International Journal of Information Science 2016, 6(1): 1-19
DOI: 10.5923/j.ijis.20160601.01

How Does Exploration Impact IR Performance in Large Document Collections?

Harvey Hyman[1,*], **Rick Will**[2], **Terry Sincich**[2], **Warren Fridy III**[3]

[1]Library of Congress, National Library Service, Washington, DC, USA
[2]Information Systems/ Decision Sciences, University of South Florida, Tampa, Florida, USA
[3]H_2 & WF_3 Research, LLC., Tampa, Florida, USA

Abstract A significant problem for IR researchers is how to efficiently handle large collections of electronic documents. Manual review is time consuming and expensive. Automated methods can be imprecise and fail to yield relevant documents in the retrieval set. This problem is receiving significant attention in two major domains: the legal community, with the increase in search of electronic documents in litigation (eDiscovery), and in the medical community, with the increase of mandates for ehealth systems such as electronic patient records (EMR and EHR) and health informatics. This paper examines how the construct of exploration may be implemented as a methodology to improve user performance when searching and sorting through large electronic document collections by facilitating context and content understanding through multiple iterations. The study reported in this paper examines the research questions of: How does exploration impact IR performance, and how can exploration be implemented to achieve improvement in IR results? The study examines the correlation between an individual user's exploration of iterated sample selections from a large corpus of electronic documents and the individual's IR performance. Our findings support that: (1) IR performance can be manipulated by using an exploration method, (2) Time spent exploring a collection is correlated with performance in both recall and precision, and (3) Number of documents viewed in a collection is correlated with performance in precision.

Keywords Information Retrieval, Exploration, IR Performance, Sorting, Knowledge Acquisition, Large Document Collections

1. Introduction and Background of Research Question and Designed Experiment

Intuitively one would think that the relationship between exploration and performance is positive and direct – the more discoveries a user uncovers from items in a corpus collection, the greater the performance in the IR result. But what is it about exploration that affects IR, and how does exploration work exactly? Are certain factors more significant than others, and if so, how can they be measured against correlated performance metrics such as recall and precision? For instance, how much time does a user need to spend searching and sorting through a collection to make a difference? How many documents need to be reviewed before an effect can be detected in order to observe a measureable or meaningful difference in the user's performance? At what point does further exploration of the collection no longer support improvement in performance?

This study examines these questions with the goal of developing insights that might be used to create a predictive model to explain the correlation between exploration and IR performance. We conduct an experiment involving 120 users to evaluate differences in IR performance of individuals using three independent variable measures for the behavioral construct of exploration and two dependent variable measures (Recall and Precision) for performance.

The behavioral construct of exploration is measured using three independent variables: *Total Time Exploring, Per Document Time Exploring, and Total Number of Documents Explored*. These independent variables are used to measure operational IR tasks of: The aggregate time spent exploring a collection, the number of documents explored in a collection, and the average time spent exploring (viewing) a single document.

The experiment described herein is designed to support the study of how users interact with a collection of documents, and how that interaction can be manipulated to lead to predictable differences in performance, as measured in recall (percentage of potential relevant documents actually recovered), and precision (percentage of relevant documents in the result set). The designed experiment evaluates the exploration method described in this study against two base-line methods for comparison.

The first base-line method is a *random extraction* of

[*] Corresponding author:
hymanphd@gmail.com (Harvey Hyman)

Published online at http://journal.sapub.org/ijis

documents from the corpus equal to the average number of documents in the participants' final IR extractions. The purpose of this base-line is to validate whether human in the loop fairs any better in performance over a simple random extraction of documents taken from the collection.

The second base-line is an extraction of documents using a technique we call "verbatim." This method uses the specific words from the IR task itself (typically in the form of a narrative). Think of this as a request narrative such as "Find me all documents that have to do with…a particular subject." The verbatim method injects the non-function words from the narrative and performs a simple "bag of words" matching against all documents in the collection. The purpose of this base-line method is to validate whether human in the loop supported by exploration fairs any better than using the simple plain language of the IR request by itself (with no user involvement).

2. Motivation and Practical Application

In the end, all documents are settled by human inspection – whether conducting a personal web search, or exploring a complex bounded collection of documents, automated retrieval only gets the user so far – it is the user who must review each document to gain the knowledge sought. So, how should a user spend their time and energy? How much of the search and sorting process should be left up to the machine and how much must be human in the loop?

In this study, we explore the behavioral side of IR in so far as we seek to explain how the user's performance in IR can be improved by leveraging the natural human activity of exploration. All information retrieval activities ultimately lead to manual human review. After all, that is the purpose of IR, to provide the user with a selection of documents meeting the user's criteria, from a larger collection. There is a significant financial and management interest in reducing the number of non-relevant documents to be reviewed. Reducing the number can result in significant time savings to the document reviewer and cost savings to the entity paying for that review.

This study addresses the objective cost savings in human review by a process for user exploration to reduce the search space, improve sorting, and produce a more effective and efficient document retrieval set. The goal is to improve the retrieval result developed by user exploration using small iterated samples from the full corpus.

The desired effect is to reduce the search space such that fewer documents are needed for human review. The study uses three measures for exploration as a predictor of IR results. The data collected are recorded using an interface browser developed for the study and evaluated for its utility based on common performance measures recall and precision.

The research conducted by this study is intended to provide insight into two areas: The first is the relationship between recall and precision previously validated in the literature but never explained; the second is how the identified exploration variables for measuring number of documents and amount of time can be used to predict productivity in IR results.

3. Organization of this Paper

This paper is organized in three parts. Part One is our introduction and background presented above. Part Two is our literature review. Part III is the narrative and description of our design and experiment conducted.

The work reported in this paper is an extension of a series of experiments studying performance and processes in information retrieval tasks. The background and underlying technology of this work leading to the research model described here, builds upon foundations previously established and validated in Hyman et al., 2015, and adapted for the study reported in this paper. We spend the next several sections of this paper (Part Two) providing a narrative of the concepts and techniques we considered in developing our approach.

4. Review of Literature and Approaches

This portion of the paper narrates the foundations, concepts and techniques we considered and that influenced the design for our research model and the series of experiments conducted.

4.1. Brief Review of 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 (van Rijsbergen, 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 study 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.

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).

4.2. The Context of Relevance in this Study

"Relevance is a subjective notion" (van Rijsbergen, 1979). In this study 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. Quite often judgments fall to a panel of experts in the domain for determining relevance of a document (van Rijsbergen, 1979; *TREC Proceedings 2009, 2010, 2011*). This is quite different from general broad-based IR – where a single user seeks documents on a subject matter, and the user determines whether the retrieval satisfies his/her information need.

4.3. Brief Background on Electronic Document Collections

Electronic documents refer to information created, manipulated, communicated, or stored in digital form requiring the use of computer hardware and/or software (*Electronically Stored Information: The December 2006 Amendments to the Federal Rules of Civil Procedure*, Kenneth J. Withers, Northwestern Journal of Technology and Intellectual Property, Vol.4 (2), 171).

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 significant volume of electronic documents 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 files) and complexity (varied types of information files) in document retrieval.

4.4. Brief Review of IR from Knowledge Based Approaches

Document representation has been identified as a key component in IR (van Rijsbergen, 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 and more frequent terms revealing noise (Grossman and Frieder, 1998).

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.

In this paper we seek to evaluate knowledge and understanding of content and context as document representation, and the use of exploration to acquire and then leverage that knowledge and understanding about the collection to making better choices for sorting and selecting the IR set.

4.5. Common Approaches for Reducing the Search Space

Methods and techniques for reducing the search space have been the subject of study going back as far as the 1950s (Singhal, 2001). They include distributional structure of documents (Harris, 1954), indexing words in documents and using their occurrences as criteria for relevance (Luhn, 1957), and early attempts to use probabilistic methods to rank relevancy of documents based on estimations (Maron and Kuhns, 1960).

Probabilistic based IR saw a significant increase in research and use in the 1980s and added conditional candidacy of documents. Cluster methods have also 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.

4.6. Techniques for Modeling and Performance Previously Considered

Research suggests that using fewer terms is more effective and will produce better recall and precision (Grossman and Frieder, 1998). 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 study reported in this paper.

Prior research has shown that 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).

Another useful method we consider is *stemming* 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.

Our research has found that 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.'

4.7. Background on Exploration and IR

The research conducted here is focused on 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: *Exploration* focuses on these two conditions. These conditions of IR go largely unexploited by users. Our work has been a series of studies addressing the gap between brute force, trial and error techniques, and test collection reviews presently employed by IR practitioners. The study reported in this paper seeks to explain how exploration can be used most effectively to improve IR performance.

Most behaviorists would agree that 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 (Hyman et al., 2015). We see common and frequent examples every day 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.

Our previous work has found that 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 natural human 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.

We draw upon the goals of exploration from the point of view of the human instinct as a means for *making sense* of our environment and satisfying curiosity are represented in the problem domain of information retrieval. (Debowski et al., 2001) view exploratory search as a "screening process," and state that exploration identifies items "to become the focus of attention." This suggests that exploration may be able to be leveraged to improve searching and sorting.

Our research examined prior reports of strategies that users formulate for web searches have examined the types of knowledge and strategies involved in web-based information seeking (Holschler and Strube, 2000). Prior findings include 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 and that the information space as "diverse and often poorly organized content" (Holschler and Strube, 2000). This use case scenario contrasts with the bounded space which is typically organized around the subject matter in question. Holschler and Strube's finding that experts can outperform less experienced users is adapted to this study by evaluating whether knowledge acquired by exploration can improve a user's ability to tackle the search problem.

Muramatsu and Pratt, 2001, proposed 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."

Catledge and Pitkow 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).

We are influenced by Broder (2002), who 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. Our research studies the user's informational need, and also seeks to explain his/her navigational behavior that may affect the IR result produced.

Our work in this area has also been influenced by Muylle et al., 2010, who 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. Our previous research reported in Hyman et al., 2015, adapted these constructs to measure scanning, skimming and scrutinizing behavior by the users conducting. We extend that work in this study of the exploration construct and its ability to influence IR performance.

We consider the work of Navarro-Prieto et al., 1999. They 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.

4.8. Considerations for Generating Search Terms

We take the view that search terms are approximations for the user's mental model for describing the attributes of relevance for a document (Salton and Buckley, 1988). In many automated IR systems, we see the use of weighting of terms to enhance the effectiveness of the selected attributes to describe a relevant document. This is often modeled by using statistical occurrences and term frequencies. We find this technique in common indexing methods going back to the fundamental foundations of indexing and IR performance (Luhn, 1957; Spark-Jones, 1971).

The research model in this paper seeks to address the major limitation associated with term frequency: the difficulty of distinguishing between the frequency of occurrence in the relevant documents and the frequency of occurrence in the entire collection – often referred to as "context dependency." We rely on an implied assumption that search terms *may* be known *a priori* or may surface as a result of patterns discovered during exploration of the collection, hence the use of exploration as a discovery methodology.

This leads to the main hypothesis in this study that exploration of the collection will provide the IR user with the additional insight to be able to better describe the document he/she is seeking, and therefore select better search terms.

A variation on this hypothesis 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.

The research and findings we have discovered during the course of this exploration study have found: (1) search and sorting experience matters, (2) subject matter experience matters, and (3) experience affects the complexity of search strategy and choice of search terms. Our research described here investigates these findings further by evaluating how users can improve their knowledge and understanding of a corpus and its contents through exploration, and leverage that knowledge and understanding through an automated tool to improve IR results.

5. The Study Itself

This portion of the paper narrates the study we conducted, the design, data collection, analysis and results produced.

5.1. Prototype System Built to Support the Study

We developed an original system architecture depicted in Figure 1, to support our exploration study. The system consists of a user interface to facilitate the exploration of a large collection of documents. We use a browser screen to present the user with small, manageable sample sets extracted from the large collection, based on the user's exploratory queries. As a side note, we found that the optimal number to present is 10 documents per set. We use the system to manage the user-system interactions and record the sessions.

The system works by having users explore small samples of a large collection and submit selection criteria based on the conclusions drawn from the explorations. It produces the sample sets from the full collection of documents. The hypothesis tested here is that *user exploration will produce better understanding of the nature and context of the collection and therefore, better decisions can be made for selecting search criteria*. The browser interface allows the user to choose search terms and then explore the resulting sample set retrieved.

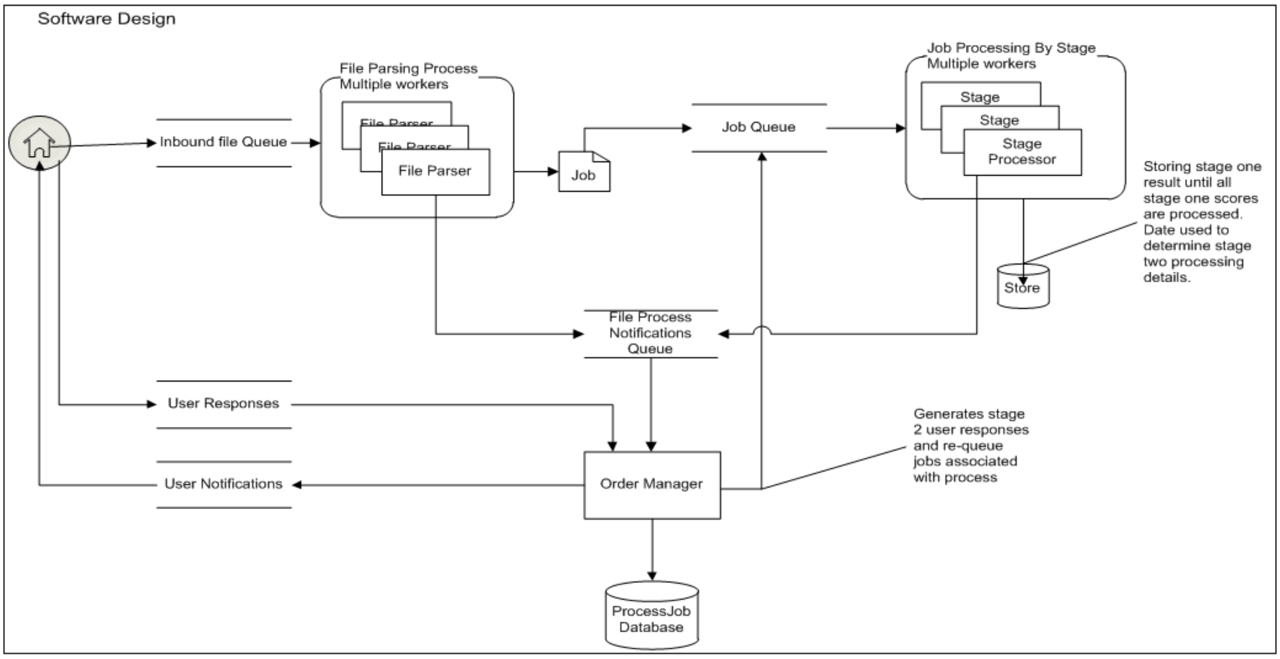

Figure 1. Architectural Design of Exploration System

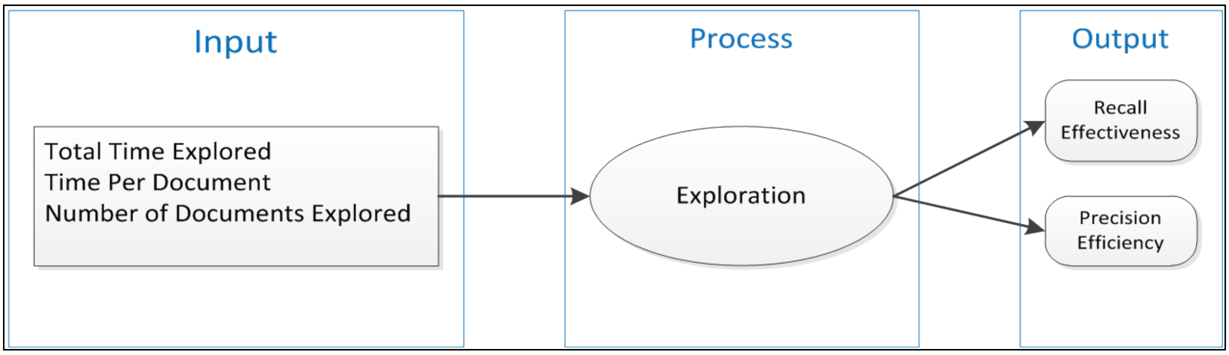

Figure 2. Research Model

5.2. IPO Research Model

Our research model here seeks to describe and explain the relationship of exploration and the IR process. The value proposition here is that a user will gain *knowledge* and *understanding* about the content of the overall collection and the context of the documents contained within, supporting user IR goals of: (1) seeking to understand the nature of the subject matter of interest, (2) context of the corpus containing the documents of interest, and (3) structure of the relevant documents sought for extraction. The objectives supporting the IR goals are: (1) learn about the context, and (2) organization of subject documents contained within the collection.

The user applies what he/she has learned during exploration about the subject matter, corpus and documents to produce the greatest number of relevant documents (recall) and fewest number of non-relevant documents (precision).

The system designed for this study facilitates user exploration of the collection. We model the system using an IPO (input-process-output) approach, whereby the inputs are our independent variables (IVs) and our outputs are our dependent variables (DVs); exploration is modeled as the process facilitating the IR result.

The model for the exploration construct 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*. The output variables of effectiveness and efficiency are operationalized using recall and precision.

5.3. Population, Frame and Task Selection

The population of interest in this research is comprised of end-users engaged in searching and sorting of large electronic document collections. The study focuses on exploration as an IR method to enhance user performance in returning a retrieval set of relevant documents to fulfill an IR task request.

Our research here is motivated to learn more about the IR user who *does not* have an *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.

The study conducted to support this research tests whether a significant relationship exists between exploration and IR user performance results. An assumption here is that the user acquires knowledge and understanding about the nature and content of the document collection through repeated iterative samples, selected by an automated system based on user search terms. The proposition is that the knowledge and understanding acquired from exploration of small sample sets can be successfully leveraged for insight about content and context of the data collection as a whole, and thereby result in improved IR performance as measured by *recall* and *precision.*

The task is divided into two parts. The first task is to have the user provide an initial set of search terms for the system to generate a sample of documents from the large collection. The user's terms are based upon what they have learned from their initial exploration of the system selected set. The second task is for the user to provide feedback to the system in the form of relevance judgments for the document set retrieved. This process is iterated over a set of cycles. The selected sets are updated for each iteration based on the user's feedback. The user declares relevance and non-relevance to the system by use of radio buttons along with additional search terms using text boxes. The terms are absorbed by the system and a new selection of documents is presented to the user for judgment.

5.4. Document Corpus Used

The document corpus 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 where sorting is an important criterion. The collection is a corpus of emails formatted in the PST file type. The collection is robust enough to provide a reasonable approximation of the *complexity* problem in sorting, given the email files within 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 *volume* problem in sorting.

5.5. Why Study a Bounded Corpus?

Exploration of a bounded collection 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 this task 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. As such, 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 motivates the main research question of *whether a user can acquire knowledge about context and content through exploration of the bounded corpus and apply that knowledge to make better sorting decisions, resulting in improved IR performance*. The hypothesis is that exploration produces insight, insight produces knowledge, and knowledge leads to better sorting decisions. An underlying assumption is that more knowledge on a topic will produce better IR results than less knowledge on a topic.

Why is the exploration of a bounded corpus an important phenomenon to study? Well, 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 sort through and extract relevant documents from large collections of similar items will continue to be a value proposition in terms of time and cost (end-user resources).

5.6. Method of Study

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.

The task, treatment and data collection are conducted via the prototype system developed for this study. The system is housed on a server and accessed by the participants using a URL link from self-provided laptop computers.

Participants are assigned an IR task. Informed consent, task instruction and data collection instrument are presented to the users through the interface (U/I) computer screens.

All participants are given the same task. The task has been adapted from the TREC Legal Track 2011 Conference Problem Set #401 (TREC, 2011). This task is robust enough to provide necessary complexity in the IR request, is well matched to the data set, and has been previously validated in the literature (prior TREC Conferences). The *exploration* independent and dependent variables are listed in Table 1.

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

Linear regression analysis is used to measure the following relationships: Correlations of independent variables with dependent variable *Recall*; Correlations of independent variables with dependent variable *Precision*; Possible interactive effect of independent variables upon the dependent variables. The results are reported in the analysis and results sections of this paper.

5.7. Dataset Adapted to this Study

The dataset in this study is a subset of the Enron Collection. The subset that we use to measure the exploration effect is a collection of 10,000 randomly selected documents from the full corpus. 1,000 of the documents have been selected from the previously validated set marked relevant, and 9,000 documents have been selected from the previously validated set marked not relevant.

This allows us to make certain assumptions. The first assumption is that a random extraction (base-line comparison) 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 the normalized relevance judgements from the validated documents.

5.8. Processing and Preparation

The objects in the collection which consist of emails and attachments first needed to be prepared in such a way that the text could be read by the engines of the system. Considerable time went into this preparation. 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 we use for our exploration study is depicted in Figure 2.

The design of the system has 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. 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 a small document sample to get them started.

5.9. Manipulation of Time as the Treatment

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 to avoid a tight clustering and thereby increase the explanatory power of the variables.

The participants supply their query terms through the interface screen. The user selections are logged per user and submitted to the job queue for processing as depicted in Figure 2.

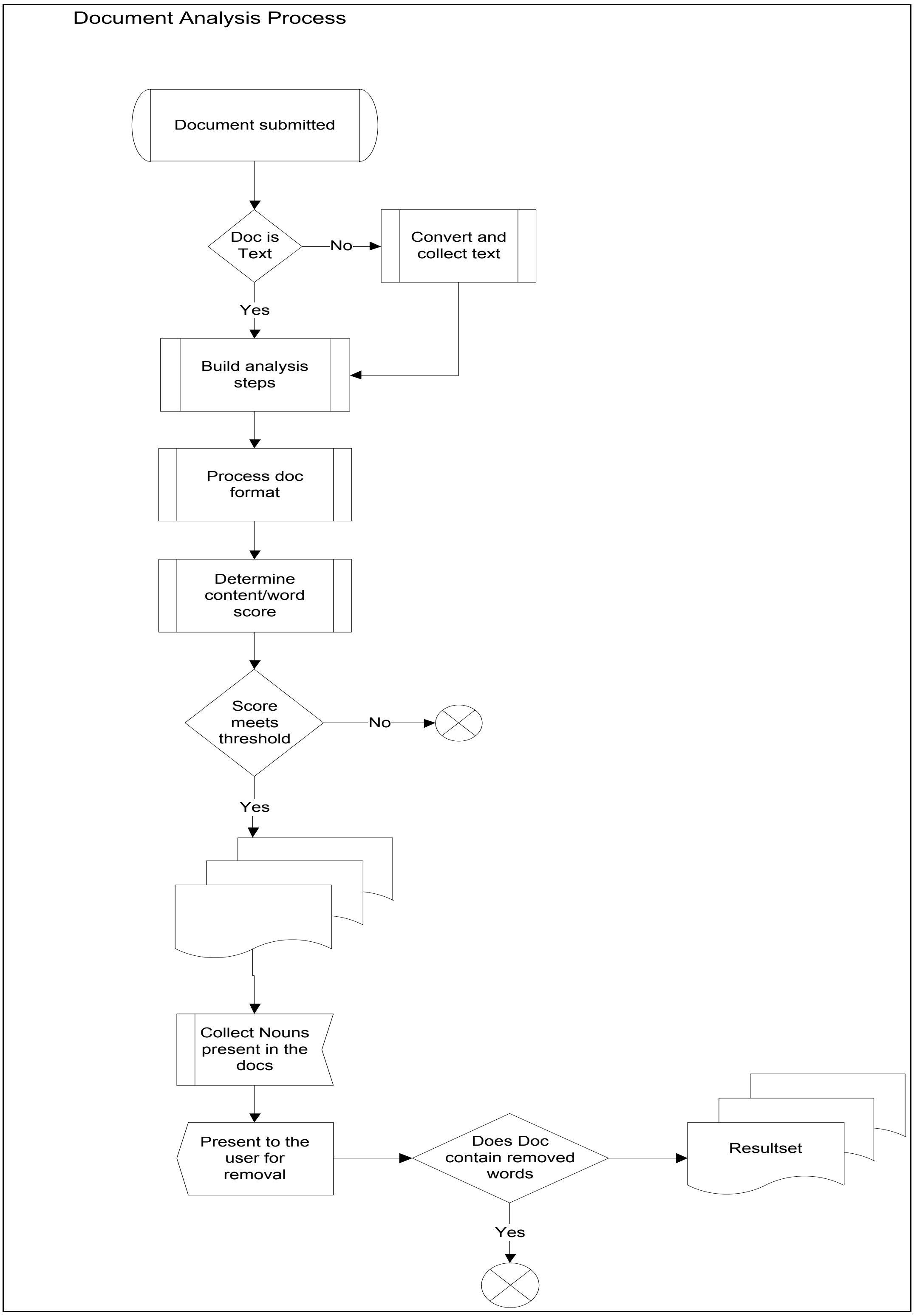
Document Analysis Process
Document submitted
Doc is Text
No
Convert and collect text
Yes
Build analysis steps
Process doc format
Determine content/word score
Score meets threshold
No
Yes
Collect Nouns present in the docs
Present to the user for removal
Does Doc contain removed words
Resultset
Yes

Figure 3. Job Run Process (Hyman et al., 2015)

5.10. Initial Pilot Studies

An initial pilot was conducted using 10 volunteers for the study with the purpose of receiving feedback regarding presentation, clarity and ease of use of the interface. The current version of the system used for the full study incorporates the feedback received from the first pilot. A second pilot was conducted with 24 participants divided into two groups, a control group which was given no time to explore and a treatment group given up to 60 minutes to explore a small sample of documents to get them started. 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.

This data was inconclusive on the issue of significant difference in recall or precision between the 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 made 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 had to do with the tight clustering of exploration time, making it difficult to draw conclusions about relationships of the variables. From that experience we designed the manipulation of time over a three group randomized assignment. An additional benefit of the pilot was that it confirmed the design, ease of use and quality issues of the system used for participant interaction and collection of our session data.

5.11. The Full Study

120 participants were randomly assigned to three groups manipulated to spread out time performance over a broader range to avoid clustering as occurred in the pilot. The individuals within each of the groups were 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. We hypothesized that participants would mentally anchor themselves if given a target time limit (as an interesting side note, the data seemed to support that notion).

The total time for exploration ranged from 10 minutes to 45 minutes. The participants' sessions have been recorded by a server hosting the prototype system.

Independent variables (IVs) representing *Total Time Exploring*, *Total Documents Viewed*, and *Time Spent per Document* have been assigned to track user interaction with the system as depicted in the research model in Figure 2.

5.11.1. Hypotheses Generated in this Study

The general proposition of this study is that an exploration method will outperform both random extraction and verbatim 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.

Hypotheses have been generated to study the effects of the exploration behavior independent variables 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 to test for interactive effects:

- H_0: *Total time exploring a corpus* affects *Recall* and *Precision* independent of *Number of documents viewed* and *Time per document.*
- H_a: *Total time exploring a corpus* affects *Recall* and *Precision* depending on *Number of documents viewed* or *Time per document.*
- H_0: *Number of documents viewed* affects *Recall* and *Precision* independent of *Total time exploring a corpus* and *Time per document.*
- H_a: *Number of documents viewed* affects *Recall* and *Precision* depending on *Total time exploring a corpus* or *Time per document.*
- H_0: *Time per document* affects *Recall* and *Precision* independent of *Total time exploring a corpus* and *Number of documents viewed.*
- H_a: *Time per document* affects *Recall and Precision* depending on *Total time exploring a corpus* or *Number of documents viewed.*

5.11.2. 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. A global F-test was used to evaluate the overall model and partial F-tests were used for testing interactive effects.

We conducted analysis upon the exploration research model comprised of the independent variables *Total Time Explored* (TTE), *Time per Document* (PER) and *Number of Documents Explored* (NUM), and the dependent variables recall and precision. The initial proposition is that individuals' IR performance can be predicted based on their exploration behavior measured by the independent variables. The proposition is represented by the hypotheses stated in the previous section and reduced to the equations for the exploration research model indicated below:

Main Effects Model:

$$DV_{Recall}, DV_{Precision} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + e$$

Full Model:

$$DV_{Recall}, DV_{Precision} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_1X_2 + B_5X_1X_3 + B_6X_2X_3 + e$$

Where:

X_1 = Total Time Explored,
X_2 = Time per Document,
X_3 = Number of Documents Explored.

5.12. Results

The average number of documents reviewed was 43, with a time per document average of 27.5 minutes total time and 58 seconds – just under one minute. 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 investigated in this research, measured by average *Recall* and average *Precision* based upon incremental units of the *Total Time Explored* variable, (2) A random extraction of a 503 document set, representing the average number of documents produced by the participants' in the study, (3) A raw extraction of documents based on the "verbatim" approach.

The graph of *Recall* against time appears in Figure 4. *Recall* performance for participants exploring the corpus for less than 15 minutes produced results in the .2 to .4 range, with an outlier at the 14-minute data point. Conclusions about a general trend within this time frame are difficult to draw.

Participants exploring the corpus in the 23 - 30-minute time frame produced results in the .45 to .5 range and followed a mostly flat trend line. There is a gap up from .2 at the 15-minute mark to .45 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 (.5 to .7) and no data point at 41 minutes. The recall performance results follow a fairly flat trend after an initial jump in performance occurring between 30 and 42 minutes. This warrants further study to determine if there is a diminishing return in a range beyond 42 minutes. A future experiment is planned to add a new time frame set for 60 minutes to investigate this relationship. We also plan to manipulate time using a forced, minute-by-minute user interface.

Exploration outperformed random extraction at every data point. However, the verbatim method was a better choice for users who spent less than 23 minutes and was competitive against exploration at several data points in the 23 – 30-minute range. This could suggest that for a user not committed to a minimum time effort for exploration, the better option is to not explore. This could also suggest that further study of exploiting the verbatim method is justified.

Exploration outperformed verbatim at all data points over 30 minutes; the effect also seems to flatten after an initial jump in this time range. This should be studied further to determine if the trend in effect remains flat, meaning that no further exploration yields improvement, or whether a time range beyond 45 minutes may continue to improve *Recall*. A future study is being designed to address this with a new sample of users and a minute-by-minute time treatment.

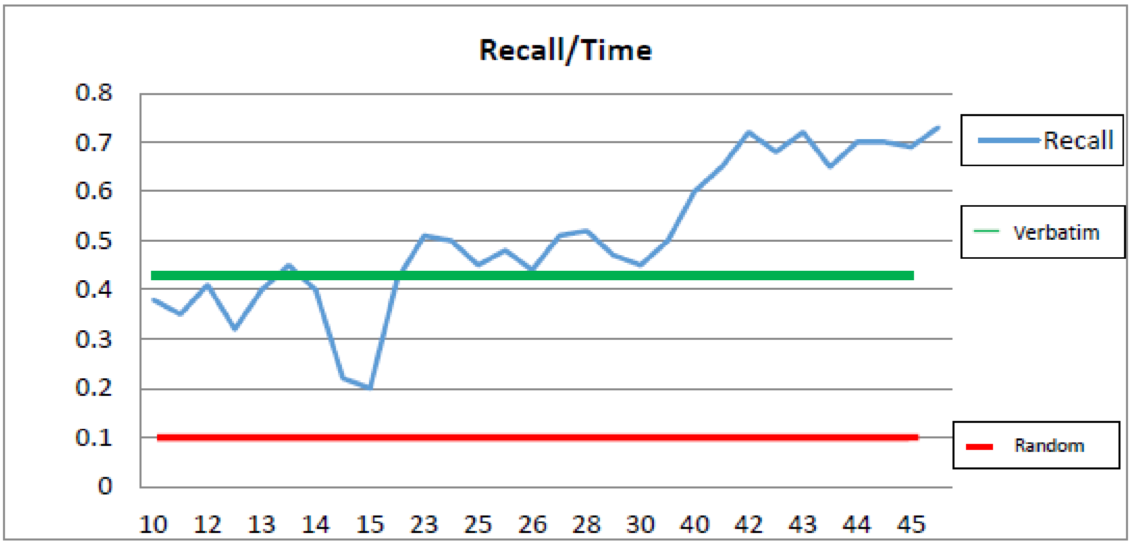

Figure 4. Recall over Time

The graph of precision against time appears in Figure 5. Precision performance follows a different trend than recall. Participants exploring the corpus on the shorter end of the timeline, 13 minutes or less, produced results in a range of .7 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. This is certainly confounding to say the least – how is it that less time produces a better result than more time? Perhaps less time is in fact better when it comes to precision? There are some plausible explanations that may explain this, but they are speculative at this time.

Participants exploring the corpus in the middle of the timeline, between 15 and 40 minutes, produced moderate results, over a wider range, .44 to .65 range. This is also a strange result given that we intuitively expected participants to improve positively and linearly as time increased, and in this time range performance followed a relatively flat trend. At this moment, we have no explanation for why this may be so, but this is yet another interesting phenomenon we discovered during the study. We plan to further study this effect and investigate whether in fact a middle time range exists that should be avoided by IR users, meaning the greatest effects are produced with short and long time exposures, but not medium.

The main results indicate that there is certainly an "exploration effect" and that it can be an effective method for producing improved precision IR results than random extraction or verbatim methods, and perhaps it may be most effective (results consistently above .60) when a user spends longer time periods (over 40 minutes indicated in this study). However, it is difficult to justify such a universal conclusion with just one study, even though this is a series. We still have an issue with data point gaps within the ranges analyzed. Therefore, a future study has been planned to investigate this effect with a new sample of users and a focus on filling in the current gaps in the observations.

5.12.1. Better than Chance?

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 120 participants, we performed 120 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.

5.12.2. Verbatim, When You Have No Idea Where to Begin

In situations where the IR 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 underlying thinking is that the terms in the request may in fact be significant indicators of relevant context (note the previous discussions about search terms as proxies for relevance attributes).

When we performed this type of extraction (verbatim) 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 verbatim in both measures.

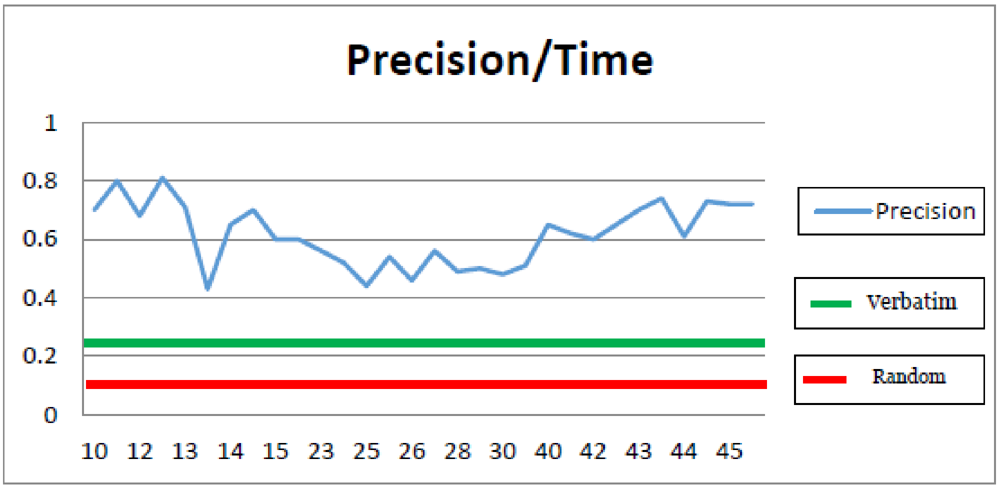

Figure 5. Precision over Time

5.12.3. Hypotheses Tested

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. The 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 also be that, after a certain amount of time is spent exploring, there is a significant leap in knowledge acquired about the corpus and the documents within it. The study suggests that the number may be as little as 40 minutes of exploration to achieve this leap.

The flatter (non-linear) 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, until greater than 40 minutes is invested. A list of the hypotheses with their measured variables and associated betas is listed in Table 8.

5.12.4. Statistical Analysis Models

The model was analyzed separately for *Recall* and for *Precision* using null and alternative hypotheses. Two tables summarizing the results for the exploration hypotheses are contained in the discussion section. A summary of results appears in Table 2 and Table 3.

The null and alternative hypotheses are as follows:

Recall	**Precision**
H_0: $B_1 = B_2 = B_3 = 0$	H_0: $B_1 = B_2 = B_3 = 0$
Ha: At least one Beta $\neq 0$	Ha: At least one Beta $\neq 0$

Where:

B_1 = Slope for Total time explored,
B_2 = Slope for Number of documents viewed,
B_3 = Slope for Time per document.

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*.

5.12.5. What About Interaction?

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 in Table 5 and Table 6.

Our analysis found no significant correlation between *Recall* and *Precision*. A printout of the Pearson Correlation appears in Table 7. 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 – tradeoff. 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 this is only one study of 120 users, but none the less, we have a confirmed observation here of this effect.

An initial indication here may suggest support for exploration as a method to influence the recall/precision trade-off. We believe that the results produced here certainly justify further study into the Recall-Precision relationship, especially if, in fact, precision can be enhanced without significant reduction in recall, through the application of an exploration methodology.

6. 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:

$$\text{Precision} = B_0 + B_1T + B_2N + B_3T*N$$

$$\text{Effect of Time on Precision} = (B_1 + B_3N)$$

$$\text{Effect of Documents on Precision} = (B_2 + B_3T)$$

Where:
T = Total Time Explored
N = Number of Documents Viewed

6.1. 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 sample size may have been too small to detect a result even though our N was 120. A more likely reason may be that we had gaps in our time line. We plan to address this by conducting additional studies with a more aggressive time manipulation treatment.

A second limitation in this study is the choice of document collection and the choice of IR task. Both of these items are narrow in their application. They are specifically related to IR tasks in the legal domain. As such, they may not serve as good approximations for IR generally – assuming there is such a data set and task that can be generalized for IR as a whole. None the less, we plan to expand our research by conducting our next studies using a medical oriented document collection and a medical IR task. This will expand the coverage of application of the exploration methodology. We plan to implement this design in our next study and compare the results to the study reported here.

6.2. Contribution

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

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 (IR) information retrieval. The results produced by our series of experiments supports the findings that user exploration of a small portion of a collection will yield improvement when implemented over various time interval ranges. There is a clear relationship established between time exploring and number of documents explored, and the corresponding IR results produced. Now, how much time and how many documents are needed for a minimum effect is the subject of future investigation.

The results that have been produced by this experiment indicate that there are ranges within which performance improves, ranges within which performance suffers, and ranges within which there is no effect (don't bother). The study also demonstrates how an exploration model approach to IR can improve performance, specifically when measured against random extraction and verbatim methods.

The 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.

7. Future Work and Conclusions

We are encouraged by the results produced in this study, particularly in the possible explanation for the recall-precision inverse relationship and in the exploration effect observed by differing retrieval results for ranges of time and number of documents. We plan to continue with additional series of experiments using alternative document collections to cross-validate the results produced here; our next data set will most likely involve a medical records database.

This study was designed to measure the significance of the relationship between exploration of iterated sample sets from a large collection and the corresponding IR result – how exploration impacts user performance. Conventional wisdom has suggested a direct and positive correlation between exploration and result. The 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 measured variables used in this study help explain user actions and strategies developed during corpus and document exploration and their significance upon IR performance – exploration variables *Total Time* invested in exploring, the *Number of Documents* viewed and *Time spent per document*.

Appendix of Tables

Table 1. List of Variables Tracked in the Study

Independent Variables	Dependent Variables
Exploration/Artifact Variables:	**Performance Measures:**
Number of documents viewed	Recall
Total Viewing Time	Precision
Viewing time per document	

Table 2. SAS 9.2 Printout for Recall Variables

```
                              The REG Procedure
                              Model: Exploration
                         Dependent Variable: RECALL
                     Number of Observations Read          120
                     Number of Observations Used          120

                             Analysis of Variance

                              Sum of         Mean
Source             DF        Squares        Square     F Value    Pr > F
Model               3        1.08166        0.36055     105.21    <.0001
Error              56        0.19191        0.00343
Corrected Total    59        1.27357

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

                             Parameter Estimates
                         Parameter     Standard
        Variable      DF   Estimate          Error          t Value     Pr > |t|
        Intercept      1     0.23835         0.03069          7.77       <.0001
        TOTAL TIME     1     0.00947         0.00142          6.69       <.0001
        PERDOCTIME     1    -0.02839         0.03653         -0.78       0.4404
        TOTDOCVIEW     1     0.00055612      0.00071032       0.78       0.4370
```

Table 3. SAS 9.2 Printout for Precision Variables

```
                              The REG Procedure
                              Model: Exploration
                        Dependent Variable: PRECISION

                     Number of Observations Read          120
                     Number of Observations Used          120

                             Analysis of Variance

                                Sum of          Mean
Source                  DF      Squares         Square    F Value    Pr > F
Model                    3      0.26712         0.08904     12.68    <.0001
Error                   56      0.39312         0.00702
Corrected Total         59      0.66024

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

                             Parameter Estimates

                               Parameter       Standard
Variable                  DF   Estimate        Error           t Value    Pr > |t|
Intercept                 1      0.66597       0.04392         15.16      <.0001
TOTAL TIME                1     -0.00975       0.00203         -4.81      <.0001
PER DOC TIME              1     -0.00128       0.05229         -0.02      0.9805
TOTAL DOCS VIEWED         1      0.00502       0.00102          4.94      <.0001
```

Table 4. Summary of Exploration Model Results

Independent Variables	Alpha	Dependent Variable
Total Time Exploring*	.01	Recall, Precision
Number Documents*	.01	Precision
Time per Document	Not Significant	

*- Interactive Effect upon Precision

Table 5. Results from SAS 9.2 printout for interactive effect upon Recall

```
                              The REG Procedure
                              Model: Exploration
                          Dependent Variable: RECALL
                      Number of Observations Read        120
                      Number of Observations Used        120
                             Analysis of Variance

                                 Sum of          Mean
Source               DF          Squares         Square       F Value     Pr > F
Model                 6          1.09217         0.18203       53.18       <.0001
Error                53          0.18140         0.00342
Corrected Total      59           1.27357

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

                             Parameter Estimates
                          Parameter        Standard
Variable          DF       Estimate           Error             t Value     Pr > |t|
Intercept         1          0.33478            0.12329           2.72        0.0089
TOTALTIM          1          0.00912            0.00404           2.26        0.0280
PERDOCTI          1         -0.04297            0.15645          -0.27        0.7847
TOTALDOC          1         -0.00474            0.00328          -1.45        0.1540
TTPD              1         -0.00175            0.00644          -0.27        0.7864
TTTD              1          0.00009655         0.00005987        1.61        0.1128
TDPD              1           0.00160            0.00300           0.53        0.5965

                   Test 1 Results for Dependent Variable RECALL
                                      Mean
              Source            DF        Square            F Value     Pr > F
              Numerator          3        0.00350            1.02         0.3897
              Denominator       53         0.00342
```

Table 6. Results from SAS 9.2 printout for interactive effect upon Precision

```
                              The REG Procedure
                             Model: Exploration
                       Dependent Variable: PRECISION

                    Number of Observations Read         120
                    Number of Observations Used         120

                            Analysis of Variance

                                Sum of          Mean
Source                DF        Squares         Square        F Value     Pr > F
Model                  6         0.38057        0.06343        12.02      <.0001
Error                 53         0.27967        0.00528
Corrected Total       59         0.66024

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

                            Parameter Estimates
                                Parameter       Standard
Variable              DF        Estimate        Error            t Value     Pr > |t|

Intercept              1        1.15779         0.15308           7.56       <.0001
TOTALTIM               1        -0.01673        0.00501          -3.34       0.0015
PERDOCTI               1        -0.33524        0.19426          -1.73       0.0902
TOTALDOC               1        -0.01290        0.00407          -3.17       0.0026
TTPD                   1        0.00460         0.00799           0.58       0.5675
TTTD                   1        0.00033831      0.00007434        4.55       <.0001
TDPD                   1        0.00467         0.00373           1.25       0.2157

                  Test 1 Results for Dependent Variable PRECISION

                                     Mean
          Source            DF       Square        F Value     Pr > F
          Numerator          3       0.03782       7.17        0.0004
          Denominator       53       0.00528
```

Table 7. Recall-Precision Correlation

```
                    Pearson Correlation Coefficients, N = 120

                         Prob > |r| under H0: Rho=0

                              RECALL        PRECISIO

          RECALL              1.00000        0.07847
          RECALL                             0.5512

          PRECISION           0.07847        1.00000
          PRECISION     0.5512
```

Table 8. List of Hypotheses Supported and Not

Hypothesis	Supported/Not	Variable	Alpha	Relationship Recall/Precision
H1a	Supported	TTE	.01	Recall: Direct and Pos
H1b	Supported	TTE*	.01	Precision: Direct and Neg*
H2a	Not	NUM		
H2b	Supported	NUM*	.01	Precision: Direct and Neg*
H3a	Not	PER		
H3b	Not	PER		

*- Interactive effect upon Precision supported

REFERENCES

[1] Baeza-Yates, R., Ribeiro-Neto, B., Modern Information Retrieval, ACM Press, New York. 1999.

[2] Bates, M., J., "Information Search Tactics," Journal of the American Society for Information Science, July, (1979).

[3] Barnett, S., A., A Study in Behavior. London: Methuen (1963).

[4] Berlyne, D., E., Conflict, Arousal and Curiosity, New York: McGraw Hill (1960).

[5] Berlyne, D., E., "Motivational Problems Raised by Exploratory and Epistemic Behavior," Psychology: A Study of Science, Vol. 5, pp. 284-364, New York: McGraw Hill (1963).

[6] Brisboa, N.R., Luances, M.R., Places, A., S., Seco, D., "Exploiting Geographic References of Documents in a Geographical Information Retrieval System Using an Ontology-Based Index," Geoinformatica, 14:307-331, (2010).

[7] Broder, A., "A Taxonomy of Web Search," IBM Research, SIGIR Forum, Vol. 36, No. 2, (Fall, 2002).

[8] Catledge, L., D., Pitkow, J., E., "Characterizing Browsing Strategies in the World-Wide Web," Computer Networks and ISDN Systems, Vol. 27, (1995).

[9] Chi-Ren, S., Klaric, M., Scott, G., J., Barb, A., S., Davis, C., H., Palaniappan, K., "GeoIRIS: Geospatial Information Retrieval and Indexing System – Content Mining, Semantics Modeling, and Complex Queries," IEEE Transactions on Geoscience and Remote Sensing, Volume 45, Number 4, April (2007).

[10] Cove, J., F., Walsh, B., C., "Online Text Retrieval via Browsing," Information Processing and Management, Vol. 24, No. 1, (1988).

[11] Crestiani, F., Lalmas, M., Van Rijsbergen, C.J., Campbell, I., "'Is This Document Relevant?...Probably': A Survey of Probabilistic Models in Information Retrieval," ACM Computing Surveys. Vol. 30, No. 4 (1998).

[12] Debowski, S., Wood, R., E., Bandura, A., "Impact of Guided Exploration and Enactive Exploration on Self-Regulatory Mechanisms and Information Acquisition through Electronic Search," Journal of Applied Psychology, Vol. 86, No.6, (2001).

[13] Demangeot, C., Broderick, A., J., "Exploration and Its Manifestations in the Context of Online Shopping," Journal of Marketing Management, Vol. 26, No. 13 – 14, (December, 2010).

[14] Giger, H. P., "Concept Based Retrieval in Classical IR Systems," SIGIR '88 Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York (1988).

[15] Grossman, M. R., Cormack, G., V., "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review," Richmond Journal of Law and Technology, Volume 27, Issue 3 (2011).

[16] Grossman, M. R., Cormack, G., V., "Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery," SIGIR'14.

[17] Grossman, M. R., Cormack, G., V., "The Grossman-Cormack Glossary of Technology-Assisted Review," Federal Courts Law Review, Vol. 7, Issue 1 (2013).

[18] Grossman, D. A., Frieder, O., Information Retrieval Algorithms and Heuristics, Kluwer Academic Publishers, Boston, Dordrecht, London. 1998.

[19] Guo, D., Berry, M. W., Thompson, B. B., Bailin, S., "Knowledge-Enhanced Latent Semantic Indexing," Information Retrieval. April (2003).

[20] Harris, Z., "Distributional Structure," Word, Vol. 10, Pg. 146–62 (1954).

[21] Holscher, C., Strube, G., "Web Search Behavior of Internet Experts and Newbies," (2000). Cite as: www9.org/w9cdrom/81/81.html.

[22] Hyman, H. S., Sincich, T., Will, R., Agrawal, M., Padmanabhan, B., Fridy, W., "A Process Model for Information Retrieval, Context Learning and Knowledge Discovery," Artificial Intelligence and Law, Volume 23, Number 2 (2015).

[23] Jarman, J., "Combining natural Language Processing and Statistical Text Mining: A Study of Specialized Versus Common Languages," Working Paper (2011).

[24] Kaplan, S., Kaplan, R., Cognition and Environment. New York: Praeger (1982).

[25] Kostoff, R., Block, J. A., "Factor Matrix Text Filtering and Clustering," Journal of The American Society for Information Science and Technology, 56(9):946-968, (2005).

[26] Luhn, H. P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM Journal of Research and Development, (1957).

[27] March, J. G., "Exploration and Exploitation in Organizational Learning," Organizational Science, 2(1), (1991).

[28] Maron, M. E., Kuhns, J. L., "On Relevance, Probabilistic Indexing and Information Retrieval," Journal of the ACM, 7:216-244 (1960).

[29] Muramatsu, J., Pratt, W., "Transparent Queries: Investigating Users' Mental Models of Search Engines," SIGIR 2001, ACM.

[30] Muylle, S., Moenaert, R., Despontin, M., "A Grounded Theory of World Wide Web Search Behavior," Journal of Marketing Communications, Available Online (09 Dec 2010).

[31] Navarro-Prieto, R., Scaife, M., Rogers, Y., "Cognitive Strategies in Web Searching," Cited as: zing.ncsl.nist.gov/hfweb/proceedings/Navarro-Prieto/index.html. (June 3, 1999).

[32] Oussalaleh, M., Khan, S., Nefti, S., "Personalized Information retrieval System in the Framework of Fuzzy Logic," Expert Systems with Applications, Volume 35, Page 423 (2008).

[33] Richel, T., Perkins, L. A., Yenduri, S., Zand, F., "Determining the Context of Text Using Augmented Latent Semantic Indexing," Journal of the American Society for Information Science and Technology, Vol. 58, No. 14 (2007).

[34] Robertson, S., E., "Progress in Documentation Theories and Models in Information Retrieval," Journal of Documentation, 33 (1977).

[35] Rooney, N., Patterson, D., Galushka, M., Dobrynin, V., "A Scalable Document Clustering Approach for Large Document Corpora," Information Processing and Management, 42: 1163-1175, (2006).

[36] Runkler, T., A., Bezedek, J., C., "Alternating Cluster Estimation: A New Tool for Clustering and Function Approximation," IEEE Transactions on Fuzzy Systems, Volume 7, Page 377 (1999).

[37] Salton, G., Buckley, C., "Term Weighting Approaches in Automatic Text Retrieval," Information Processing and Management, Volume 24, Number 5, (1988).

[38] Singhal, A., Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, (2001).

[39] Spark-Jones, K., Automatic keyword classification for retrieval, Butterworth, London (1971).

[40] Trembley, M., Berndt, D. J., Luther, S. L., Foulis, P. R., French, D. D., "Identifying Fall-Related Injuries: Text Mining the Electronic Medical Record." Information Technology Management. Vol. 10, Page 253 (Nov. 2009).

[41] van Rijsbergen, C. J, Information Retrieval, Butterworth, London, Boston. 1979.