precision, but not for Recall; and Time per document was not supported for either Recall or Precision.
**Table 7: Summary of Exploration Model Results**

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Alpha</th>
<th>Dependent Variable Effected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Time Exploring*</td>
<td>.01</td>
<td>Recall &amp; Precision</td>
</tr>
<tr>
<td>Number Documents*</td>
<td>.01</td>
<td>Precision</td>
</tr>
<tr>
<td>Time per Document</td>
<td>Not Significant</td>
<td></td>
</tr>
</tbody>
</table>

*-* Interactive Effect upon Precision

The exploration independent variables have been analyzed for interactive effects. Total Time Explored and Total Number of Documents Viewed were found to have an interactive effect upon Precision and the relationship was significant at alpha .01. This suggests that the impact upon Precision by the total time spent in exploration depends on the total number of documents viewed, and the impact upon Precision by the total number of documents viewed depends on the total time explored. No other interactive effect was found to be supported. SAS 9.2 printout results from interactive tests appear on the next two pages in Table 8 and Table 9.
Table 8: Results from SAS 9.2 printout for interactive effect upon Recall

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>1.09217</td>
<td>0.18203</td>
<td>53.18</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>53</td>
<td>0.18140</td>
<td>0.00342</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>59</td>
<td>1.27357</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.05850 R-Square 0.8576
Dependent Mean 0.50733 Adj R-Sq 0.8414
Coeff Var 11.53156

| Variable          | Label           | DF | Estimate   | Standard Error | t Value | Pr > |t| |
|-------------------|-----------------|----|------------|----------------|---------|------|---|
| Intercept         | Intercept       | 1  | 0.33478    | 0.12329        | 2.72    | 0.0089 |
| TOTALTIM          | TOTAL TIME      | 1  | 0.00912    | 0.00404        | 2.26    | 0.0280 |
| PERDOCTI          | PER DOC TIME    | 1  | -0.004297  | 0.15645        | -0.27   | 0.7847 |
| TOTALDOC          | TOTAL DOCS VIEWED | 1 | -0.00474  | 0.00328        | -1.45   | 0.1540 |
| TTPD              |                 | 1  | -0.00175   | 0.00644        | -0.27   | 0.7864 |
| TTD                |                 | 1  | 0.00009655 | 0.00005987     | 1.61    | 0.1128 |
| TDPD               |                 | 1  | 0.00160    | 0.00300        | 0.53    | 0.5965 |

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable RECALL

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>3</td>
<td>0.00350</td>
<td>1.02</td>
<td>0.3897</td>
</tr>
<tr>
<td>Denominator</td>
<td>53</td>
<td>0.00342</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 9: Results from SAS 9.2 printout for interactive effect upon Precision

The REG Procedure
Model: MODEL1
Dependent Variable: PRECISION

Number of Observations Read 60
Number of Observations Used 60

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>0.38057</td>
<td>0.06343</td>
<td>12.02</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>53</td>
<td>0.27967</td>
<td>0.00528</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>59</td>
<td>0.66024</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.07264
R-Square 0.5764
Dependent Mean 0.61600
Adj R-Sq 0.5285
Coeff Var 11.79238

Parameter Estimates

| Variable     | Label                  | DF | Estimate | Standard Error | t Value | Pr > |t| |
|--------------|------------------------|----|----------|----------------|---------|------|---|
| Intercept    |                        | 1  | 1.15779  | 0.15308        | 7.56    | <.0001|
| TOTALTIM     | TOTAL TIME             | 1  | -0.01673 | 0.00501        | -3.34   | 0.0015|
| PERDOCTI     | PER DOC TIME           | 1  | -0.33524 | 0.19426        | -1.73   | 0.0902|
| TOTALDOC     | TOTAL DOCS VIEWED      | 1  | -0.01290 | 0.00407        | -3.17   | 0.0026|
| TTPD         |                        | 1  | 0.00460  | 0.00799        | 0.58    | 0.5675|
| TTD          |                        | 1  | 0.00033831| 0.00007434    | 4.55    | <.0001|
| TDPD         |                        | 1  | 0.000467 | 0.00373        | 1.25    | 0.2157|

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable PRECISIO

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>3</td>
<td>0.03782</td>
<td>7.17</td>
<td>0.0004</td>
<td></td>
</tr>
<tr>
<td>Denominator</td>
<td>53</td>
<td>0.00528</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A global F-test has been performed upon the behavioral model. A summary of results appear in Table 10 on the next page.

The null and alternative hypotheses are:

**Recall**

H<sub>0</sub>: B<sub>1</sub> = B<sub>2</sub> = B<sub>3</sub> = B<sub>4</sub> = 0

Ha: At least one Beta ≠ 0

**Precision**

H<sub>0</sub>: B<sub>1</sub> = B<sub>2</sub> = B<sub>3</sub> = B<sub>4</sub> = 0

Ha: At least one Beta ≠ 0

Where:

X<sub>1</sub> = Tolerance for ambiguity (TOA),

X<sub>2</sub> = Locus of control (LOC),

X<sub>3</sub> = Disposition toward innovativeness (DISPO),

X<sub>4</sub> = Personal innovativeness in information technology (PIIT).

**Table 10: Summary of Behavioral Model Results**

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Alpha</th>
<th>Dependent Variable Effected</th>
<th>Beta Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOA</td>
<td>.01</td>
<td>Precision</td>
<td>.005</td>
</tr>
<tr>
<td>LOC</td>
<td>.01</td>
<td>Recall</td>
<td>-.013</td>
</tr>
<tr>
<td>DISPO</td>
<td>.05</td>
<td>Precision</td>
<td>.008</td>
</tr>
<tr>
<td>PIIT</td>
<td></td>
<td>Not Significant</td>
<td></td>
</tr>
</tbody>
</table>

The global F-test for the *Recall* behavioral model and the *Precision* behavioral model are both significant at alpha .01. However, just like the exploration model, the behavioral model differed in which variables were significant for *Recall* and which were significant for *Precision*. LOC was significant for *Recall* at alpha .01.
TOA was significant for Precision at alpha .01 and DISPO was significant for Precision at alpha .05. PIIT was not supported for Recall or Precision. The printouts for these results appear on the next pages in Table 11 and Table 12.

Table 11: SAS 9.2 Printout for Recall Variables

<table>
<thead>
<tr>
<th>The REG Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model: MODELL</td>
</tr>
<tr>
<td>Dependent Variable: RECALL</td>
</tr>
</tbody>
</table>

Number of Observations Read: 60
Number of Observations Used: 60

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>1.16472</td>
<td>0.29118</td>
<td>147.12</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>55</td>
<td>0.10885</td>
<td>0.00198</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>59</td>
<td>1.27357</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE: 0.04449
R-Square: 0.9145
Dependent Mean: 0.50733
Adj R-Sq: 0.9083
Coeff Var: 8.7697

Parameter Estimates

| Variable   | Label    | DF | Estimate      | Error       | t Value | Pr > |t| |
|------------|----------|----|---------------|-------------|---------|------|---|
| Intercept  | Intercept| 1  | 0.52230       | 0.04589     | 11.38   | <.0001|
| LOC        | LOC      | 1  | -0.01291      | 0.00194     | -6.64   | <.0001|
| TOA        | TOA      | 1  | 0.00043654    | 0.00149     | 0.29    | 0.7782|
| DISPO      | DISPO    | 1  | -0.00091858   | 0.00293     | -0.31   | 0.7547|
| PIITSUM    | PIIT SUM | 1  | 0.00320       | 0.00124     | 2.59    | 0.0124|
The behavioral variables have been analyzed for interactive effects. Interaction between the independent variables was not found to be supported in the individual p-values but was support at alpha .01 in the partial F test. This conflicting result suggests there may be multi-collinearity among two or more of the variables. To account for this possibility we have tested whether any of the IVs correlate.

The Pearson Coefficient results indicate that DISPO and TOA are highly correlated. We plan to study this effect in future experiments to determine if one of the variables should be removed from the equation for parsimony. We also found that LOC and PIIT are highly negatively correlated. PIIT was not found to be significant as a main effect; however, this relationship suggests that we need to be careful drawing conclusions about the IVs’ effects on Recall and Precision and we will need to further investigate this
effect in our future work with larger populations. The SAS 9.2 results reports for interactive effects and multi-collinearity have been reproduced on the next pages in Table 13, Table 13.1 and Table 14.

Table 13: SAS 9.2 Printout for Recall Variables

| Variable   | Label     | Parameter | Estimate | Error  | t Value | Pr > |t| |
|------------|-----------|-----------|----------|--------|---------|------|---|
| Intercept  | Intercept | 1         | 0.38595  | 0.13234| 2.92    | 0.0053|
| LOC        | LOC       | 1         | 0.01269  | 0.01158| 1.10    | 0.2782|
| TOA        | TOA       | 1         | 0.00620  | 0.00397| 1.56    | 0.1244|
| DISPO      | DISPO     | 1         | -0.00244 | 0.00566| -0.43   | 0.6687|
| PIITSUM    | PIIT SUM  | 1         | 0.00908  | 0.00787| 1.15    | 0.2941|
| PIITSUMTOA |           | 1         | -0.00039540 | 0.00022606 | -1.75 | 0.0864|
| PIITSUMDISPO |        | 1         | 0.00025541 | 0.00048162 | 0.53 | 0.5982 |
| LOCDISPO   |           | 1         | -0.0008662 | 0.00073713 | -0.12 | 0.9069|
| LOCTOA     |           | 1         | -0.00068173 | 0.00035911 | -1.90 | 0.0634|
| DISPOTOA   |           | 1         | -0.0002182 | 0.00011459 | -0.19 | 0.8498|

The REG Procedure
Model: MODEL1
Test 1 Results for Dependent Variable RECALL

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>5</td>
<td>0.01005</td>
<td>8.57</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Denominator</td>
<td>50</td>
<td>0.00117</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 13.1: SAS 9.2 Printout for Precision Variables

The REG Procedure
Model: MODEL1
Dependent Variable: PRECISION

Number of Observations Read 60
Number of Observations Used 60

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>9</td>
<td>0.61878</td>
<td>0.06875</td>
<td>82.91</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>50</td>
<td>0.04146</td>
<td>0.00082926</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>59</td>
<td>0.66024</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.02880
R-Square 0.9372
Dependent Mean 0.61600
Adj R-Sq 0.9259
Coeff Var 4.67482

Parameter Estimates

| Variable | Label | DF | Estimate | Error | t Value | Pr > |t|
|----------|-------|----|----------|-------|---------|------|
| Intercept| Intercept | 1 | 0.38182  | 0.11131| 3.43    | 0.0012|
| LOC      | LOC   | 1 | -0.01392 | 0.00974| -1.43   | 0.1589|
| TOA      | TOA   | 1 | -0.00226 | 0.00334| -0.68   | 0.5006|
| DISPO    | DISPO | 1 | 0.00607  | 0.00476| 1.28    | 0.2082|
| PIITSUMTOA| PIIT SUM | 1 | 0.00063453| 0.00662| 0.10    | 0.9240|
| PIITSUMDISPO| PIIT SUM | 1 | 0.00014418| 0.00019014| 0.76 | 0.4518|

The REG Procedure
Model: MODEL1
Test 1 Results for Dependent Variable PRECISION

<table>
<thead>
<tr>
<th>Mean Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>5</td>
<td>0.000367</td>
<td>4.42</td>
<td>0.0021</td>
</tr>
<tr>
<td>Denominator</td>
<td>50</td>
<td>0.00082926</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 14: SAS 9.2 Printout of Multi-Collinearity Analysis

Pearson Correlation Coefficients, N = 60
Prob > |r| under H0: Rho=0

<table>
<thead>
<tr>
<th></th>
<th>PIIT</th>
<th>LOC</th>
<th>TOA</th>
<th>DISPO</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIIT</td>
<td>1.00000</td>
<td>-0.89706</td>
<td>-0.00623</td>
<td>-0.12841</td>
</tr>
<tr>
<td>PIIT</td>
<td>-0.89706</td>
<td>1.00000</td>
<td>-0.22654</td>
<td>-0.07217</td>
</tr>
<tr>
<td>LOC</td>
<td>-0.00623</td>
<td>-0.22654</td>
<td>1.00000</td>
<td>0.91590</td>
</tr>
<tr>
<td>TOA</td>
<td>0.9623</td>
<td>0.0818</td>
<td>-0.07217</td>
<td>1.00000</td>
</tr>
<tr>
<td>DISPO</td>
<td>-0.12841</td>
<td>-0.91590</td>
<td>0.5837</td>
<td>1.00000</td>
</tr>
<tr>
<td>DISPO</td>
<td>0.3282</td>
<td>0.5837</td>
<td>0.91590</td>
<td>1.00000</td>
</tr>
</tbody>
</table>
As indicated previously, a third model has been developed to investigate whether a significant relationship exists between the behavioral and exploration variables. The behavioral variables have been set up as the independent and the exploration variables have been set up as the dependent. A summary of the results appear in Table 15 below. The null and alternative hypotheses are the same from the behavioral model given that the same betas are being investigated (TOA, LOC, DISPO, and PIIT).

**Table 15: Summary of Behavioral-Exploration Model Results**

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Alpha</th>
<th>Dependent Variable Effected</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOA</td>
<td>Not Significant</td>
<td></td>
</tr>
<tr>
<td>LOC</td>
<td>.01</td>
<td>Number of Documents</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>Time Per Document</td>
</tr>
<tr>
<td>DISPO</td>
<td>Not Significant</td>
<td></td>
</tr>
<tr>
<td>PIIT</td>
<td>Not Significant</td>
<td></td>
</tr>
</tbody>
</table>

In terms of the impact of the behavioral scales upon exploration behavior, LOC was significant for *Time per Document* at alpha .05, and for *Number of Documents* viewed at alpha .01. The other three scales were not significant for either *Recall* or *Precision*. SAS 9.2 printouts for these results are reproduced in Table 16 and Table 17 on the next page.
### Table 16: SAS 9.2 Printout for Number of Documents Viewed

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>36717</td>
<td>9179.18102</td>
<td>47.37</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>55</td>
<td>10657</td>
<td>193.75653</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>59</td>
<td>47373</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE: 13.91965  R-Square: 0.7751  Dependent Mean: 43.66667  Adj R-Sq: 0.7587  Coeff Var: 31.87705

#### Parameter Estimates

| Parameter | Label       | DF | Estimate | Standard Error | t Value | Pr > |t| |
|-----------|-------------|----|----------|----------------|---------|------|---|
| Intercept | Intercept   | 1  | 25.05723 | 14.35748       | 1.75    | 0.0865|
| LOC       | LOC         | 1  | -3.41746 | 0.60807        | -5.62   | <0.0001|
| TOA       | TOA         | 1  | 0.02524  | 0.46524        | 0.05    | 0.9569|
| DISPO     | DISPO       | 1  | 0.97680  | 0.91530        | 1.07    | 0.2905|
| PIITSUM   | PIIT SUM    | 1  | -0.35271 | 0.38716        | -0.91   | 0.3663|

### Table 17: SAS 9.2 Printout for Time per Document

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>1.47097</td>
<td>0.36774</td>
<td>6.64</td>
<td>0.0002</td>
</tr>
<tr>
<td>Error</td>
<td>55</td>
<td>3.04387</td>
<td>0.05534</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>59</td>
<td>4.51483</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE: 0.23525  R-Square: 0.3258  Dependent Mean: 0.56833  Adj R-Sq: 0.2768  Coeff Var: 41.39313

#### Parameter Estimates

| Parameter | Label       | DF | Estimate | Standard Error | t Value | Pr > |t| |
|-----------|-------------|----|----------|----------------|---------|------|---|
| Intercept | Intercept   | 1  | 0.59497  | 0.24265        | 2.45    | 0.0174|
| LOC       | LOC         | 1  | 0.03403  | 0.01028        | 3.31    | 0.0016|
| TOA       | TOA         | 1  | 0.00260  | 0.00786        | 0.33    | 0.7425|
| DISPO     | DISPO       | 1  | -0.00908 | 0.01547        | -0.59   | 0.5595|
| PIITSUM   | PIIT SUM    | 1  | 0.01221  | 0.00654        | 1.87    | 0.0674|
The behavioral variables have been analyzed for interactive effects upon *Time per Document* and *Number of Documents Viewed*. The SAS 9.2 results report for interactive effects has been reproduced in Table 18 and Table 19 on the next page.

An interactive effect has been found to exist between DISPO and TOA upon *Time per Document*. This effect is supported at alpha .05. No other interactive effects were supported.

**Table 18: SAS 9.2 Printout for Interaction (Time per Document)**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>9</td>
<td>2.09396</td>
<td>0.23266</td>
<td>4.81</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>50</td>
<td>2.42088</td>
<td>0.04842</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>59</td>
<td>4.51483</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE = 0.22004 R-Square = 0.4638
Dependent Mean = 0.56833 Adj R-Sq = 0.3673
Coeff Var = 38.71668

**Parameter Estimates**

| Variable   | Label       | DF | Estimate | Error | t Value | Pr > |t| |
|------------|-------------|----|----------|-------|---------|------|--|
| Intercept  | Intercept   | 1  | -1.70402 | 0.85049 | -2.00 | 0.0505 |
| LOC        | LOC         | 1  | 0.20550  | 0.07439 | 2.76   | 0.0080 |
| TOA        | TOA         | 1  | 0.06818  | 0.02550 | 2.67   | 0.0101 |
| DISPO      | DISPO       | 1  | 0.06719  | 0.03638 | 1.85   | 0.0707 |
| PIITSUM    | PIIT SUM    | 1  | 0.07675  | 0.05057 | 1.52   | 0.1354 |
| PIITSUMTOA |            | 1  | 0.00020629 | 0.00145 | 0.14   | 0.8877 |
| PIITSUMDISPO |        | 1  | -0.00242 | 0.00310 | -0.78  | 0.4381 |
| LOCDISPO   |            | 1  | -0.00641 | 0.00474 | -1.35  | 0.1824 |
| LOCTOA     |            | 1  | 0.00055853 | 0.00231 | 0.24   | 0.8098 |
| DISPOTOA   |            | 1  | -0.00211 | 0.00073644 | -2.87 | 0.0060 |

**Test 1 Results for Dependent Variable PERDOCTI**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Mean Square</th>
<th>Mean F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>5</td>
<td>0.12460</td>
<td>2.57</td>
<td>0.0379</td>
</tr>
<tr>
<td>Denominator</td>
<td>50</td>
<td>0.04842</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 19: SAS 9.2 Printout for Interaction (Total Documents Viewed)

The REG Procedure
Model: MODEL1
Dependent Variable: TOTALDOC TOTAL DOCS VIEWED

Number of Observations Read 60
Number of Observations Used 60

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>9</td>
<td>39295</td>
<td>4366.08555</td>
<td>27.02</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>50</td>
<td>8078.56338</td>
<td>161.57127</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>59</td>
<td>47373</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 12.71107 R-Square 0.8295
Dependent Mean 43.66667 Adj R-Sq 0.7988
Coeff Var 29.10932

Parameter Estimates

| Variable | Label  | Parameter | DF | Estimate  | Standard Error | t Value | Pr > |t| |
|----------|--------|-----------|----|-----------|----------------|---------|-----|---|
| Intercept| Intercept | 1 | 58.61794 | 49.13050 | 1.19 | 0.2385 |
| LOC      | LOC | 1 | -0.73322 | 4.29734 | -0.17 | 0.8652 |
| TOA      | TOA | 1 | -0.80526 | 1.47287 | -0.55 | 0.5870 |
| DISPO    | DISPO | 1 | -0.88471 | 2.10130 | -0.42 | 0.6755 |
| PIITSUM  | PIIT SU | 1 | -0.76073 | 2.92120 | -0.26 | 0.7956 |
| PIITSUMTOA | PIIT SU | 1 | 0.01383 | 0.08393 | 0.16 | 0.8697 |
| PIITSUMDISPO | PIIT SU | 1 | -0.00732 | 0.17880 | -0.04 | 0.9675 |
| LOCDISPO | 1 | -0.01698 | 0.27366 | -0.06 | 0.9508 |
| LOCCTOA | 1 | -0.06714 | 0.13332 | -0.50 | 0.6168 |
| DISPOTOA | 1 | 0.04745 | 0.04254 | 1.12 | 0.2700 |

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable TOTALDOC

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>5</td>
<td>515.60918</td>
<td>3.19</td>
<td>0.0141</td>
</tr>
<tr>
<td>Denominator</td>
<td>50</td>
<td>161.57127</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Our analysis found no significant correlation between Recall and Precision. A printout of the Pearson Correlation appears in Table 20. Conventional wisdom has always been that Recall and Precision have an inverse relationship, in so far as, when one increases, it does so at the expense of the other. The reader will remember that this assumed relationship has fostered the alternative F-measures which discount for particularly lopsided Recall-Precision performance trade-offs. The findings here are
limited in that our sample size is only 60. However, we believe that the results produced here certainly warrant further study into the Recall-Precision relationship especially in light of our experiments in the next chapter which support a finding that precision can be enhanced without significant reduction in recall.

Table 20: Recall-Precision Correlation

|                  | Prob > |r| under H0: Rho=0 |
|------------------|---------|------------------|
|                  | RECALL  | PRECISION        |
| RECALL           | 1.00000 | 0.07847          |
| RECALL           | 0.5512  |                  |
| PRECISION        | 0.07847 | 1.00000          |
| PRECISION        | 0.5512  |                  |

6.9 Discussion

Perhaps the most interesting and significant result produced in this study is that although Total Time Spent Exploring (TTE) is significant for both Recall and Precision, it is positively correlated for recall but negatively correlated for precision. This supports the claim that more time spent exploring the corpus leads to greater recall, but also leads to less precision. This result is consistent with prior research establishing the inverse relationship between Recall and Precision however, prior to this study no empirical explanation has been put forth. The result produced in this study provides a possible explanation for why this relationship is this way. The beta associated with Total Time for Recall was .009 and -0.097 for Precision, suggesting that for every minute increase in
Total Time we should expect to see an increase in Recall by almost .01 and a decrease in Precision by almost .10.

However, the study found that Precision is positively correlated with Number of Documents Viewed; the associated beta of .005, suggests that for every additional document viewed we should expect to see an increase in Precision by .005 units (a two document increase will produce a .01 increase in Precision).

The study found an interactive effect upon Precision by Total time and Number of Documents Viewed with a beta of -.016 for Total Time, a beta of -.013 for Number of Documents Viewed, and a beta for the interactive effect of .0003. This implies that for every 1 minute increase in Total Time, Precision will increase (or decrease) by -.016 + (.0003*number of documents viewed), and for every 1 document increase in the Total documents viewed precision will increase (or decrease) by -.013 + (.0003*time explored).

The linear equation looks like this:

\[
\text{Precision} = B_0 + B_1 T + B_2 N + B_3 T N
\]

Effect of Time on Precision = (B_1 + B_3N)

Effect of Documents on Precision = (B_2 + B_3T)

Where:

T = Total Time Explored

N = Number of Documents Viewed

In terms of behavioral factors impacting Precision, TOA reports a beta value of .005. The TOA inventory used in this study is scored based upon a person’s lack of tolerance, the higher someone scores, the less tolerant they are. This suggests that for every 1 point increase in an individual’s TOA score Precision will increase by .005 units.
This intuitively makes sense, given that people less tolerant of ambiguity are going to focus their search narrowly, resulting in less non-relevant documents being returned. However, TOA was not significant in Recall. DISPO was significant in precision at alpha .05. The associated beta of .002 suggests that for every 1 point increase in DISPO score an individual will produce .002 more units of Precision.

In terms of Recall, the only significant behavioral variable was LOC, at alpha .01. The associated beta of -0.01 suggests that for every 1 point increase in LOC score an individual will produce .01 less units of Recall. A lower LOC score indicates the individual believes he/she controls their fate rather than external factors. Therefore, a higher LOC should lead to less recall and a lower LOC should lead to greater recall.

The results produced are consistent with our original hypothesis that people with greater internal LOC will be inclined to search broader and therefore produce higher recall. One example of perceived control and its effect upon IR came up during our post-task interviews. Subject PG1 indicated that he was; “less concerned about missing documents.” Whereas subject MG2 indicated that; “I feel I may miss ‘the smoking gun.’”

A list of the hypotheses with their measured variables and associated betas is listed in Table 21 below.
Table 21: List of Hypotheses Supported and Not

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Supported/Not</th>
<th>Variable</th>
<th>Alpha</th>
<th>Relationship to Recall/Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1a</td>
<td>Supported</td>
<td>TTE</td>
<td>.01</td>
<td>Recall: Direct and Pos</td>
</tr>
<tr>
<td>H1b</td>
<td>Supported</td>
<td>TTE*</td>
<td>.01</td>
<td>Precision: Direct and Neg*</td>
</tr>
<tr>
<td>H2a</td>
<td>Not</td>
<td>NUM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H2b</td>
<td>Supported</td>
<td>NUM*</td>
<td>.01</td>
<td>Precision: Direct and Neg*</td>
</tr>
<tr>
<td>H3a</td>
<td>Not</td>
<td>PER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3b</td>
<td>Not</td>
<td>PER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H4a</td>
<td>Not</td>
<td>TOA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H4b</td>
<td>Supported</td>
<td>TOA</td>
<td>.01</td>
<td>Precision: Direct and Pos</td>
</tr>
<tr>
<td>H5a</td>
<td>Supported</td>
<td>LOC</td>
<td>.01</td>
<td>Recall: Direct and Pos</td>
</tr>
<tr>
<td>H5b</td>
<td>Not</td>
<td>LOC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H6a</td>
<td>Not</td>
<td>DISPO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H6b</td>
<td>Supported</td>
<td>DISPO</td>
<td>.05</td>
<td>Precision: Direct and Pos</td>
</tr>
<tr>
<td>H7a</td>
<td>Not</td>
<td>PIIT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H7b</td>
<td>Not</td>
<td>PIIT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*- Interactive effect upon Precision supported

As previously mentioned, possible links between the behavioral scales and exploratory behavior were also evaluated for significance. The only significant behavioral variable was LOC. *Time per Document* was affected by LOC with a beta of .034 at alpha .05. This means that for every 1 point change in LOC, an individual will, on average, spend .034 more minutes per document; a significant, but perhaps not meaningful amount of time differential. However, the real insight comes in the form of recognizing that a relationship exists between these variables that can be exploited by the eDiscovery practitioner. Remember that a higher LOC score translates into less internal and greater external LOC. This may be important in situations where the length or complexity of documents in a corpus is of particular criticality. Persons with less internal and greater external LOC will, on average spend more time per document.

The opposite effect was found in *Total documents viewed* (NUM), where the effect had a beta of -3.41 at alpha .01. In this instance, individuals with higher scores on
the LOC scale having less internal locus will, on average, view 3.4 fewer documents for every 1 point increase in their LOC score. This result is both significant and meaningful. The practitioner will be informed that users with higher LOC scores can be expected to view fewer documents.

An unanticipated interactive effect upon *Time spent per document* by DISPO and TOA was discovered to exist. This effect is supported at alpha .05; with a beta of .067 for DISPO, .068 for TOA, and a beta for the interactive effect of -.002. This implies that for every 1 unit change in DISPO score, *Time per document* will increase (or decrease) by .067 + (-.002*TOA score), and for every 1 unit change in TOA, *Time per document* will increase (or decrease) by .068 + (-.002*DISPO score).

The linear equations look like this:

\[
\text{Time per document} = B_0 + B_1D + B_2T + B_3D*T
\]

\[\text{Effect of DISPO upon Time per document=} (B_1 + B_3T)\]

\[\text{Effect of TOA upon Time per document} = (B_2 + B_3D)\]

Where:

\[D = \text{DISPO}\]

\[T = \text{TOA}\]

A beta value of .002 minutes is a rather small number, so this effect although significant, may not be meaningful. Further studies will need to be conducted to determine the impact of this relationship.

Table 29 contains a complete list of the Independent and Dependent Variables used in this study, along with their scales, ranges, means and standard deviations. It has been included in Appendix – K.
Table 30 contains a printout of the Pearson Coefficient Correlations among the Independent Variables. It has been included in Appendix – L.

6.10 Limitations

This study like all studies has limitations. The first limitation lies in the sample size. Several variables were found to not be significant. One possible reason for this is our small sample size; N=60. We plan to address this by collecting more data in future studies.

A second limitation in this study is the use of law students as an approximation for legal professionals such as lawyers and paralegals. In this case, the use of law students was helpful to us because they had the requisite understanding of legal terminology and strategies in litigation, but they were not jaded by years of legal experience that may impact the study. We plan to conduct future studies with paralegals and lawyers to determine if legal experience matters in this form of IR. For that reason we have designed covariates to track data such as this. We plan to implement this design in our next study.

6.11 Contribution

This study has demonstrated the feasibility of an exploration method instantiated through an automated tool that allows users to acquire knowledge about the context of a corpus and apply that knowledge in their search strategy, thereby addressing a major issue researched in eDiscovery IR – how to resolve the dilemma of context to improve recall and precision in large corpora.

The study reported in this chapter makes several significant contributions to theory. The main contribution is the investigation into how exploration can be useful for
large collection information retrieval. The results produced by our experiment support the finding that user exploration of a small portion of a collection will yield improvement at various time intervals. There is clearly a relationship between time and number of documents and IR results produced. How much time and how many documents are needed for a minimum effect will be investigated in future experiments. The results that have been produced by this experiment indicate that there are ranges within which performance improves and ranges within which performance suffers.

We have investigated behavioral and exploration relationships and discovered some new relationships. First, this study has provided insight into how IR behavioral variables may be used to predict a user’s result. Second, this study demonstrates how an exploration model approach to IR can improve performance, specifically when measured against random and non-function word methods.

This study provides insight into which exploration variables can be used to predict IR outcomes and more importantly, how these variables can be used to enhance user productivity.

This study has investigated the underlying constructs of IR in the eDiscovery domain and reported on initial relationships that appear to be present. The next step will be to develop the unresolved questions in future experiments that have come out of this study.
6.12 Future Work

We are encouraged by the results we have produced in this study, particularly in the possible explanation for the recall-precision inverse relationship and in the differing retrieval results for the ranges of time and the number of documents explored. We plan to continue with an additional series of experiments using alternative document collections to cross-validate the results produced here; our next data set will involve a medical records database. We also plan to conduct further behavioral experiments in IR using (RFT) regulatory focus theory. The goal with RFT is to determine whether individuals can be primed to prefer recall or precision preferences; this will provide an additional tool for practitioners who wish to design an eDiscovery strategy that favors recall or precision.

6.13 Conclusion

Study One is designed to measure the significance of the relationship between exploration of a sample collection and the IR result, and how exploration impacts user performance. Conventional wisdom suggests there should be a direct and positive correlation between exploration and result. This study produced results that showed that the relationship is not linear and in fact, at some ranges performance suffers and exploration should be avoided.

The results produced by this study help explain which behavioral preferences have significant impact on exploration and on IR performance. This study also provides an explanation for why recall and precision are correlated in an inverse relationship. The measured variables used in this study help explain user actions and strategies developed during corpus and document exploration and their significance upon IR performance.
The IT artifact developed for this study is a prototype system designed to support the exploration process for eDiscovery IR. Proof of concept is instantiated via the Design Science Paradigm. The contribution of this study lies in its insights of how differences in exploration variables *Total Time* invested in exploring, the *Number of Documents* viewed and *time spent per documents*, and behavioral variables *locus of control*, *tolerance for ambiguity* and *disposition toward innovativeness* impact the IR result as evaluated by *Recall* and *Precision*. 
Chapter 7

Study Two: Using Elimination Terms for IR Document Filtering

7.1 Abstract

Given the large volumes of information contained in electronic stores, tools need to support the retrieval of relevant documents from large collections without producing too many non-relevant documents (Oussalah et al., 2008; Oard et al., 2010; Grossman and Cormack, 2011). A significant concern here is about disclosing too much, such as privileged (non-relevant) documents. An additional concern is about the high cost associated with human review of documents. If precision can be increased, a smaller and more precise collection can be produced by the automated system for human review. This portion of the dissertation focuses on using elimination terms as a method to reduce the number of non-relevant documents in the IR result. The objective is to provide insight into how elimination terms can be used to improve IR performance. The results of this experiment demonstrate that the addition of an elimination component to a document search can significantly reduce the number of non-relevant documents in the retrieval.

7.2 Introduction

Under the rules of evidence, a party may seek documents that are not relevant but may lead to relevance. This is especially true in cases where one party is “fishing” for information (Oard et al., 2010). This can lead to the problem of over inclusion in the IR
result. One effective technique in the literature that can be adapted to this problem is *stop words*.

Stop-words traditionally are *non*-informative words such as “the, a, is” usually ignored by an IR algorithm (Singhal 2001). Informative stop words could be used as filters. This study evaluates the use of filters to represent user selected elimination terms to remove a document from consideration.

User selected elimination criteria can improve precision by preventing a non-relevant document from being considered by the Recall term module. For example, the user may know of a certain type of header or footer, the number of words in a document, or specific terms (elimination terms) that will eliminate a document (containing positive search terms) from consideration otherwise retrieved by a query using Recall terms alone.

### 7.3 Background

An example of a technique used to produce broader recall but can lead to the problem of producing too many non-relevant documents is *stemming*. Stemming is the reduction of different forms of the same word down to its stem or root (Singhal 2001). For instance stemming “fall, falling, and falls” to find documents associated with a person falling (slip and fall cases). The limitation with stemming is that it can lead to non-relevant documents being included due to the polysemy problem; in a “slip and fall” case the user does not want documents about an autumn day in the month of September. Therefore, in order to be effective, terms chosen must translate some characteristic that distinguish the relevant documents from the rest of the collection. This concept has been identified as term discrimination (Spark-Jones, 1972).
The hypothesis here is that term discrimination can be achieved in this research by separating internal context of the relevant document (Recall) from the external content of the corpus (Precision) – eliminating non-relevant documents.

7.4 Methods

The method used is a controlled experiment. The data collected from the participants in Study One are divided into Recall terms and Elimination terms. The data is analyzed using a pair differences test also called a random block design (RBD). The prototype for Study One has specifically been designed to support this data collection effort. The IR task prompts the participants to provide Recall terms and Elimination terms using two different user screens as displayed in Appendix- F.

This experiment is designed to evaluate how elimination terms as a separate module of an algorithm can impact performance in terms of precision as measured by the difference in non-relevant documents produced between samples. By having the entire population of exploration participants provide both Recall terms and Elimination terms the study avoids the possibility that some other input produced the reduction in non-relevant documents. The reader is reminded that the prototype built for Study One is housed on a server and accessed by participants using a URL link from their self provided laptop computers.

7.4.1 Task/Treatment

The task is the eDiscovery retrieval task displayed in the Appendix - C. The treatment is the use of elimination terms. The dependent variable in this case is Precision. Covariates are not tracked in this experiment. The effect evaluated for significance in this
study is the use of *elimination terms* and non-relevant documents retrieved. The variables are listed in Table 7.

**Table 22: List of Variables and Descriptions**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recall Term (IV)</strong>*</td>
<td>User selected and submitted to the system for IR.</td>
</tr>
<tr>
<td><strong>Elimination Term (IV)</strong></td>
<td>User selected and submitted to the system for IR.</td>
</tr>
<tr>
<td><strong>Recall (DV)</strong>*</td>
<td>Not analyzed in this experiment.</td>
</tr>
<tr>
<td><strong>Precision (DV)</strong></td>
<td>Percentage of relevant documents in the retrieval set measured by the difference in non-relevant documents between samples.</td>
</tr>
</tbody>
</table>

* - Collected for Exploration Study

### 7.4.2 Dataset

The data set used is the EDRM version 2 of the Enron collection. The full corpus of this version contains approximately 650,000 to 680,000 email objects depending on the counting of attachments. This data set has been previously validated in the literature (TREC Legal Track Proceedings 2010, 2011).

#### 7.4.2.1 Process

The description of process and architecture depicted in Figure 3 and used for the exploration experiment in Study One is repeated for the elimination experiment reported in this chapter. The main difference in this run process lies in the procedure used after the data has been prepared. A file watcher is used to begin the procedure calls to process the user selected terms for IR. The IR result is saved to separate system bins for export, usually in an Excel spreadsheet, but sometimes raw output is reviewed during the tuning process. The system is designed to run the IR selections from the user for recall terms (saved in the R1 bin) and elimination terms (saved in the R2 bin). A depiction of the system model is displayed in Figure 16.
The retrieval task described in the Appendix – C is presented to the participants via their laptop access to the server application using a URL link.

![Diagram](image1)

**Figure 16: System Design Model**

### 7.4.3 Pilot Study

There have been two pilots performed in this experiment to refine the system and the data collection instrument in preparation for the full population study. The first pilot study consisted of 5 paralegals selected by convenience from Hillsborough County, Florida. Feedback from the pilot study helped us address shortcomings in the initial design. For example, our initial design measured across four methods; it proved to be difficult to measure and hard to isolate effects. Also, in the first study participants were told there were up to 100 relevant documents in the collection. This caused confusion among the participants. Additionally, our panel of experts felt that this may have introduced a bias into the study.

The experiment was redesigned to measure differences between the applications of search terms alone (Recall terms) versus the application of search terms combined with Elimination terms. After feedback was received from both pilot participants and our panel of experts, we redesigned the experiment to the current format. We use a random block design (RBD) for data analysis. In this case the blocks are the participants. Each
block contains two observations. We are measuring the differences between the two observations: paired differences.

A second pilot was conducted with the current design using the artifact with RBD for data analysis; it consisted of 10 paralegals who have worked as document reviewers in eDiscovery. They volunteered their time to assist with the project. All participants were administered the questionnaire displayed in Appendix – G, for demographic information (covariates were not tracked in this experiment). The IR results produced from the second pilot are reported in Table 23.

Precision significantly improved using the elimination component over recall search terms alone. Non-relevant documents were reduced on average from 40.4 to 30. Average reduction in non-relevant documents was 10.8, with the greatest reduction being 18 documents and the least reduction being 4 documents. The standard deviation between non-relevant document samples was 4.4 for Recall terms, 4.3 for Elimination terms. Using this information we obtain the following 95% confidence interval for the true mean reduction in non-relevant documents: \( 10.8 \pm 1.56 = (9.24, 12.36) \).
Table 23: Reduction in Non-Relevant Documents

<table>
<thead>
<tr>
<th>Participant</th>
<th>Non-Relevant Documents Using Recall Terms Alone</th>
<th>Non-Relevant Documents Using Elimination Terms</th>
<th>Reduction in Non-Relevant Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36</td>
<td>27</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>34</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>44</td>
<td>30</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>45</td>
<td>32</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>42</td>
<td>38</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>45</td>
<td>27</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>44</td>
<td>31</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>37</td>
<td>29</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>34</td>
<td>22</td>
<td>12</td>
</tr>
<tr>
<td>Average</td>
<td>40.4</td>
<td>30</td>
<td>10.8</td>
</tr>
<tr>
<td>STDev</td>
<td>4.4</td>
<td>4.3</td>
<td>4.3</td>
</tr>
</tbody>
</table>

CI $10.8 \pm 2.26^* \sqrt{4.3/9} = 1.56$

7.4.4 Full Study Design, Methods and Data Analysis

The method of analysis in this case is a random block design (RBD). In this case we are measuring paired differences; there are two observations for each participant. The study utilizes 30 of the participants from Study One. The user selections are captured using the server based application from the exploration experiment. The difference lies in the implementation of the Recall versus the Recall plus Elimination terms. There are a total of 30 participants with 2 samples from each, representing a total of 60 observations. The blocks are the participants. SAS 9.2 was used to perform the statistical analysis for the random block design/paired difference test. The test is represented by the following model:
\[ E(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 \\
+ \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17} \\
+ \beta_{18} X_{18} + \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21} + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \beta_{25} X_{25} \\
+ \beta_{26} X_{26} + \beta_{27} X_{27} + \beta_{28} X_{28} + \beta_{29} X_{29} + \beta_{30} X_{30} \]

Where:

- \( X_1 \) = A dummy variable representing (0) for Recall and (1) for Elimination,
- \( X_2 \) through \( X_{30} \) = Dummy variables representing participants (1) for each level of participant and (0) for not.

The null and alternative hypotheses are as follows:

- \( H_0: B_1 = 0 \), meaning there is no difference between the mean number of non-relevant documents using recall and the mean number of documents using elimination (e.g. \( \mu_{\text{recall}} = \mu_{\text{elimination}} \)).

- \( H_a: B_1 \neq 0 \) meaning there is a significant difference the mean number of non-relevant documents using \textit{Recall} and the mean number of documents using \textit{Recall} plus \textit{Elimination} (e.g. \( \mu_{\text{recall}} \neq \mu_{\text{elimination}} \)).

### 7.5 Results

The finding produced by the experiment is that the use of elimination terms produced a statistically significant reduction in non-relevant documents resulting in an improvement in precision. The effect was significant at alpha .01, with a 95% confidence interval for the true mean reduction in non-relevant documents: 10.37 +/- 0.949 = (9.42, 11.32), meaning the average reduction in documents we expect to see would be a high of 11.32 and a low of 9.42. The SAS 9.2 printout reporting the results of the RBD/Paired
Difference analysis has been reproduced in Table 24 and the reduction in non-relevant documents have been reproduced in Table 25 on the next pages.

Table 24: Paired Difference/RBD

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARTICIP</td>
<td>29</td>
<td>244304.3500</td>
<td>8424.2879</td>
<td>645.48</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>SAMPLE</td>
<td>1</td>
<td>1612.0167</td>
<td>1612.0167</td>
<td>123.52</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

The GLM Procedure
Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARTICP</td>
<td>30</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30</td>
</tr>
<tr>
<td>SAMPLE</td>
<td>2</td>
<td>1 2</td>
</tr>
</tbody>
</table>

Number of Observations Read: 60
Number of Observations Used: 60
### Table 25: Reduction in Non-Relevant Documents

<table>
<thead>
<tr>
<th>Participant</th>
<th>Non-Relevant Docs Recall</th>
<th>Non-Relevant Docs Elimination</th>
<th>Difference in Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80</td>
<td>76</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>66</td>
<td>61</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>61</td>
<td>56</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>65</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>116</td>
<td>107</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>114</td>
<td>106</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>131</td>
<td>121</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>140</td>
<td>133</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>168</td>
<td>160</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>257</td>
<td>245</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>245</td>
<td>233</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>235</td>
<td>224</td>
<td>11</td>
</tr>
<tr>
<td>13</td>
<td>234</td>
<td>223</td>
<td>11</td>
</tr>
<tr>
<td>14</td>
<td>266</td>
<td>254</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td>236</td>
<td>225</td>
<td>11</td>
</tr>
<tr>
<td>16</td>
<td>252</td>
<td>240</td>
<td>12</td>
</tr>
<tr>
<td>17</td>
<td>226</td>
<td>215</td>
<td>11</td>
</tr>
<tr>
<td>18</td>
<td>220</td>
<td>210</td>
<td>10</td>
</tr>
<tr>
<td>19</td>
<td>239</td>
<td>231</td>
<td>8</td>
</tr>
<tr>
<td>20</td>
<td>223</td>
<td>216</td>
<td>7</td>
</tr>
<tr>
<td>21</td>
<td>195</td>
<td>189</td>
<td>6</td>
</tr>
<tr>
<td>22</td>
<td>189</td>
<td>183</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>169</td>
<td>163</td>
<td>6</td>
</tr>
<tr>
<td>24</td>
<td>238</td>
<td>230</td>
<td>8</td>
</tr>
<tr>
<td>25</td>
<td>205</td>
<td>198</td>
<td>7</td>
</tr>
<tr>
<td>26</td>
<td>216</td>
<td>199</td>
<td>17</td>
</tr>
<tr>
<td>27</td>
<td>288</td>
<td>264</td>
<td>24</td>
</tr>
<tr>
<td>28</td>
<td>273</td>
<td>251</td>
<td>22</td>
</tr>
<tr>
<td>29</td>
<td>247</td>
<td>227</td>
<td>20</td>
</tr>
<tr>
<td>30</td>
<td>210</td>
<td>193</td>
<td>17</td>
</tr>
</tbody>
</table>

Average: **10.37**

Standard Deviation: **5.11**

Confidence Interval: \(10.37 \pm 2.26\)\(*\)

\[
\sqrt{(5.11/29)} = .949
\]
7.6 Discussion

The similarity in recall term selection across participants is consistent with other findings in this area (TREC Proceedings 2009, 2010). Some search terms produced significantly better recall than others. For instance, participants submitting the recall term ‘EOL’ resulted in a significantly higher recall than the other participants in their group who did not submit that term. This is also consistent with other research in the field indicating that retrieval is highly sensitive to choice of search terms (Oard et al., 2010; Grossman and Cormack, 2011). The significant results produced for non-relevant document reduction due to elimination terms are encouraging given that the time and cost associated with human review can be reduced if there are less non-relevant documents in the retrieval (Grossman and Cormack, 2011).

Precision was significantly improved, possibly due to the fact that Elimination terms are applied as exclusionary without regard to weights or threshold. The use of Elimination terms allowed for non-relevant documents to be “eliminated” from the retrieval that otherwise would have been included based on (Recall) search terms alone. However, there were instances when relevant documents were eliminated from the retrieval resulting in a reduction in recall (of relevant documents). This will need to be addressed in future studies to determine if the reduction in recall can be controlled.
7.6.1 Limitations

This experiment has similar limitations to the exploration experiment in that the use of law students as proxies for eDiscovery users may not be a true representation of that population. This limitation is not as severe as with other studies using students. As previously mentioned in the exploration experiment, law students have similar training (albeit, a lack of experience) as eDiscovery users and therefore are a reasonable approximation for the general population in the domain. A larger population would also be more desirable. Our goal is to repeat this experiment with a group of 120 or greater.

A more important limitation with this experiment is that the results may be ad hoc; that is the result produced may be peculiar to this data set or the design of the task itself. This reduces the confidence in being able to generalize the results produced so far. This limitation will be addressed in a future study using a different data set, and different task to improve the universality of the phenomenon found in this experiment.

7.6.2 Future Work

Initial results from the pilots and the full study are encouraging for the use of elimination terms. The next step will be to repeat this experiment on a different data set, using a different task to see if the results produced here can be repeated across circumstances, populations, and environments.
7.6.3 Conclusion

This experiment demonstrated that non-relevant documents can be reduced in IR by the use of elimination terms thereby improving precision in the retrieval result. If non-relevant documents can be reduced without the loss of relevant documents, use of an elimination component can successfully reduce the time and cost associated with human review. Further study of the relationship of elimination terms with recall and precision is certainly warranted. Recall did not vary widely by participant. This is consistent with prior findings that relevance is very context and content dependent. What is new in this dissertation is the evidence that precision can be manipulated through the use of an algorithm designed to accept elimination terms as a separate module in the IR algorithm.
Chapter 8

Using Learning and Relevance Feedback to Improve IR Performance

8.1 Abstract

Consider the following scenario. You are assigned the task of extracting relevant documents from an email collection. The collection contains 32 million objects. You have completed a two week investigation completely immersing yourself into the organization of the custodian of the data. You are now ready to begin the extraction process. What do you do now? You are armed with what seemed yesterday to be a complete and exhaustive set of terms including local vocabularies and document structures. You stare at the computer interface and think. You enter some initial terms based on your theory about the organization of the relevant documents. The result produces 250,000 documents. Now what do you do? Do you choose to review a concordance report of all documents? Do you rely on an inverted index to reformulate your search terms? Your frustration grows as multiple trial and error attempts produce equally ambiguous extractions of documents that must now be painstakingly reviewed for days or weeks by dozens of contracted reviewers.

This vignette is not fiction. It is based on the Philip Morris tobacco cases which, in one instance, resulted in 250,000 documents reviewed by 25 full time workers for 6 months. The example illustrated above describes the problems that exist in large volume IR: How to reduce the volume in the return set? How to measure differences in
performance between return sets? How to change direction of the search structure before too much effort and time has been wasted on a wrong-headed search theory?

The problem of volume associated with IR efforts involving large collections is addressed by the use of an automated tool as an improvement beyond manual review. The use of automation allows the reviewer to cull large numbers of documents within a collection down to a more manageable level for inspection. This action is implemented through rules and algorithms that select the most likely documents relevant to the user’s information need. While automation provides a solution to the problem of volume by reducing the search space, it creates a secondary problem of uncertainty.

Uncertainty describes the limitation in the use of IR tools. Two main practitioner uses of automated tools in IR are predictive matching and exact matching. Tools that are predictive in nature typically rely on a probabilistic method. The automated system must correctly approximate the mental model of the user and apply rules for prioritizing and sorting documents contained in the collection. When the approximation is poor, the document retrieval is imprecise leading to many false positives and false negatives. This limitation is particularly acute in domains where the IR is highly context dependent, and the user is a domain or subject matter expert, and approximation is difficult to achieve. In situations involving highly contextual IR, users often use exact matching. There is an obvious lack of flexibility in this type of search method.

Typically, such IR users formulate queries using Boolean operators in an attempt to fine tune the retrieval (Interview with Bill Hamilton, November 2011) rather than rely on predictive methods. The user will conduct initial IR searches using concordance reports or inverted indexing to map terms and review “hit reports.” This is an imperfect
approach using brute force methods to produce results. This study addresses the limitation of uncertainty by introducing a prototype system allowing users to translate their IR criteria to an automated system via an iterative process by adapting relevance feedback (Schweighofer and Geist, 2008), and Active Learning (Zheng and Padmanabhan, 2006). The system allows the user the ability to judge the structure of their search query by providing a small sample of the IR result. The user is able to review the sample and make adjustments to their search structure through an iterative relevance feedback interface. The learning method here is not “active learning” per se, given that the human and not the machine is doing the learning.

This study focuses on user knowledge of the structure of the document, the context of relevance, and the nature of the corpus. The IT artifact in this study is the prototype system providing a process for the user to leverage domain, content, and context knowledge to improve retrieval performed by the automated system, thereby reducing uncertainty in the process and improving the precision of the IR result.

The system has been developed by adapting two techniques from the machine learning domain and applying them to the domain of context based IR – eDiscovery. The first machine learning technique is Active Learning (Zheng and Padmanabhan, 2006). The second technique is Relevance Feedback presented by Schweighofer and Geist, in their 2008 TREC proceeding paper. The system is evaluated based on the operationalized dependent variables Recall and Precision (Vanrijsbergen, 1979).

The two techniques have been used as methods for an automated tool to “learn” rules and patterns about relevant items and improve performance results. In this study the techniques have been adapted to a system designed to present retrieval sets to the user so
that the human, rather than the machine, conducts the learning. The boundary condition here is an IR domain such as eDiscovery where all items are settled by human review, relevance is highly contingent on context knowledge, the search is ad hoc, and the corpus is a constrained set of similar documents.

8.2 Introduction to Problem Domain

“A user confronted with an automatic retrieval system is unlikely to be able to express his information need in one go. He is more likely to want to indulge in a trial and error process in which he formulates his query in the light of what the system can tell him about his query” (Vanrijsbergen, 1979). Interviews with practitioners of eDiscovery confirm that this claim is still true today over thirty years later. Study Three addresses this problem by developing a prototype system that supports an efficient iterative process as a replacement for the traditional trial and error process.

At some point in the eDiscovery process all ESI documents will have to be reviewed by the decision maker (Sedona Conference 2007). Given the potential penalties for failure to disclose, the decision maker must balance the time and cost associated with the size of the extraction to review against the potential of missing relevant documents.

The first two studies in this dissertation address the problem of how knowledge can be acquired to improve retrieval performance through exploration and elimination terms. This study demonstrates how an automated tool may incorporate exploration and elimination terms, in an iterative method, giving the user greater control and increased interaction with the system as an alternative to brute force trial and error approaches such as reliance on concordance reports or inverted indexing.
The first thing an eDiscovery user must do is investigate the nature of the subject matter and the context of the documents within the collection. Typically this involves the legal team sitting down with the client to brainstorm about what search terms might produce relevant documents from the collection. This process involves in-depth interviews of the client and their support personnel in order to make predictions about the nature of the subject matter, format of document storage, context of relevant document search terms, and content of the collection both structural and organizational (Interview Bill Hamilton, November 2011).

At the conclusion of the investigation, the eDiscovery user believes that he/she has a confident mental model of the collection and the criteria for document relevance. However, the user is limited in search techniques to brute force, trial and error methods using Boolean search terms and operators (Interview with Quarles and Brady Paralegal, March 2011). If the user wants to accomplish a more complex search method for retrieval, a commercial specialist vendor is brought in to apply a statistical or probabilistic approach such as SVM or a variant of LSI. The use of a commercial vendor results in a significant increase in cost many law firm clients will not undertake, resulting in many small to medium level commercial cases being left unresolved (Interview with Quarles and Brady Associate, Douglas Knox, March 2011).

8.3 Research Questions in the Area

Study One informs that exploration improves IR performance and that certain variables are good predictors of IR results. Study One reports that total time exploring a corpus and total documents viewed are significant predictors of performance. Verbal protocols and post-test interviews from the exploration study provided additional insight
into how users determine their strategies for search and what items they focus upon during exploration (Discussed further in Chapter 9).

Study Two, the elimination term experiment, informs us that precision in the IR result can be improved by using specific terms to cull non-relevant documents from the return set. The random block design (RBD)/ Paired Differences analysis demonstrated that reduction of non-relevant documents can be accomplished through the use of elimination terms, without significantly reducing recall, thus saving time and money in the human review stage. This study demonstrates how a system can be designed to translate the performance gains from exploration and elimination to improve an automation tool used for large volume IR.

In Study One and Study Two we created a system designed to acquire knowledge through exploration. Study Three builds on these results by allowing the user to fine tune their knowledge about the corpus and subject matter, and leverage that knowledge through an iterative system based on learning and relevance feedback.

Study Three addresses the lack of flexibility problem mentioned at the beginning of this chapter by operationalizing user knowledge and supporting the iterative process in an automated system. This study extends the work of the first two studies by developing and implementing the above mentioned system as an instantiated IT artifact addressing a business need in eDiscovery of how to harness user knowledge through exploration and iteration to improve retrieval performance.
8.4 Importance of Research

This study evaluates a new approach to eDiscovery IR. The approach is demonstrated as an IT artifact. The artifact supports user learning by presenting a retrieval set to the user. The set is generated by an automated tool based on the user’s selected search structure. The artifact offers the user the ability to adjust his/her search structure based on their evaluation of the retrieval set using an iterative relevance feedback approach. This research focuses on a specific problem in the legal domain of how to provide the IR user with flexibility in an automated tool to go beyond brute force trial and error and addresses uncertainty associated with predictive methods.

The artifact in this study provides the user with “insight” into the consequences of his/her choice of search criteria by presenting a small collection of the search results for iterative relevance judgments. It allows for the user to tune the system by providing rules and hints that are factored into the next extraction sample.

This iterative feature allows the user to have hands-on ability to manipulate the search process, rather than taking a black box approach where the user is confronted with a limited “hit report” disclosing term frequency, concordance, or having the entire search result returned and the user making culling decisions for refining the next iteration.
8.5 Approach

We answer the business problem posed in this study by developing a prototype system to instantiate the IT artifact. The artifact is based on a method of learning (Debowski et al., 2001; Hills et al., 2010) adapted from Active Learning (Zheng and Padmanabhan, 2006) using relevance feedback (Schweighofer and Geist, 2008).

We adapt the learning method here from active learning by shifting the focus of the learner. The traditional active learning technique is based on machine learning. The system “learns” the patterns and improves performance. In this case, it is the user who is learning; the system simply supports the process.

One can think of our method as compared to power steering on a vehicle. It is still the driver doing the steering, but he/she has the leverage of the machine power; versus the traditional active learning approach in which the car does the driving after receiving input from the GPS or other selective data input.

The system developed in this study is evaluated using document reviewers as participants to measure system performance in the IR result. The objective of the system is to return a relevant document set to the user based on the user’s judgments. This approximates the real life situation where the user is a legal professional attempting to explore a client’s corpus of documents with the goal of extracting only those documents judged in the user’s mind to be relevant. This scenario involves the problem of how to translate a user’s mental model to an automated tool based on context knowledge.

The solution presented here is a system that allows a user to continuously refine their search structure through an iterative process (Vanrijisbergen, 1979). The user is
presented with a retrieval set and makes adjustments to his/her search structure based on what he/she has learned.

The artifact in this study is not a learning algorithm; it is an algorithm for learning. The foundation that underlies this learning system is based on prior research on the exploratory nature of search and frameworks describing how machine learning has been used as a solution in this area.

8.5.1 Learning

“The search for information is often a cyclical, exploratory process” (Debowsk et al., 2001). Search has also been compared to problem solving techniques similar to foraging (Hills et al., 2010). Hills et al., characterize problem solving itself as a search process. The decision regarding when to exploit – stay with the current position or strategy, versus when to explore – move on to a new search or location is a trade-off that has been studied in problem solving and learning (Robbins, 1952; March 1991; Hills et al., 2010). This is especially true in the domain of eDiscovery IR where the search can be very complex in terms of strategy and structure (Debrowski et al., 2001) and the domain is highly context driven (Grossman and Cormack, 2011). The learning taking place here is done by the user based on the feedback received from the system in the form of documents retrieved pursuant to the user’s inputted search structure.

The artifact developed in this chapter addresses the issues presented above. Once the user has acquired knowledge about the corpus being searched via the first artifact, the second artifact supports the user via an iterative learning method by presenting retrieval results in small sets to the user. He/she can assess the results and adjust the search structure to improve the retrieval result. The goal of the second artifact is to address the
gap in electronic search identified by (Dembroski et al., 2001) as; “not highly informative regarding the effectiveness of strategies.” They suggest that in order to achieve successful retrieval, the search structure and alternative strategies must be continually evaluated. This is addressed by the second artifact using a learning tool with iterative feedback.

8.5.2 Active Learning

“Active Learning” is a subfield of machine learning. It is sometimes called query learning or optimal experimental design (Settles, 2009). The fundamental principle is about allowing a learning algorithm to choose the data from which it learns about the over-all collection. Study Three adapts the active learning approach by re-focusing the learning to the user. In this implementation, the user chooses the data from which he/she learns about the over-all collection.

The difference in the implementation of active learning in this study lies in the source of the learning. In the traditional active learning model the algorithm learns from selected data in the collection. In this study the user learns from the algorithm via relevance feedback. Rather than relying on pattern recognition and probabilistic methods, this system relies on the user inputted criteria and human judgment based on context knowledge. This instantiation is appropriate due to the fact that the IR user in this situation has a special and significant grasp of the subject matter and the context of relevance. The conventional modeling method of this knowledge is automated indexing or extensive manual interviews of users by search specialists.

The traditional active learning system “selectively acquires” items (Zheng and Padmanabhan, 2006). The system developed here acquires items for presentation to the user based on user chosen search structure. The user labels the documents as relevant or
not relevant. The labeled documents are stored in bins accordingly. The user then enters
modifications to the search structure. The system makes the next round of extractions
based on the user criteria and presents the set for iterative feedback. In this study the
number of iterations is fixed. In the real world application the system is evaluated by the
number of iterations required for achieving a predetermined recall and precision
threshold.

(Zheng and Padmanabhan, 2006) present an information acquisition problem
where data has to be acquired with a specific modeling objective in mind. They develop
an active learning technique for the acquisition problem using machine learning and
optimal experimental design. The characteristics defining their problem are: (1) Some
readily available data is generated as a natural consequence of a firm doing business, (2)
There is a specific modeling objective and a target variable, and (3) Additional useful
information is not readily available.

eDiscovery and other expert user IR share these three characteristics with Zheng
and Padmanabhan’s information acquisition problem. The main differences are: (1) The
central problem in IR is not that there is SOME readily available data, but that there is
TOO MUCH readily available data – however the data is generated during the normal
course of business; in fact, quite often the data generated is directly related to the cause of
action in the legal case. (2) While there is a specific modeling objective in mind, the
target variable in this case is relevance – the matching of certain documents based on
criteria often illusive to define, resulting in the use of terms vulnerable to ambiguities. (3)
In eDiscovery, TOO MUCH “un-useful” information is available – the contra-positive of
the Zheng-Padmanabhan criteria. (4) The eDiscovery user is doing the learning, not the system.

Zheng and Padmanabhan address the problem of determining how to selectively acquire additional data. Study Three adapts active learning to eDiscovery in order for the user to selectively acquire additional data. This is done using an iterative algorithm and small retrieval sets (our research suggests 10–20 items) from the collection.

8.5.3 Relevance Feedback

“Users generally seek information in an iterative manner” (McCay et al., 2004). Relevance feedback is a process whereby the system is given an information need -- usually in the form of a query, and an initial set of hints from the user regarding exemplar relevant and sometimes non-relevant documents (Zhao et al., 2008). The process can also take the form of iterative feedback whereby the user participates over several rounds (Harman, 1992).

This study develops a prototype system to support an iterative method of user relevance feedback whereby the system selects a set of documents and the user assesses the documents for relevancy and makes adjustments to the search criteria based on what is learned from the retrieval set.

The vast majority of research on relevance feedback has focused on probabilistic methods using vector space and re-weighting based on the feedback (Rocchio, 1971; Salton and Buckley, 1990; Harman, 1992; Zhao et al., 2008).

(Schweighofer and Geist, 2008) proposed a model for legal information retrieval using relevance feedback. The purpose of their research was to improve Boolean search with query expansion. They developed a prototype based on legal vocabularies and “legal
language” to represent descriptions of legal concepts. Their focus however, was on legal information retrieval of legal texts based on ontologies.

Schweighofer and Geist published their model in the Text Retrieval Conference Legal Track (TREC) 2008 Proceedings. This model proposes an explanation of the relationship of users to documents through the medium of ontology and relevance feedback. Their model displayed in Figure 7 depicts the impact of relevance feedback and ontology, leading to query expansion.

![Schweighofer and Geist Relevance Feedback Model](image)

**Figure 17:** Schweighofer and Geist Relevance Feedback Model

### 8.6 Significant Prior Research

Significant research has been done in an attempt to apply the above techniques to improve IR. (Zhang and Chen, 2002) proposed a general active learning framework for content-based IR. Their empirical results found that their active learning algorithm outperformed random sampling. They made a significant assumption that semantic meaning is inferable and that semantic meanings of objects can be characterized by a multilevel attribute tree.

Their model relies on prior probabilities to initialize their list, and their first selection set is based on a random extraction of objects. They also assume that their
model will infer knowledge about items from nearby neighbors. The limitation of their work lies in how they define the reduction in uncertainty. They view it from a probabilistic and predictive approach — by the improvement in attribute probabilities.

Another limitation in their research lies in their data set and goals. Their data set was limited to 1750 objects. Their goal was a content-based experiment to distinguish between two types — aircraft and non-aircraft — and then further distinguish within those groups into sub categories.

eDiscovery IR is much more content intensive and context rich insofar as it cannot be easily modeled in such a simple binary fashion. Zhang’s and Chen’s work serves as a foundation to demonstrate that the framework of active learning can be applied to content-based IR but their method was completely automated. eDiscovery users require a more ad hoc approach to reducing uncertainty. The approach taken in this study is a variant on active learning insofar as it implements the user as learner.

We address the uncertainty problem by including the human in the loop instead of a probabilistic approach. We also do not use a random extraction to initialize our first iteration. Our assumption is that user knowledge gained from exploration provides the initial criteria for the first iteration retrieval set.

Relevance feedback has been investigated as a technique to improve active learning methods (Raghaven et al., 2006; Onoda et al., 2003). Research done by (Raghavan et al., 2006) found that human in the loop feedback can be used to improve active learning methods where the “human teacher can have significant knowledge on the relevance of features.” eDiscovery is just such an area. (Raghaven et al., 2006) extended
work done by (Sebastiani, 2002) and (Lewis, 1998). Their research concentrated on improving classifier performance by using humans to label features.

Labeling features has the potential to be very helpful in IR searches that are organizational dependent such as emails, where the user may be able to target features such as length of emails, or emails containing attachments. Also feature labeling would be helpful when the ‘To,’ ‘From,’ or ‘Subject’ lines could be identified. Raghaven et al.’s experiments used robust data sets containing 12,902; 20,000; and 67,111 objects. However, the limitation with their research has to do with the parameters of their experiment. They conduct one-versus-rest and binary classification experiments. They also use Support Vector Machines (SVM) as their method for classification. By contrast the domain of eDiscovery does not fit neatly within one-versus-rest such as hurricane versus not a hurricane, or binary classifications such as football versus baseball, as implemented in the Raghaven et al. experiments.

Our approach also uses human in the loop, but we do not create vectors to predict matches between features. Our human in the loop experiment is designed for the user to gain leverage over the search process. The reason for this lies in the uniqueness of context-based IR such as eDiscovery.

(Xu et al., 2007) investigated the document set given to the user for feedback. Their research explored the question of how to best determine the initial set of documents to be given to the user and how to incorporate the user feedback for the next set of results. They defined the problem as an application of active learning in ad hoc information retrieval which is exactly how eDiscovery is described by practitioners and
researchers alike. They use the parameters of document relevancy, density, and diversity to select a sample set.

(Xu et al., 2007) apply language modeling to calculate a relevance factor. They estimate document density by measuring the average distance from the current document to all other documents. For diversity, they measure the distance between a document and the document set, similar to single linkage in hierarchical clustering.

By contrast, our feedback approach uses direct input from the user selection of initial search terms. We provide the user with a sample of results which they judge as relevant or not, and add terms or remove terms. Our method addresses the ad hoc nature of the IR directly. We also address the context and content intensive nature of the IR by focusing on those with the knowledge – expertise through exploration.

8.7 Methods (Development of Prototype)

Our method is an alternative to the probabilistic approaches as indicated in the above examples. We focus on the human/user and keep the human in the loop for the duration of the process. We apply automation as a means for the user to reduce the search space and shorten the time in review from iteration to iteration. Instead of relying on concordance and hit reports, the user is able to have instant, real time results from their search structure criteria. We substitute a probabilistic algorithm based on a predictive approach with an ad hoc structure of user search criteria.

The IT artifact for this study is a learning tool prototype designed to present a small set of extracted documents from a targeted corpus based upon user inputted criteria. The prototype provides the user with the opportunity to explore and exploit via iterative relevance feedback. The use of the iterative technique addresses the flexibility problem in
conventional trial and error methods. The technique of relevance feedback also addresses the problem of imprecision resulting from uncertainty when using predictive automated IR tools.

The artifact combines approaches from the earlier experiments of exploration and elimination and instantiates them by implementing a learning approach adapted from active learning (Zheng and Padmanabhan, 2006) through relevance feedback (Schweighofer and Geist, 2008). The model for this prototype system is displayed in Figure 18 (Legal Intelligence® Model); it solves a specific business problem existing within legal informatics – investigation and retrieval of ESI documents for eDiscovery. The user screens of the prototype are displayed in the Appendix - F.

Study One and Study Two report on a series of experiments. The exploration study has a behavioral component that seeks to explain how user performance in IR can be improved with exploration and which independent variables help predict IR results. The experiments also have a design science approach by evaluating the impact of elimination terms upon the IR algorithm.

Study Three is both behavioral and design science. It is behavioral, insofar as it seeks to determine if a new process will improve user selections, and also design science in that it evaluates the artifact by demonstrating how the process and tool supporting improvement in the IR result is feasible (Hevner and March, 2003).
8.7.1 How the System Works

The system is a web based application. It is implemented using a .NET framework with JSON and MS Silverlight to improve transfer of data between the server and the client. There are two front end interface screens supporting user interaction with the system. The screens are displayed in Appendix - F.

The first screen is a simple user input screen displayed in Figure 31. We decided to use a clean, simple user display. The screen provides text boxes for the user to enter selections for recall terms and elimination terms based on their knowledge acquired from Study One and Study Two. The text boxes accept natural language search terms meaning it ignores spaces and matches the terms to words in the documents by specific location of each letter of the word, relative to the letter before it and the letter after it.

Each time the user enters a phrase or term and hits the enter button, the term or phrase appears in a ribbon below the text box. Recall terms are displayed with a green underline and elimination terms are displayed with a red underline. Once the user has

Figure 18: Legal Intelligence® Model for eDiscovery IR (Study Three)
entered all of their initial search terms, the system extracts the first twenty documents meeting the criteria and presents the results in the second screen.

The second screen is the *eDiscovery Learning Interface* prototype displayed in Figure 32. There are many things happening in this screen. There are two list boxes at the top of the screen. These boxes contain the recall and elimination terms. The user has the option of removing terms by selecting the ‘X’ button located next to the term. The green and red underlines help the user keep track of recall versus elimination terms. There is a counter that displays the current iteration. For this study we used a finite number of 10 iterations.

Below the list boxes are two additional displays. The left side of the screen displays the list of documents retrieved. The user selects a document and its contents are displayed in the reading window on the right side. The document list is presented based on titles. In this case the titles are extracted from the subject lines of the emails as contained in the corpus. On the right side of the screen are the radio buttons for relevance and not relevant.

When the user selects a document and determines it is relevant or not, the system places that document in a corresponding bin. The user is able to tune the IR structure by removing terms from the list boxes up top or by entering additional terms in the text boxes below the relevance radio buttons. When the user has completed their review of the sample, the *Next* button, located at the bottom of the screen, is selected. This completes the current iteration and the system presents the next sample of documents based on the user criteria adjustments.
An additional feature included in the system is the user’s ability to short cut the iteration. The user may proceed to the next iteration at any time by selecting the *By-Pass* checkbox. This will result in the system activating the *Next* button for shortcutting the present iteration and conducting a new extraction based on the terms present in the ribbon above.

In this study the number of iterations has been fixed at ten. However, the real life version of the system will allow the user to continue with iterations until a final selection of search structure has been determined.

### 8.8 Design for System Testing

This study is a controlled experiment in eDiscovery learning, using relevance feedback. The experimental design is a repeated measures/random block design (RBD). The study is designed to measure whether the system produces significant results among the iterations. In this study we use 30 participants with 10 iterations each, producing a total of 300 observations. We also calculate the average recall and precision per iteration and display the results in a bar chart for visual illustration.

The reason here for fixing the number of iterations at ten is that the purpose of this experiment is to demonstrate feasibility in the design and implementation of the method. Future work will extend this research by fixing a level of precision or a lever of recall and measuring the number of iterations needed to achieve the chosen performance levels.
8.8.1 Participants

The participants for this study are 30 third year law students conveniently selected from a pool of 60 cohorts who had participated in Study One. The purpose of the study is to evaluate a system that provides a method for user relevance feedback based on corpus knowledge gained through exploration. Therefore, the correct population frame to draw upon would be the participants who have already acquired knowledge about the corpus through exploration. For this reason we enlisted the assistance of the participants from Study One.

8.8.2 Task and Treatment

Participants are assigned the same eDiscovery task from Study One. This allows us to use participants who are already familiar with the application, methods, and exercise. Informed consent, task instruction and data collection instrument are displayed using computer screens from the application – they are the same display screens from Study One in Appendices – A, B, and C.

All participants are given the same task. The task is divided into three requirements. The first requirement is for the user to provide an initial set of recall (search) terms and optional elimination terms (filters) in response to an eDiscovery request based on the user’s acquired knowledge from exploration (Study One). The second requirement is for the user to review a set of documents identified by the system as relevant based on the initial user input. The third requirement is for the user to assess the retrieval set and adjust his/her search structure based on the relevance feedback. The experiment iterates for 10 rounds. The eDiscovery task is the same task as Study One; it has been adapted from the TREC Legal Track 2011 Conference Problem Set #401.
8.8.3 Data Set

The data set used is 10,000 documents randomly selected from the EDRM version 2 of the Enron collection. The set has been seeded with 1,000 known relevant and 9,000 known non-relevant documents. The full corpus of this version contains approximately 650,000 to 680,000 email objects depending on the counting of attachments. This data set has been previously validated in the literature (TREC Legal Track Proceedings 2010, 2011).

8.8.4 Process

The artifact in this experiment is a server based application designed to accept user selections via an internet browser as an interface. The application stores the user selections and then pushes the data files. The system then generates twenty randomly selected documents from the retrieval for presentation to the user for relevance feedback. The twenty documents are presented to the user via the browser window. The presentation screen displays a list of documents by title. The user selects on a document, and its content is displayed in a reader window next to the document list. Two radio buttons are presented to the user to select relevant or not relevant for each document.

Documents identified as relevant by the user are saved in the retrieval bin. Documents identified as non-relevant by the user are marked for discard so that they are not selected by the system in the next round. The additional terms provided by the user are absorbed by the system and used to select the next set of documents. This experiment terminates after ten iterations. However, the real world application is designed to continue to make iterative selections until terminated by the user. At the termination of the iterations, the system pushes the entire data set through the collected user terms
search array and extracts the retrieval documents. The application implements MS Silverlight in the browser and JSON to process the data between the client and the server.

8.9 Experimental Design

In this case we are using a random block design (RBD). The dependent variable of interest in this study is the individual iteration. The blocks are the participants. There are 10 iterations performed by each participant for a total of 30 blocks, with 10 observations within each block for a total of 300 observations.

We are concerned with finding out if a significant difference exists between the iterations as measured by the dependent variable Precision. The null and alternative hypothesis is as follows:

\[ H_0: \text{No difference exists among the iterations.} \]

\[ H_a: \text{At least one of the iterations is different.} \]

8.10 Results

30 participants produced 10 relevance feedback iterations. This gave us a total of 300 observations to analyze. SAS 9.2 was the package used to perform the analysis using a RBD/Repeated Measures designed experiment. The result for Precision was significant at alpha .01. The result for recall was not significant. The SAS 9.2 output reports for this experiment are displayed on the next page in Table 25 and Table 26.
**Table 26: SAS 9.2 Report for Precision for eDiscovery Learning Experiment**

```latex
\begin{tabular}{lrrrrr}
  \hline
  Class & Levels & Values \\
  \hline
  PARTICIP & 30 & 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 \\
  ITERATIO & 10 & 1 2 3 4 5 6 7 8 9 10 \\
  \hline
  Number of Observations Read & 300 \\
  Number of Observations Used & 300 \\
  \hline
\end{tabular}
```

```latex
\begin{tabular}{llllll}
  \hline
  The GLM Procedure & & & & & \\
  Dependent Variable: PRECISION \\
  \hline
  Source & DF & Type I SS & Mean Square & F Value & Pr > F \\
  \hline
  Model & 38 & 1.73085333 & 0.04554877 & 8.43 & <.0001 \\
  Error & 261 & 1.40941333 & 0.00540005 & & \\
  Corrected Total & 299 & 3.14026667 & & & \\
  \hline
  R-Square & Coeff Var & Root MSE & \multicolumn{2}{l}{PRECISION Mean} \\
  \hline
  0.551180 & 9.585005 & 0.073485 & \multicolumn{2}{l}{0.76667} \\
  \hline
  \end{tabular}
```

**Table 27: SAS 9.2 Report for Recall for eDiscovery Learning Experiment**

```latex
\begin{tabular}{lrrrrr}
  \hline
  Class & Levels & Values \\
  \hline
  PARTICIP & 29 & \multicolumn{3}{l}{29} \\
  ITERATIO & 9 & \multicolumn{3}{l}{9} \\
  \hline
  Number of Observations Read & 300 \\
  Number of Observations Used & 300 \\
  \hline
\end{tabular}
```

```latex
\begin{tabular}{llllll}
  \hline
  The GLM Procedure & & & & & \\
  Dependent Variable: RECALL \\
  \hline
  Source & DF & Type I SS & Mean Square & F Value & Pr > F \\
  \hline
  PARTICIP & 29 & 0.37309467 & 0.01286533 & 5.95 & <.0001 \\
  ITERATIO & 9 & 0.03204800 & 0.00356089 & 1.65 & 0.1022 \\
  \hline
  R-Square & Coeff Var & Root MSE & \multicolumn{2}{l}{RECALL Mean} \\
  \hline
  0.417968 & 7.085874 & 0.046493 & \multicolumn{2}{l}{0.656133} \\
  \hline
  \end{tabular}
```
Our focus in this study was to improve precision in the retrieval. However, there are several explanations for why there was a failure to detect a significant result in recall. The most obvious would be that the sample was not large enough to detect a difference. Another possibility is that there is in fact no significant effect upon recall. This will be investigated in greater depth in future experiments where we will be conducting large scale experiments designed to further study the recall-precision relationship in IR.

Each iteration level was averaged across the 30 participants. An additional and unanticipated observation was that average precision improved for successive iterations. This result is illustrated in Figure 19 below. Average precision for the first iteration was 0.70. Average precision for the final iteration exceeded 0.80. The explanation we offer is that actual learning took place and users simply got better at focusing their search structure resulting in improved precision as iterations increased. This is not to suggest that the tool is responsible for the learning. The tool is simply a tool, and the learning is facilitated through its use.

![Figure 19: Average Recall and Precision by Iteration](image)
While precision (indicated in red) continued to improve, recall (indicated in blue) remained fairly consistent. This is important to note, given the prior assumptions regarding the trade-offs between precision and recall. The results produced in this study suggest that under certain circumstances, recall and precision may not have to be traded-off against each other. This result is also consistent with the results produced in Study One, indicating that recall and precision may not always be negatively correlated.

8.11 Discussion

The main goal of this experiment was to demonstrate that a learning tool based on the principles of learning and relevance feedback is feasible and can be implemented to produce significant results in eDiscovery IR. We achieved that by producing significant results from a random block design/repeated measures experiment. An additional interesting observation we discovered during our analysis has to do with the recall and precision results as seen in Figure 19.

The bar chart in Figure 19 reveals that although precision continued to improve with each successive iteration, recall remained in a fairly consistent range and, more importantly, did not appear to suffer significant loss. Precision on average improved from .7 to .8, while recall remained in a range of .6 to .7. This is encouraging and we plan to further examine this relationship in future studies to determine whether precision can be improved while keeping recall stable. The ability of the user to review sample sets from their search criteria allows the user to refine and reapply their IR search structure in real time. This phenomenon is consistent with findings from the exploration study regarding how individuals perform search generally.
There are two advantages to the method implemented by the prototype in this study. The first advantage is in the design and presentation of the return set. By presenting a limited number of documents, in this case 20, the user is given a more manageable amount of data to review, versus the traditional approach of “hit reports” and concordance reports. This method can be likened to reviewing the first page from a web search and then using the knowledge gained from that small set to refine and craft a new search structure.

The second advantage we give the user is in the form of metric analysis. The user can evaluate the changes in the quality of the result produced by reviewing the precision for each set. This method could also be applied to an unbounded, scale free approach which currently presents a fresh set of documents to the user with no base-line measure for improvement between iteration. In such instances the user is left with imprecise reliance on their own memory or other ambiguous methods to track differences in results produced between IR sets. A tool designed to track the precision for each subsequent search could be a useful improvement as a plug-in to web search tools such as google.

8.12 Limitations

One of the limitations in this study is the insular nature of the population studied. Although third year law students make a good approximation for legal professionals, they are not in fact legal professionals. They have no experience in litigation or in the budgeting constraints that come with document extraction in support of litigation. Another potential limitation is the small number of participants used which decreases the power in our analysis and our ability to make generalizations about our results. We
addressed this limitation by increasing the number of iterations to 10 thus increasing our total observations to 300.

8.13 Contributions

If we reflect on the results produced by this study with respect to the scenario presented in the vignette at the beginning of this chapter, the artifact as presented provides the user an opportunity to learn in real time from his/her search structure by reviewing the return set. This simple procedural device gives the user the ability to examine the impact of their search criteria and make refinements and changes to their search structure without having to sink significant amounts of time and effort into a wrongheaded search theory. This addresses the age old dilemma of exploration versus exploitation presented in Chapter Six (March, 1991; Debowski et al., 2001).

The specific outcome of this study has been the creation of an artifact, implementing an eDiscovery learning tool based on an adapting a learning technique through relevance feedback. This study applies the learning technique to a new domain of IR and a new method for implementation. The new domain is eDiscovery, characterized by subject matter expertise, acquired knowledge, and relevance defined heavily by content and context dependence. The method for implementation is user learning instead of machine learning.

We also extend research in active learning and relevance feedback by using human in loop in a new way. The use of the human to learn the relevance of documents and the system to collect recall and elimination terms demonstrates the feasibility of how human learning can be used successfully in this type of IR. The next step will be to
compare the results achieved using human learning to results achieved using a probabilistic method such as SVM or another method.

8.14 Future Work

As a result of the successful trials of the learning tool, we plan to follow up this study with a series of experiments to examine how to refine the tool to increase recall as well as precision. We also plan to repeat this experiment with legal professionals to determine whether our results can be validated across different populations.

We plan to conduct a new series of experiments to evaluate the performance of users who rely completely on the learning tool, versus users who acquire corpus knowledge using the exploration tool prior to using the learning tool – a real world scenario that occurs during IR.

8.15 Conclusion

This study demonstrates that an IT artifact designed to improve precision in eDiscovery IR is feasible. This study extends the technique of learning adapted from Active Learning, to the new domain of eDiscovery IR, implemented through continuous user feedback by an automated system (human learning). The results of this study establish that this method of user feedback produces a more efficient document extraction as evidenced by the significance in change in the Precision variable.
Chapter 9
Post-Test Interviews

Questionnaires measuring usability of the two systems and participants' exploration strategies were completed by all. Post-Test interviews were conducted with selected participants.

The responses from the participants indicated that the systems were easy to use and simple. Users overwhelmingly responded that the systems were efficient and “gave great results,” (Participant-ET). The majority of users responded that “what they liked most about the system [was how] easy it was to operate.” The majority of the users indicated that the system was useful in helping them complete the task.

On the scaled question of “How likely you are to use a system like this for an information retrieval task?” the average response was 7.2 on a 10 point scale. On the free response question of “In what ways did the system fail to meet your needs for the task?” a significant portion of the subjects responded that it did not disappoint them in any way.

Several subjects reported that they preferred the method of Study Three (Learning) over Study One (Exploration) because it offered them the opportunity to “instantaneously” select terms they wanted or did not want as they were browsing the collection.
On the issue of exploration strategy, most participants indicated that they mostly rely on a trial and error method using keyword terms. This is consistent with prior research done in this domain. During our post-task interviews we followed up on this issue with the question of “How do you conjure up keywords/search terms to use?”

The answers varied. Overall, we found that participants interviewed drew blanks and struggled with this question. Few were able to offer concrete answers into how they initialize their search terms. Several subjects reported that they go to forums and scan blogs to get an idea of what terms people use to describe the subject of the IR. Other answers included: “skimming blindly,” “using control F to find phrases in exemplar documents,” “sometimes just looking at titles,” “looking for terms that may stand out more than others.”

When we asked these same questions of the practitioners, we found that the main source of search terms for eDiscovery IR comes from interviews of clients as either the custodians of the collections or parties to the transactions leading to the litigation.

When we began this dissertation we did not anticipate that a lack of verbosity to the question by the subjects. We plan to further investigate this phenomenon about how people find their initial keywords by designing a new questionnaire to focus on this subject.
Chapter 10

Conclusion and Future Research

This dissertation presents research in the domain of Information Retrieval as applied to legal informatics, specifically eDiscovery. It explores methods to solve the problems of volume and uncertainty in the application of an expert user in eDiscovery IR.

The research conducts analysis over three related studies designed to develop and evaluate methods to improve user and system performance through exploration and learning.

The artifacts developed in this dissertation represent constructs (exploration, learning, and relevance feedback), methods (job run, architecture, and user interface), and instantiation (implementation of eDiscovery Learning Tool system, demonstrating feasibility and concrete assessment) of a solution to a real world business problem guided by the Information Systems Research Framework (Hevner and March, 2003).

The work done in this dissertation is planned to be extended in two additional studies: (1) Designing a suggestion system to recommend other terms and documents to the user based on selections inputted from relevance feedback iterations; (2) Using the theory of Regulatory Focus (RFT) to prime eDiscovery users to prefer recall strategies over precision strategies and vice versa, as a method to align the IR strategies of the user with the unique strategy for the type of litigation case.
The plan for future research is to focus on repeatability of the study results and to further investigate the effect of exploration upon IR performance. This will be done using three new data sets. The first data set will be the “Tobacco” set provided by the Illinois Institute of Technology. The set contains approximately 1 to 2 million objects and will be a significant leap in scale from the 680,000 objects used in the present studies. The second data set will be the TREC 2012 data set. This set is scheduled to be released within the next few months and will provide an opportunity to generalize the conclusions reached in the studies reported in this dissertation using the Enron collection from TREC 2010 and 2011. The third data set will be a medical set provided by the TREC 2012 Medical Track.
References


Borko, H., “Measuring the Reliability of Subject Classification by Men and Machines,” American Documentation, October (1964).


Xu, W., Gong, Y., “Document Clustering by Concept Factorization,” Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.


APPENDIX A

IRB Approval Letter
December 8, 2011
Harvey Hyman, ESQ
Information Systems & Decision Sciences
2410 W. Morrison Ave
Tampa, FL 33629

RE: Exempt Certification for IRB#: Pro00005763
   Title: eDiscovery IR Study.

Dear Mr. Hyman:

On 12/8/2011 the Institutional Review Board (IRB) determined that your research meets USF requirements and Federal Exemption criteria as outlined in the federal regulations at 45CFR46.101(b):

(2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless:

(i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

Please be advised data collection cannot begin until the IRB receives and approves your letter(s) of support from the non-USF/non-Affiliate site(s) which must be submitted as an amendment.

As the principal investigator for this study, it is your responsibility to ensure that this research is conducted as outlined in your application and consistent with the ethical principles outlined in the Belmont Report and with USF IRB policies and procedures. Please note that changes to this protocol may disqualify it from exempt status. Please note that you are responsible for notifying the IRB prior to implementing any changes to the currently approved protocol.

The Institutional Review Board will maintain your exemption application for a period of five years from the date of this letter or for three years after a Final Progress Report is received, whichever is longer. If you wish to continue this protocol beyond five years, you will need to submit a new application. When your study is completed, either prior to, or at the end of the five-year period, you must submit a Final Report to close this study.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

John A. Schinka, Ph.D., Chairperson USF
Institutional Review Board
APPENDIX - B

Informed Consent

We are conducting a research study on factors that impact information retrieval. My name is Harvey Hyman. I am the primary investigator. My phone number is 813-966-4278, and my email is HymanLaw@gmail.com. I am a dissertation candidate at the University of South Florida, Information Systems and Decision Sciences Department of the College of Business Administration.

The title of this study is eDiscovery IR Study. The IRB number is Pro 5763.

The purpose of this study is to determine two issues of interest: (1) Whether a significant relationship exists between user exploration of a corpus and the level of retrieval, and (2) Whether a significant relationship exists between combining the use of elimination terms with search terms and the level of retrieval.

Any person who is capable of using a computer interface and understands electronic search is eligible for this study. There is a 2 hour time commitment for this study.

In this study you will be asked to log-on to a server to begin the study. You will be prompted to input demographic information. After you have inputted your information you will be given the discovery task.

The data collected by this study will be held on a server at the USF campus. The data will be used for academic purposes only. All demographic information collected will identify users by participant number. The data collected in this study will be used to publish research articles in journals.

I thank you for your participation. Your participation is completely voluntary. The USF IRB phone number is 813-974-5638 should you have any questions or concerns.

By participating in this study you are agreeing to take part in this research.

We will share this data only for academic research purposes with detailed data description publicly catalogued for the public review through the USF libraries web site at http://lib.usf.edu, and the actual data will be freely shared with other researchers whenever permission is granted, under the agreement of the Primary Investigator, Harvey Hyman, and Senior Personnel and the University.
Researchers who will use our data set must sign an agreement that they will put in place their maximum data protection measures, and they will use this data only for their academic purposes. This data will not contain any personal information.
eDiscovery Retrieval task

eDiscovery Task adapted from TREC 2011 Legal Track Topic 401

The purpose of this task is to retrieve documents that match the below request for production. The company in this case is Enron. The company is a now defunct energy trading company that was the subject of a large body of litigation both civil and criminal.

The following is the request for production:

You are requested to produce all documents or communications that describe, discuss, refer to, report on, or relate to the design, development, operation, or marketing of enrononline, or any other online service offered, provided, or used by the Company (or any of its subsidiaries, predecessors, or successors-in-interest), for the purchase, sale, trading, or exchange of financial or other instruments or products, including but not limited to, derivative instruments, commodities, futures, and swaps.

Additional Guidance for Relevance:

The above request broadly seeks documents concerning Enron online, the Company’s general purpose trading system, or any other online financial or commodities services offered, provided, or used by the Company and its agents.

In this case attorney-client communication or otherwise privileged information is not an issue.

This request is seeking information specifically about an online system for trading financial instruments. A document is not relevant if it refers to the purchase, sale, trading, or exchange of a financial instrument or product, but does not involve the use of an online system.

A document is relevant if it describes, discusses, refers to, reports on, or relates to: the design, development, operation, or marketing of “enrononline,” or any other online services offered, provided or used. This includes, how the system was set up, how the system worked on a day-to-day basis, how the Company developed or modified the system, how the Company marketed or advertised the system, and the actual use of the system by the Company, its subsidiaries, predecessors, or successors in interest.
A relevant document can be for the purchase, sale, trading, or exchange of: financial instruments, financial products, including, derivative instruments, commodities, futures, or swaps. These instruments and products are distinguished from other goods and services by the fact that their value depends on future events and their purchase incurs financial risk.

A document is relevant even if it makes only implicit reference to these parameters. No particular transaction (i.e., purchase or sale) need be cited specifically. If the document generally references such activities, transactions, or a system whose function is to execute such transactions, and it otherwise meets the criteria, it is relevant.

Examples of responsive documents include: Correspondence, Policy statements, Press releases, Contact lists, or Enronline guest access emails.

Additional Guidance for Non-Relevance

Examples of non-relevant documents include: Purchase, sale, trading or exchange of products or services other than financial instruments or products, or any documents referring to employee stock options or stock purchase plans offered as incentives or compensation, or the exercise thereof. Also documents relating to structured finance deals or swaps that are specified explicitly by written contracts, even if the contracts themselves are electronic or electronically signed are not relevant. Also documents related to the use of online systems by Enron employees for their personal use are outside this request and are not relevant.
APPENDIX - D

Instructions for Participant Groups

Purpose of Study Group 1

The purpose of the discovery task is for you to identify key words that will result in returning the maximum number of documents from a collection to match a discovery request.

First you will be given a discovery request. After you have viewed the discovery request you will be asked to provide key words that you believe will return documents that match the request.

You will also be asked to provide elimination words that will filter out non-responsive documents. The goal you should use for providing elimination words is to reduce the number of non-relevant documents from being returned.

Purpose of Study Group 2

The purpose of the discovery task is for you to identify key words that will result in returning the maximum number of documents from a collection to match a discovery request.

After you have viewed the task you will be directed to a list of PDF documents. These documents are a representative sample of documents from the larger collection. You will be given a maximum of 15 minutes to review these documents. You may review as many of the documents as you like. The purpose of your review is to determine if you are able to draw conclusions about the nature of the collection and then apply your conclusions to your selection of search terms.

After you have reviewed the sample collection you will be asked to provide key words that you believe will return documents that match the request.

You will also be asked to provide elimination words that will filter out non-responsive documents. The goal you should use for providing elimination words is to reduce the number of non-relevant documents from being returned.
Purpose of Study Group 3

The purpose of the discovery task is for you to identify key words that will result in returning the maximum number of documents from a collection to match a discovery request.

After you have viewed the task you will be directed to a list of PDF documents. These documents are a representative sample of documents from the larger collection. You will be given a maximum of 30 minutes to review these documents. You may review as many of the documents as you like. The purpose of your review is to determine if you are able to draw conclusions about the nature of the collection and then apply your conclusions to your selection of search terms.

After you have reviewed the sample collection you will be asked to provide key words that you believe will return documents that match the request.

You will also be asked to provide elimination words that will filter out non-responsive documents. The goal you should use for providing elimination words is to reduce the number of non-relevant documents from being returned.

Purpose of Study Group 4

The purpose of the discovery task is for you to identify key words that will result in returning the maximum number of documents from a collection to match a discovery request.

After you have viewed the task you will be directed to a list of PDF documents. These documents are a representative sample of documents from the larger collection. You will be given a maximum of 45 minutes to review these documents. You may review as many of the documents as you like. The purpose of your review is to determine if you are able to draw conclusions about the nature of the collection and then apply your conclusions to your selection of search terms.

After you have reviewed the sample collection you will be asked to provide key words that you believe will return documents that match the request.

You will also be asked to provide elimination words that will filter out non-responsive documents. The goal you should use for providing elimination words is to reduce the number of non-relevant documents from being returned.
APPENDIX - E

Exploration Study Computer Print Screens

Figure 20: eDiscovery Study User Informed Consent Screen
Figure 21: eDiscovery Study User Group Selection Screen

Figure 22: eDiscovery Study User Group 1 Instruction Screen
**Figure 23:** eDiscovery Study User Group 2 Instruction Screen

**Figure 24:** eDiscovery Study User Group 3 Instruction Screen
**Figure 25:** eDiscovery Study User Group 4 Instruction Screen

**Figure 26:** eDiscovery Study User Demographics Screen
The purpose of this task is to retrieve documents that match the below request for production. The company in this case is Enron. The company is a now defunct energy trading company that was the subject of a large body of litigation both civil and criminal.

The following is the request for production:

You are requested to produce all documents or communications that describe, discuss, refer to, report on, or relate to the design, development, operation, or marketing of e-rononline, or any other online service offered, provided, or used by the Company (or any of its subsidiaries, predecessors, or successors in interest), including but not limited to, derivative instruments, commodities, futures, and swaps.

Additional Guidance for Relevance:

The above request broadly seeks documents concerning Enron online, the Company’s general purpose trading system, or any other online financial or commodities services offered, provided, or used by the Company and its agents.

In this case you attorney-client communications or otherwise privileged information is not an issue.

This request is seeking information specifically about an online system for trading financial instruments. A document is not relevant if it refers to the purchase, sale, trading, or exchange of a financial instrument or product, but does not involve the use of an online system.

A document is relevant if it describes, discusses, refers to, reports on, or relates to: the design, development, operation, or marketing of e-rononline, or any other online services offered, provided or used. This includes, how the system was set up, how the system worked on a day-to-day basis, how the Company developed or modified the system, how the Company marketed or advertised the system, and the actual use of the system by the Company, its subsidiaries, predecessors, or successors in interest.

A relevant document can be for the purchase, sale, trading, or exchange of financial instruments, financial products, including, derivative instruments, commodities, futures, or swaps. These instruments and products are distinguished from other goods and services by the fact that their value depends on future events and their purchase occurs financially.

A document is relevant even if it makes only implicit reference to these parameters. No particular transaction (i.e. purchase or sale) need be cited specifically if the document generally references such activities, transactions, or a system whose function is to execute such transactions, and it otherwise meets the criteria, it is relevant.

Examples of Responsive documents include: Correspondence, Policy statements, Press releases, Contact lists, and Enronline guest access emails.

Additional Guidance for Non-Relevance:

Examples of non-relevant documents include: Purchase, sale, trading or exchange of products or services other than financial instruments or products, or any documents referring to employees, stock options, or stock whose basis offered as transactions or compensation, or to the exercise thereof. Also documents relating to structural finance deals or assets that are

Figure 27: eDiscovery Study Retrieval Task Screen

Figure 28: Exploration Screen with Count-Down Timer
Figure 29: Recall Terms User Input Screen

Figure 30: Elimination Terms User Input Screen
APPENDIX - F

Learning/Relevance Feedback Computer Interface Screens

Figure 31: eDiscovery Learning Tool Search Term Selection Screen
Figure 32: eDiscovery Learning Document Retrieval Interface Screen
APPENDIX – G

Pre-Task Questionnaire for User Understanding of Request

Pre-Task Strategy Questionnaire

1. Summarize in one or two sentences what the request is seeking?

2. What concepts do you believe define the documents that satisfy the request?

3. What order of steps will you use to formulate a strategy to find and identify the documents to match the request? First I will… Next I will…
APPENDIX – H

Post-Task Exploration Questionnaire

Narrative Questions

1. When I conduct an information search, the type of information I expect to find is?

2. If I had to choose between being efficient or being thorough, I would choose __________.

3. When I conduct an information search, the format I expect the information to be found is in: Web page, Web Site, PDF, Email, Other?

4. When I find an information item, I evaluate it to determine if it meets my need by?

5. When conducting a specific search for documents, my search method differs from a search for web pages or web sites because?

6. When I select a document for review I focus on:

7. I search for documents contained within a collection of documents to meet my information need by doing the following:

8. I use the following criteria to evaluate whether a document meets my information need:

9. When I search for documents within a collection of documents, I define/determine what I am looking for by?
10. When viewing a document in a collection, the items I focus upon within that document that help me determine if that document meets my requirement (information need) are?

Scaled Agree/Disagree Questions  (-3 to +3)

1. When I search for information, I am most concerned with being efficient.

2. When I search for information, my first/primary method of sorting between documents that meet my need and documents that do not meet my need is to scan the titles of documents.

3. When I search for information, my ONLY method of sorting between documents that meet my need and documents that do not meet my need is to scan the titles of documents.

4. When I select a document I almost always review the entire document.

5. When I search for information, I prefer to skim (quick review of a portion of the contents) the documents whose titles seem to meet my information need.

6. My only method of sorting is to scan titles.

7. When I search for information, I am most concerned with being thorough.

8. When I search for information, I prefer to scrutinize (review entire content) the documents whose titles seem to meet my information need.

9. My first/immediate method of sorting is to scan titles.

10. I use titles to base my selection of documents.

11. When I select a document for further review I rarely need to go beyond the first paragraph before deciding that it does or does not meet my need.

12. When I select a document I rarely review the entire document.
Scaled Agree/Disagree Questions (-3 to +3)

When I search for documents:

1. I limit the depth of my exploration to scanning of titles of documents alone.

2. I scan titles and then skim selected documents based on the content of the titles.

3. I select documents based on titles, but I also randomly select documents for a broad exploration of the collection.

When I select a document:

1. I prefer to limit my review to the first paragraph of the document.

2. I prefer to skim the entire document to get a general understanding of the content.

3. I prefer to scrutinize the entire document to get an in depth understanding of the content.
APPENDIX – I

System Usability Study

Usability Questionnaire

1. Overall, what is your impression of the system?

2. How did you like the way system performed for you?

3. What did you like best about the system?

4. Was the system useful in helping you complete the information retrieval task?

5. On a scale of 1 to 10, how likely are you to use a system like this for an information retrieval task?

6. What ways did the system fail to meet your needs for the task?

7. What are the additional features you would like to have on this system to make your task easier to complete?
## APPENDIX – J

**Table 28**: Covariates, Levels, and Descriptions

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Litigation Experience (LIT)</td>
<td>0</td>
<td>No litigation experience.</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Less than 1 year.</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1 to 2 years.</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>More than 2 years.</td>
</tr>
<tr>
<td>Financial Experience</td>
<td>0</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Some knowledge of terms</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Significant knowledge of terms, amateur level investing.</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Expert knowledge, professional level investing.</td>
</tr>
<tr>
<td>Corpus Knowledge</td>
<td>0</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Have seen items from corpus previously.</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Done previous exploration with this corpus.</td>
</tr>
</tbody>
</table>
### APPENDIX – K

**Table 29:** Independent and Dependent Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>DV / IV</th>
<th>Scale/Measure</th>
<th>Range Low, High</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>DV</td>
<td>.1 – 1.0 Ratio of Documents</td>
<td>.2, .73</td>
<td>.51</td>
<td>.15</td>
</tr>
<tr>
<td>Precision</td>
<td>DV</td>
<td>.1 – 1.0 Ratio of Documents</td>
<td>.43, .81</td>
<td>.62</td>
<td>.11</td>
</tr>
<tr>
<td>Total Time Explored</td>
<td>IV</td>
<td>Minutes 1 - 45</td>
<td>10, 45</td>
<td>28</td>
<td>12.47</td>
</tr>
<tr>
<td>Total Number of Documents Viewed</td>
<td>IV</td>
<td>Ordinal Number of Items</td>
<td>15, 120</td>
<td>44</td>
<td>28.34</td>
</tr>
<tr>
<td>Time Spent Per Document</td>
<td>IV</td>
<td>Minutes 1 - 45</td>
<td>.2, 1.3</td>
<td>.57</td>
<td>.28</td>
</tr>
<tr>
<td>Tolerance for Ambiguity</td>
<td>IV</td>
<td>Ordinal Scored Questionnaire</td>
<td>10, 48</td>
<td>30.28</td>
<td>10.9</td>
</tr>
<tr>
<td>Locus of Control</td>
<td>IV</td>
<td>Ordinal Scored Questionnaire</td>
<td>-12, 14</td>
<td>1.83</td>
<td>8.05</td>
</tr>
<tr>
<td>Disposition Toward Innovation</td>
<td>IV</td>
<td>Ordinal Scored Questionnaire</td>
<td>18, 35</td>
<td>26.97</td>
<td>5.26</td>
</tr>
<tr>
<td>Personal Innovation toward Information Technology</td>
<td>IV</td>
<td>Ordinal Scored Questionnaire</td>
<td>-20, 25</td>
<td>6.33</td>
<td>12.13</td>
</tr>
</tbody>
</table>
## APPENDIX – L

### Table 30: Pearson Coefficient Correlations

<table>
<thead>
<tr>
<th></th>
<th>PIIT</th>
<th>LOC</th>
<th>TOA</th>
<th>DISPO</th>
<th>TOTALTIM</th>
<th>PERDOCTI</th>
<th>TOTALDOC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PIIT</strong></td>
<td>1.00000</td>
<td>-0.89706</td>
<td>-0.00623</td>
<td>-0.12841</td>
<td>0.81612</td>
<td>-0.33136</td>
<td>0.69643</td>
</tr>
<tr>
<td><strong>PIIT</strong></td>
<td>&lt;.0001</td>
<td>0.9623</td>
<td>0.3282</td>
<td>&lt;.0001</td>
<td>0.0097</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td><strong>LOC</strong></td>
<td>-0.89706</td>
<td>1.00000</td>
<td>-0.22654</td>
<td>-0.07217</td>
<td>-0.89525</td>
<td>0.49924</td>
<td>-0.85054</td>
</tr>
<tr>
<td><strong>LOC</strong></td>
<td>&lt;.0001</td>
<td>0.0818</td>
<td>0.5837</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td><strong>TOA</strong></td>
<td>-0.00623</td>
<td>-0.22654</td>
<td>1.00000</td>
<td>0.91590</td>
<td>0.11939</td>
<td>-0.28352</td>
<td>0.39666</td>
</tr>
<tr>
<td><strong>TOA</strong></td>
<td>0.9623</td>
<td>0.0818</td>
<td>&lt;.0001</td>
<td>0.3635</td>
<td>0.0281</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td><strong>DISPO</strong></td>
<td>-0.12841</td>
<td>-0.07217</td>
<td>0.91590</td>
<td>1.00000</td>
<td>-0.01952</td>
<td>-0.21921</td>
<td>0.27970</td>
</tr>
<tr>
<td><strong>DISPO</strong></td>
<td>0.3282</td>
<td>0.5837</td>
<td>&lt;.0001</td>
<td>0.8823</td>
<td>0.0924</td>
<td>0.0304</td>
<td></td>
</tr>
<tr>
<td><strong>TOTALTIME</strong></td>
<td>0.81612</td>
<td>-0.89525</td>
<td>0.11939</td>
<td>-0.01952</td>
<td>1.00000</td>
<td>-0.34633</td>
<td>0.88212</td>
</tr>
<tr>
<td><strong>TOTALTIME</strong></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.3635</td>
<td>0.8823</td>
<td>0.0067</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td><strong>PERDOCTI</strong></td>
<td>-0.33136</td>
<td>0.49924</td>
<td>-0.28352</td>
<td>-0.21921</td>
<td>-0.34633</td>
<td>1.00000</td>
<td>-0.56832</td>
</tr>
<tr>
<td><strong>PERDOCTI</strong></td>
<td>0.0097</td>
<td>&lt;.0001</td>
<td>0.0281</td>
<td>0.0924</td>
<td>0.0067</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td><strong>TOTALDOCS</strong></td>
<td>0.69643</td>
<td>-0.85054</td>
<td>0.39666</td>
<td>0.27970</td>
<td>0.88212</td>
<td>-0.56832</td>
<td>1.00000</td>
</tr>
<tr>
<td><strong>TOTALDOCS</strong></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.0017</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
</tbody>
</table>