

THE RELATIONSHIP BETWEEN USER PREFERENCES AND IR PERFORMANCE: EXPERIMENTAL USE OF BEHAVIORAL SCALES FOR GOAL ALIGNMENT IN IR PROJECTS

HARVEY HYMAN¹, RICK WILL² & TERRY SINCICH³

¹New College of Florida, United States

^{2,3}University of South Florida, United States

ABSTRACT

This paper tells the story of a series of experiments designed to explore the relationship between behavioral preferences and user performance in information retrieval projects. The experiments are a set of monitored user interactions with a randomly selected set of documents from a large corpus. Users' behavioral preferences are recorded in a pre-test questionnaire, and their subsequent sessions are measured against standardized IR performance metrics of Recall and Precision. User IR performance is analyzed for significant correlations with a set of behavioral scales. The scales are designed to measure user preferences in the areas of tolerance for ambiguity, locus of control, innovativeness in technology, and dispositional innovativeness.

Our findings support that a relationship exists between IR performance measures of recall and precision, and a user's behavioral preferences. Our findings also suggest that behavioral preferences may be used to create a predictive model to forecast a user's IR performance. These findings can be applied to organizations that prioritize strategies depending on the orientation of the searching and sorting goals for an electronic document collection being reviewed

KEYWORDS: Information Retrieval, User Behavior, Recall, Precision, Locus of Control (LOC), Tolerance for Ambiguity (TOA), Personal Innovativeness (PIIT), Dispositional Innovativeness

INTRODUCTION OF THE PROBLEM AND RESEARCH QUESTION STATED

IR projects tend to reflect the stakeholder's interest in finding documents meeting their particular mental model of *relevance* as related to the specific subject matter being reviewed within a corpus of documents. The construct of *Relevance* in this research is defined as a document containing the closest similarity, in content and context, to the subject matter of focus. In this application, an IR system employed to search, sort and select documents from an electronic collection does not inform on the subject matter being queried, but instead, the IR

System informs about the existence of documents containing elements of the subject matter being Queried (Vanrijsbergen, 1979).

To the extent that a system helps to produce documents that are the *most* relevant, and avoid producing documents that are not relevant or less relevant, an IR system supports two objectives: First, it should fulfill the stakeholder's information need, by providing the desired documents, and second, it should save time and cost in the reviewing process, by reducing the number of unwanted documents.

The scenario we explore in this paper is the case of *Relevance* in terms of a set of documents matching a particular information need (relevance criteria) ultimately settled by the judgement of a requester (stakeholder) in a multi-user IR project. In this case the stakeholder is an expert or semi-expert on the subject matter being queried. He/she engages the use of “reviewers” as proxies to scale-up production of the “humans in the loop” of a searching and sorting IR project for processing large collections of electronic documents.

The general problem described herein is both a maximization and a minimization problem: How can the stakeholder communicate his or her mental model of relevance to the reviewers of document collections such that the greatest number of the most relevant documents are retrieved and such that the fewest number of the least relevant documents are retrieved?

We model this problem as a case of leveraging the constructs of *knowledge* and *exploration* (Hyman et al., 2015). When we discuss *knowledge* we are referring to the tacit (know how) mental model of the stakeholder who has a keen understanding of the nature of the context and content of the subject matter being queried for the IR task. The boundary of the stakeholder’s knowledge lies in his or her lack of insight about the contents of *the collection* being queried and the *context of the documents* matching the relevance criteria. The stakeholder knows something about the subject matter, and has a general idea of *what he/she is looking for* – this motivates the first of two research questions: How can we design a tool to support reviewers’ exploration of the content of a collection being queried to develop an understanding of the context of the documents comprising it? This was addressed in a paper by (Hyman, et al., 2015).

Of course, training the reviewer about the content of the collection and context of the documents is not enough. We must also align the skill sets of the reviewer with the strategic goals of the IR task being performed. This motivates the second research question: How can we use behavioral preferences to best align the skill sets of the reviewers with the strategic IR goals of the stakeholder? This is the question addressed by this paper.

Exploration-Exploitation Theory

Our experiments in this area have been following a line of research on the theory of *exploration* – leveraging the user’s natural curiosity and sense making skills (Debowski et al., 2001; Demangeot and Broderick, 2010). When we discuss *exploration* we are referring to a user’s natural tendency to weigh their course of action to drill down on a document found in a collection – represented as *exploitation* (Karimzadehgan and Zhai, 2010), versus abandoning that document in favor of searching for alternative documents that might closer match the stakeholder’s relevance criteria. This phenomenon is acknowledged in the research literature as the “exploration-exploitation dilemma” (Cohen et al., 2007; Hoffman et al., 2013).

IR Process Model

Hyman et al., 2015, developed an IR Process Model which focused on IR user behaviors identified as scanning, skimming, and scrutinizing. The experiment reported in this paper builds on the IR Process Model of Hyman et al., 2015 as a framework to support the study of user behavioral preferences as a predictor of user IR performance. The results reported in this paper provide insight into how a user’s preferences may be used to align a reviewer’s natural tendencies with the strategic goals of the IR project, to improve productivity.

An underlying assumption here is that IR projects can range along a continuum between recall centric (casting a wide net) on one end, and precision centric (executing a more selective, narrow approach) on the other end. Simply put, some stakeholders are more concerned with finding the maximum number of possibly relevant documents, whereas other stakeholders are more concerned with finding a reduced set of the most relevant documents with the understanding that there may be a trade-off of missing some potentially relevant documents.

Description of IR Problem Presented

The IR problem discussed here is modeled as two retrieval tasks: Collection and Evaluation. The first retrieval task is collection – to meet the goal of finding all possible documents that fit the requesting criteria (recall), and avoiding documents that do not fit the criteria (precision). The second retrieval task evaluation, involves the review of the documents in the extracted set.

There are many commonly used IR project examples of this two-tier procedural approach. We motivate our research here using Legal IR and Medical IR where stakeholders and reviewers are significantly represented in conditional document production efforts. In the example of Legal IR, there are two stakeholder groups. The first group is the requestor of documents from the repository of the second stakeholder group, the owner of the document collection. In essence the second group attempts to meet the requestor groups IR task as narrowly as possible – producing that which meets the relevance criteria, and yet avoid producing documents that fall outside the criteria. The motivation here can be a host of issues ranging from privacy interests associated with releasing documents outside of the requirements, to production costs associated with large volume retrieval. In the example of Medical IR, numerous moral, ethical and regulatory issues motivate the IR strategic goal of producing only that which is relevant to the stakeholder’s request.

The strategic IR goal of producing only that which meets relevance criteria is represented as maximizing the number of relevant documents (recall), and minimizing non-relevant documents (precision). We depict the competing interests of Recall versus Precision, and the trade-offs between them, in a confusion matrix—False negative/False positive table in Figure 1.

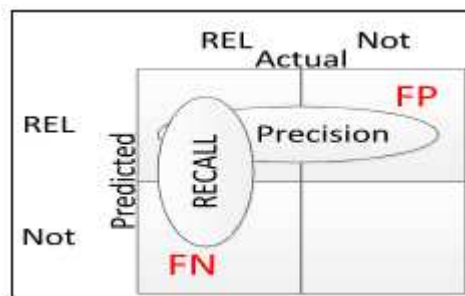


Figure 1: Recall/Precision Relevance Confusion Matrix

We assume the IR stakeholder has a significant frame of reference about the nature, structure and characteristics of the targeted documents. Another assumption is that the stakeholder has a significant frame of reference about the nature and content of the document collection being targeted (Oard et al., 2010; Grossman and Cormack, 2011; Voorhees, 2000).

Motivation to Focus on Behavioral Scales

A significant recurring problem reported in IR projects is how to balance the leverage achieved through

automated methods against the final review stage of human inspection. (Grossman and Cormack, 2014).

The behavioral experiments described in this paper are designed to address this problem by providing insight into how a user's behavioral preferences can be used to align a reviewer's skills and tendencies with the strategic goals of an IR project.

Identifying patterns and preferences, and aligning them to the over-all goals of an IR project can translate into savings in time and cost during the human review process — the most expensive portion of an IR project given that the most expert and highly compensated are assigned to the final review – of great concern to the stakeholder seeking to balance the pressure to reduce cost with the demands of production and quality in the review process.

Discussion on Information Seeking and Automated Tools

Prior research has found that information seeking can be divided into two categories: broad exploration search, and precise search specificity (Heinstrom, 2006). The concept of broad exploration has been found to be a possible indicator of an overview strategy to build knowledge, whereas precise information seeking may be an indicator of a more tightly focused search (Heinstrom, 2006). The underlying assumption here is that in the case of precision search, the user has a specific frame of reference from which to investigate and probe a collection.

Automated methods and tools are an effective way to sort through large collections. However, a recurring limitation associated with IR automated tools lies in the flat nature of using search terms. Ultimately, even the best fitted weighted algorithms and machine learning techniques, in the end only count up the occurrences and distributions of the terms in the query; “the machine” never really “knows” the meaning behind the words or what might be the greater concept of interest to the human performing the search.

Users have the luxury of assuming dependencies between concepts and expected document structures, whereas automated tools leverage knowledge through the use and process of statistical and probabilistic measures of terms in a document, and its relationship to the collection, 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 et al., 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).

There is significant research that suggests 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 can be viewed as approximating the stakeholder’s mental model of relevance. An assumption here is that this can lead to an item that informs the review as the user of the system with additional terms to improve searching and sorting of the collection of documents.

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 users to explore large document collections. The results of their study found that, “visual navigation can be easily used and understood” (Lehman et al., 2010). We adapt this underlying premise along with the IR Process Model (Hyman et al., 2015).

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 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. Some probabilistic models have been proposed for query expansion. These models are based upon the Probability Ranking Principle (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.

Ultimately, all IR tasks share in common some form of the problem of uncertainty. 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. In this research we explore behavioral preferences as a means of explaining how IR users might deal with the uncertainty problem.

Theory and Framework Guiding this Study

The research model used to guide this study is adapted from the *Executives' Information Behaviors Research Model* (Vandenbosch and Huff, 1997). The model is depicted in Figure 2. 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 IR system explored in our study. EIS and IR of an electronic document collection 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, IR users seek solutions in an uncertain environment – extracting relevant documents from a corpus of uncertainty.

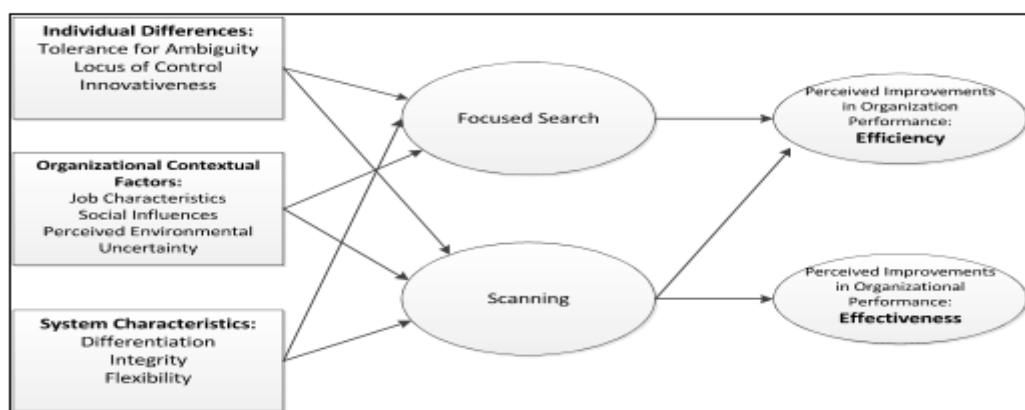


Figure 2: Executives' Information Behaviors Research Model (Vandenbosch and Huff)

In this study we seek to measure behavioral factors that impact recall and precision. The Vandenbosch and Huff Model is adapted to our research here as depicted in Figure 3. The study evaluates whether a user's behavior preferences matter when it comes to IR tasks and design.

The construct of Focused Search is adapted to approximate the search behaviors associated with the performance measure of Precision. This construct is representative of 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*, originally addressed by (Hyman et al., 2015). This construct is representative of the user who browses data looking for trends or patterns, seeking a broad, general understanding of the issue in question (Hyman, et al., 2015; Vandenbosch and Huff, 1997; Aguilar, 1967).

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).

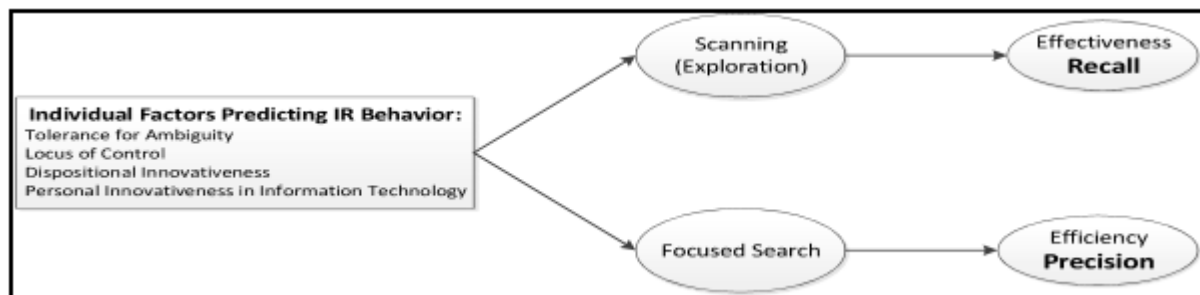


Figure 3: Adapted Information Retrieval Behavior Model

We use four scales to measure individual differences impacting the latent factors of IR performance. The scales of Tolerance for Ambiguity (TOA), Locus of Control (LOC), Dispositional Innovativeness (DISPO), and Personal Innovativeness (PIIT), are operationalized using previously validated instruments (Rydell and Rosen, 1966; Levenson, 1974; Steenkamp and Gielens, 2003; Agarwal and Prasad, 1998).

Population Frame and Sample

The population of interest in this research is made up of digital collection reviewers as IR users. The research presented here explores how behavioral scales can better align the reviewers' preferences with the strategic goals of the IR project for improving performance in the result set.

This study approximates the IR user who does not have an *a priori* mental model for relevance. Instead, he/she seeks a broad scanning/exploring of the collection to gain insight into context and meaning to better understand the model of relevance. This study explores Legal-IR as a specific subject matter of focus and employs law students to approximate legal professionals and litigation support personnel — a total of 120 third year law students representing three

Universities have volunteered to participate in the study. These students are well suited for the study because they have been exposed to Legal-IR concepts in the classroom or have experience through summer clerkships, yet they are relatively less experienced than Legal IR 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 bias that legal professionals develop during their litigation experience.

Document Collection

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.

Data Collection Methods Used

The methods have been used in this study to record the user sessions in the experiments.

They are as follows:

- Notes taken during physical observations of the users performing the IR task;
- Pen and paper questionnaires used to record the behavioral scales;
- Post-task interviews conducted to provide further insight into the testing methods;
- Verbal protocols whereby the users are asked to “think out loud” during the experiment.

We make use of a computer interface application designed to present a series of screens to support the following actions taking place in the sessions:

- Informed consent protocol which must be agreed to by the participant,
- Description of the study,
- IR task description,
- User input screen for selection of search terms,
- User interaction screen to display resulting documents and to record user relevance judgements.

The computer interface application is 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:

- The user is given radio buttons to indicate whether a document is relevant or not relevant;
- 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;
- 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 have been 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, user search techniques, IR results produced, and performance measures.

Data collected from observations, verbal protocols, and pre and post-task interviews have been used to develop quotes for useful descriptions for insight into the experiment sessions, and also to assist the authors in formulating future research questions.

Method of Analysis and Measurement

SAS 9.2 is the statistical package used for the analysis in this study. User IR performance is measured using dependent variables (DVs): *Recall* and *Precision* with a linear regression model. The model is comprised of the behavioral scales Tolerance of Ambiguity (TOA), Locus of Control (LOC), Personal Innovativeness (PIIT), and Dispositional Innovativeness (DISPO).

Data collected to measure the independent variables (IVs) of Locus of Control, Tolerance for Ambiguity, Dispositional Innovativeness, and Personal Innovativeness are analyzed for significance of impact upon the dependent variables (DVs) of *Recall* and *Precision*, in a main effects model. Interactive effects among the IVs are also analyzed using a “full model” which includes the main effects and interactive effects of the stated IVs. All four scales have been analyzed for reliability using Cronbach’s alpha measure.

Document Seeding

The research conducted here is 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. Hyman et al., 2015 assessed how well users are able to identify relevant documents using exploration as a method and manipulating time as a treatment. In that study they used “seeding” of known relevant documents to establish a base-line number of relevant documents within the data set to access *Recall* and *Precision* in the document selections. We apply the same seeding technique used by Hyman, et al. to establish base-lines in this study.

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; Fraser and Zeller, 2010). We accomplish seeding in this study 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 IR 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 (*Recall*) and the percentage of relevant documents within the retrievals (*Precision*).

Pre-Task IR Behavioral Questionnaires

In this study we use known scales previously validated in the literature to anchor our findings about individuals’ exploration search 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 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 (Agarwal and Prasad, 1998).

We also apply a technique to verify how well the participant understood the task requested by the study. After review of the IR task, the participants were 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. The pre-task and task verification questions are listed in the Appendix.

Verbal Protocols, Interviews, Post-Task Questionnaires

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 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 the Appendix.

Post-Task paper and pen questionnaires were used to gain insight into what specific techniques participants used to complete the task, how the participants characterized their chosen techniques as a form of IR solution, and the participants' attitudes toward solving IR problems for development of future research questions.

Description of Task

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. Sets of explanatory variables comprised of behavioral scales known to be associated with preferences that are predictive in the use of technology and innovativeness are recorded prior to the task.

All participants are given the same task. The task is to provide recall (search) terms and elimination terms (filters) in response to an IR project request. The task has been adapted from the TREC Legal Track 2011 Conference Problem Set #401. The problem set is reproduced in the Appendix.

Description of Behavioral Scales

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), Dispositional Innovativeness (DISPO), and Personal Innovation (PIIT). Ten (10) subjects from the participant group have been selected for verbal protocols and are encouraged to "think out loud" while performing the IR task. Post-task interviews are conducted with these subjects to develop further insights into the user IR behaviors and as a means for triangulation against the behavioral scales

Independent Variables (IVs) representing *tolerance for ambiguity (TOA)*, *locus of control (LOC)*, *dispositional innovativeness (DISPO)*, and *personal innovativeness (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 2, representing the impact of behavioral measures upon the dependent variables (DVs) *Recall* and *Precision*. The adapted model is depicted in Figure 3.

Behavior Scales Explained

Personality traits have been associated with information seeking patterns and differences in search approaches and strategies (Heinstrom, 2006). The four behavioral scales explained above 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*. The four behavioral scales and their corresponding Alpha values are listed in Table 1. They are further described and explained in a narrative in the next sections.

Table 1: List of Behavior Scales

Variable	Name	Description	Number of Items	Cronbach's Alpha
TOA	Tolerance for Ambiguity	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)	8	.80
LOC	Locus of Control	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.	5	.85
DISPO	Dispositional Innovativeness	The measure of an individual's likeliness to try a new product, or think tangentially when solving a problem.	8	.85
PIIT	Personal Innovativeness in the Domain of Information Technology	The degree to which an individual has a preference for technology use.	4	.97

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 4, below and in written form as follows;

H1a: TOA is positively related to Recall.

H1b: TOA is negatively related to Precision.

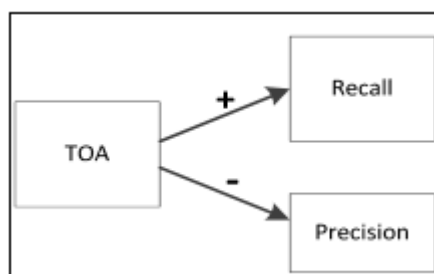


Figure 4: TOA Effect upon Recall and Precision

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 Legal IR 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

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.

We believe 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 5, and presented in written form as follows;

H2a: *LOC is positively related to Recall.*

H2b: *LOC is negatively related to Precision.*

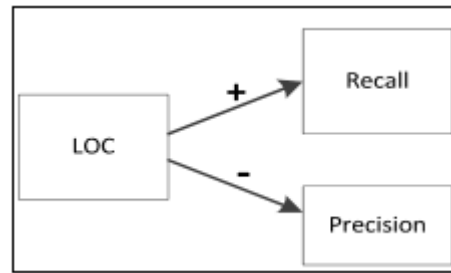


Figure 5: LOC Effect upon Recall and Precision

Dispositional Innovativeness

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 study 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 dispositional 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). One of the hypotheses in this study 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.

We believe 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 this proposition are depicted in Figure 6 and in written form as follows:

H3a: *DISPO is positively related to Recall.*

H3b: *DISPO is negatively related to Precision.*

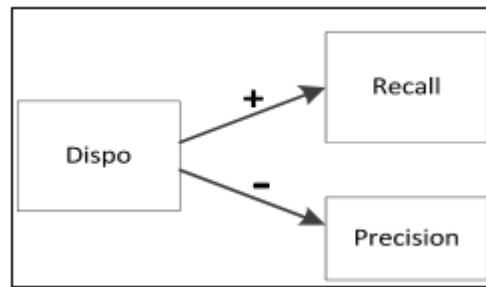


Figure 6: DISPO Effect upon Recall and Precision

Personal Innovativeness (PIIT)

“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 an IR 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 7 and in written form below:

H4a: *PIIT is negatively related to Recall.*

H4b: *PIIT is positively related to Precision.*

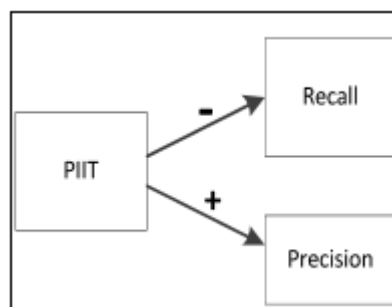


Figure 7: PIIT Effect upon Recall and Precision

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.

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 1.

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 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: $DV_{\text{Recall}}, DV_{\text{Precision}} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_1X_2 + B_6X_1X_3 + B_7X_1X_4 + B_8X_2X_3 + B_9X_2X_4 + B_{10}X_3X_4 + e$

Where

X1 = TOA,

X2 = LOC,

X3 = DISPO,

X4 = PIIT,

Statistical Analysis of Models

A global F-test has been performed upon the behavioral model for Recall and Precision.

A summary of results appears in Table 2 below. The null and alternative hypotheses are:

Recall	Precision
H0: B1 = B2 = B3 = B4 = 0	H0: B1 = B2 = B3 = B4 = 0
Ha: At least one Beta ≠ 0	Ha: At least one Beta ≠ 0

Where:

X1 = Tolerance for ambiguity (TOA),

X2 = Locus of control (LOC),

X3 = Dispositional innovativeness (DISPO),

X4 = Personal innovativeness (PIIT).

Table 2: Summary of Behavioral Model Results

Independent Variables	Alpha	Dependent Variable Effected	Beta Estimate
TOA	.01	Precision	.005
LOC	.01	Recall	-.013
DISPO	.05	Precision	.008
PIIT	Not Significant		

The global F-test for the *Recall* behavioral model and the *Precision* behavioral model are both significant at alpha .01. However, the behavioral models differ in which variables were found to be significant for *Recall* and which were found to be significant for *Precision*:

- LOC was significant for *Recall* at alpha .01.
- TOA was significant for *Precision* at alpha .01.
- DISPO was significant for *Precision* at alpha .05.
- PIIT was not supported for *Recall* or *Precision*.

The printouts for these results appear in Table 3 and Table 4.

Table 3: SAS 9.2 Printout for Recall Variables

The REG Procedure						
Model: MODEL1						
Dependent Variable: RECALL						
Number of Observations Read 120			Number of Observations Used 120			
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	4	1.16472	0.29118	147.12	<.0001	
Error	55	0.10885	0.00198			
Corrected Total	59	1.27357				
Root MSE	0.04449			R-Square	0.9145	
Dependent Mean	0.50733			Adj R-Sq	0.9083	
Coeff Var	8.76897					
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard 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	0.7702
DISPO	DISPO	1	-0.00091858	0.00293	-0.31	0.7547
PIIT	PIIT	1	0.00320	0.00124	2.59	0.0124

Table 4: SAS 9.2 Printout for Precision Variables

The REG Procedure						
Model: MODEL1						
Dependent Variable: PRECISION						
Number of Observations Read 120			Number of Observations Used 120			
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	4	0.60044	0.15011	138.06	<.0001	
Error	55	0.05980	0.00109			
Corrected Total	59	0.66024				
Root MSE		0.03297	R-Square		0.9094	
Dependent Mean		0.61600	Adj R-Sq		0.9028	
Coeff Var		5.35284				
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	0.22744	0.03401	6.69	<.0001
LOC	LOC	1	-0.00012484	0.00144	-0.09	0.9312
TOA	TOA	1	0.00542	0.00110	4.91	<.0001
DISPO	DISPO	1	0.00833	0.00217	3.84	0.0003
PIIT	PIIT	1	0.00003059	0.00091712	0.03	0.9735

Interactive Effects Analyzed

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 supported 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 for interactive effects and multi-collinearity have been produced in Table 5, Table 6 and Table 7.

Table 5: SAS 9.2 Printout for Recall Variables

The REG Procedure						
Model: MODEL1						
Dependent Variable: RECALL RECALL						
Number of Observations Read	120	Number of Observations Used	120			
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	9	1.21496	0.13500	115.16	<.0001	
Error	50	0.05861	0.00117			
Corrected Total	59	1.27357				
Root MSE	0.03424	R-Square	0.9540			
Dependent Mean	0.50733	Adj R-Sq	0.9457			
Coeff Var	6.74866					
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard 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
PIIT	PIIT	1	0.00908	0.00787	1.15	0.2541
PIIT-TOA		1	-0.00039540	0.00022606	-1.75	0.0864
PIIT-DISPO		1	0.00025541	0.00048162	0.53	0.5982
LOC-DISPO		1	-0.00008662	0.00073713	-0.12	0.9069
LOC-TOA		1	-0.00068173	0.00035911	-1.90	0.0634
DISPO-TOA		1	-0.00002182	0.00011459	-0.19	0.8498
Model: MODEL1						
Test 1 Results for Dependent Variable RECALL						
Source	DF	Mean Square	F Value	Pr > F		
Numerator	5	0.01005	8.57	<.0001		
Denominator	50	0.00117				

Table 6: SAS 9.2 Printout for Precision Variables

The REG Procedure
Model: MODEL1
Dependent Variable: PRECISIO PRECISION

Number of Observations Read	120
Number of Observations Used	120

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	0.61878	0.06875	82.91	<.0001
Error	50	0.04146	0.00082926		
Corrected Total	59	0.66024			

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	Parameter Estimate	Standard 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
PIIT	PIIT	1	0.00063453	0.00662	0.10	0.9240
PIIT-TOA		1	0.00014418	0.00019014	0.76	0.4518
PIIT-DISPO		1	-0.00018739	0.00040508	-0.46	0.6457
LOC-DISPO		1	0.00006220	0.00061998	0.10	0.9205
LOC-TOA		1	0.00033756	0.00030204	1.12	0.2691
DISPO-TOA		1	0.00017096	0.00009638	1.77	0.0822

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

Source	DF	Mean Square	F Value	Pr > F
Numerator	5	0.00367	4.42	0.0021
Denominator	50	0.00082926		

Table 7: SAS 9.2 printout of Multi-Collinearity Analysis

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

	PIIT	LOC	TOA	DISPO
PIIT	1.00000	-0.89706	-0.00623	-0.12841
PIIT		<.0001	0.9623	0.3282
LOC	-0.89706	1.00000	-0.22654	-0.07217
LOC	<.0001		0.0818	0.5837
TOA	-0.00623	-0.22654	1.00000	0.91590
TOA	0.9623	0.0818		<.0001
DISPO	-0.12841	-0.07217	0.91590	1.00000
DISPO	0.3282	0.5837	<.0001	

Summary of Findings

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 8 below.

Table 8: List of Hypotheses Supported and Not

Hypothesis	Supported/Not	Variable	Alpha	Relationship to Recall/Precision
H1a	Not	TOA		
H1b	Supported	TOA	.01	Precision: Direct and Pos
H2a	Supported	LOC	.01	Recall: Direct and Pos
H2b	Not	LOC		
H3a	Not	DISPO		
H3b	Supported	DISPO	.05	Precision: Direct and Pos
H4a	Not	PIIT		
H4b	Not	PIIT		

***Interactive effect upon Precision supported**

LIMITATIONS

This study like all studies has limitations that can be improved upon in future extensions. The first limitation lies in the finding that several variables were found to not be significant. One possible reason for this is that our sample size (N=120), might not have been large enough to detect a result. We plan to address this in future extensions by testing against alternative IR tasks, and possibly switching the task to a Medical IR project to explore the commonalities and differences in user behavioral effects between Legal and Medical IR projects.

A more critical limitation in this study might be 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 biased in their searching behaviors by years of legal experience that may impact the IR task. We plan to conduct future studies with paralegals and lawyers to determine if legal experience matters in this form of IR. This might also impact our ability to generalize these findings to other IR projects, especially if Legal IR tasks are found to have behaviors that are peculiar to Legal IR alone. This is something we also will consider to pursue in our next extension on this topic.

CONTRIBUTION

The study reported in this paper makes several significant contributions to theory. The main contribution is the investigation into how behavioral preferences can be correlated to a user's performance in multi-user IR projects.

There is clearly a relationship between user behaviors and IR performance. The significance and magnitude will remain to be seen in extension work and future experiments.

As a result of our investigation into the use of behavioral scales for IR projects, we have discovered some new relationships. The model validated here suggests that these relationships can be of significant use to the stakeholders in IR projects. By aligning the behavioral scales of the reviewer to the strategic goals of the IR project, significant performance differences may be produced, which can translate into time and cost savings, as well as better production in Recall and Precision.

CONCLUSIONS

In this paper we set out to tell the story of a series of experiments designed to explore if there is a significant relationship between user behaviors and IR performance measures, and if so, how can we create a model to apply behavioral scales to IR projects.

The results produced by this study help explain which behavioral preferences have significant impact on IR performance and which are not yet supported by evidence. The measured variables used in this study help explain user actions and strategies and their significance upon IR production.

The contribution of this study lies in the validation of the behavioral IR model, and its insights into how differences in behavioral variables *locus of control*, *tolerance for ambiguity* and *dispositional innovativeness* can have an impact on the user's IR result when evaluated by *Recall* and *Precision*.

REFERENCES

1. **Agarwal**, R., Prasad, J., "A Conceptual and Operational Definition of Personal Innovativeness in the Domain of Information Technology," *Information Systems Research*, Vol. 9 No. 2, June (1998).
2. **Aguilar**, F., J., *Scanning the Business Environment*, MacMillan, New York, 1967.
3. **Bates**, M., J., "Information Search Tactics," *Journal of the American Society for Information Science*, July, (1979).
4. **Bates**, M., J., "Subject Access in Online Catalogs: A Design Model," *Journal of the American Society for*

- Information Science*, November (1986).
5. **Bates, M., J.**, “The Design of Browsing and Berry Picking Techniques for the Online Search Interface,” *Online Review*, 13 (5), (1989).
 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. **Burke, Russell T., Esq., Rowe, Robert D., Esq.**, “Legal and Practical Issues of Electronic Information Disclosure,” Nexsen Pruet Adams Kleemeier, LLC, (2004).
 8. **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).
 9. **Cohen, J., D., McClure, S., M., Yu, A., J.**, “Should I Stay or Should I Go,” *Philosophical Transactions: Biological Sciences*, Vol. 362, No. 1481, *Mental Processes in the Human Brain* (May, 2007), The Royal Society.
 10. **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).
 11. **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).
 12. **Deerwester, S., Dumais, S. T., Furnas, G.W., Landauer, T. K., Harshman, R.**, “Indexing by Latent Semantic Analysis,” *Journal of the American Society for Information Science*. (Sep. 1990).
 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. **Fraser, G., Zeller, A.**, “Mutation-driven Generation of Unit Tests and Oracles.” *Proceedings of the 19th international symposium on Software testing and analysis*. July (2010).
 15. **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).
 16. **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).
 17. **Grossman, M., R., Cormack, G., V.**, “Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery,” *SIGIR'14*.
 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. **Heinstrom, J.**, "Broad Exploration or Precise Specificity: Two Basic Information Seeking Patterns Among Students," *Journal of The American Society for Information Science and Technology*, 57(11), 2006.
21. **Hoffmann, K.**, Whitson, S., de Rijke, M., "Balancing Exploration and Exploitation in Listwise and Pairwise online Learning to Rank for Information," *Information Retrieval*, 16:63-90 (2013).
22. **Huber, G. P.**, "Organizational Learning: The Contributing Processes and the Literatures," *Organization Science*, (2:1), (1991).
23. **Hyman, H. S.**, Fridy III, W., "Using Bag of Words (BOW) and Standard Deviations to Represent Expected Structures for Document Retrieval: A Way of Thinking that Leads to Method Choices," NIST Special Publication, Proceedings: Text Retrieval Conference (TREC) 2010.
24. **Hyman, H. S.**, Sincich, T., Will, R., Agrawal, M., Fridy, W., Padmanabhan, B., "A Process Model for Information Retrieval Context Learning and Knowledge Discovery,"
25. *Artificial Intelligence and Law Journal*, Volume 23, Issue 2, pp. 103 - 132, (2015).
26. **Jarman, J.**, "Combining natural Language Processing and Statistical Text Mining: A Study of Specialized Versus Common Languages," Working Paper (2011).
27. **Karimzadehgan, M.**, Zhai, C., X., "Exploration-Exploitation Tradeoff in Interactive Relevance Feedback," *Conference on Information and Knowledge Management* (2010).
28. **Kirton, M. J.**, "Adaptors and Innovators: A Description and Measure," *Journal of Applied Psychology*, (61:5), (1976).
29. **Lehman, S.**, Schwanecke, U, Dorner, R., "Interactive Visualization for Opportunistic Exploration of Large Document Collections," *Information Systems*, 35 (2010).
30. **Levenson, H.**, "Activism and Powerful Others: Distinctions within the Concept of Internal- External Control," *Journal of Personality Assessment*, (38), (1974).
31. **McCaskey, M. B.**, "Tolerance for Ambiguity and the Perception of Environmental Uncertainty in Organization Design," *The Management of Organization Design*, Kilman, Pondy, Slevin (eds.), (1976).
32. **Oard, D. W.**, Baron, J. R., Hedin, B., Lewis, D. D., Tomlinson, S., "Evaluation of Information Retrieval for E-discovery," *Artificial Intelligence and Law*, 18:347 (2010).
33. **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).
34. **Pearson, P. H.**, "Relationships Between Global and Specific Measures of Novelty Seeking," *Journal of Consulting and Clinical Psychology*, Vol. 34 (1970).

35. **Presson, P., K., Clark, S., C., Benassi, V., A.,** “The Levenson Locus of Control Scales: Confirmatory Factor Analysis and Evaluation,” *Social Behavior and Personality*, 25 (1), (1997).
36. **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).
37. **Robertson, S., E.,** “Progress in Documentation Theories and Models in Information Retrieval,” *Journal of Documentation*, 33 (1977).
38. **Rogers, E., M., Shoemaker, F., F.,** *Diffusion of Innovations*, Third Edition, The Free Press, New York, (1995).
39. **Roehrich, G.,** “Consumer Innovativeness Concepts and Measurements,” *Journal of Business Research*, Vol. 57 (2004).
40. **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).
41. **Rydell, S., T., Rosen, E.,** “Measurement and Some Correlates of Need Cognition,” *Psychological Reports*, (19), (1966).
42. **Salton, G.,** *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA, (1989).
43. **Steenkamp, J, Gielens, K.,** “Consumer and Market Drivers of the Trial Probability of New Consumer Packaged Goods,” *Journal of Consumer Research*, Vol. 30, No. 3 (December, 2003).
44. **Todd, P., Benbasat, I.,** “Process Tracing Methods in Decision Support Systems Research Exploring the Black Box,” *MIS Quarterly* (11:4), (1987).
45. **TREC Proceedings:** NIST Special Publication, Voorhees and Buckland Editors, (2010, 2011)
46. **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).
47. **Vandenbosch, B., Huff, S., L.,** “Searching and Scanning: How Executives Obtain Information from Executive Information Systems,” *MIS Quarterly*, Vol. 21, No. 1 (Mar. 1997).
48. **Van Rijsbergen, C. J,** *Information Retrieval*, Butterworth, London, Boston. 1979.
49. **Venkatraman, M., P., Price, L., L.,** “Differentiating Between Cognitive and Sensory Innovativeness Concepts, Measurement, and Implications,” *Journal of Business Research*, (20), (1990).
50. **Voorhees, E., M.,** “Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness,” *Information Processing and Management*, Volume 36, Page 697 (2000).

APPENDIX

Pre-Task Questionnaire for User Understanding of Request

Pre-Task Strategy Questionnaire

- Summarize in one or two sentences what the request is seeking?
- What concepts do you believe define the documents that satisfy the request?
- 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...
- Narrative Questions

Post-Task Questionnaire

- When I conduct an information search, the type of information I expect to find is?
- If I had to choose between being efficient or being thorough, I would choose.
- When I conduct an information search, the format I expect the information to be found is in: Web page, Web Site, PDF, Email, Other?
- When I find an information item, I evaluate it to determine if it meets my need by?
- When conducting a specific search for documents, my search method differs from a search for web pages or web sites because?
- When I select a document for review I focus on.
- I search for documents contained within a collection of documents to meet my information need by doing the following:
- I use the following criteria to evaluate whether a document meets my information need:
- When I search for documents within a collection of documents, I define/determine what I am looking for by?
- When viewing a document in a collection, the items I focus upon within that document that help me determine if that document meets my requirements (information need) are?
- Scaled Agree/Disagree Questions (-3 to +3)
- When I search for information, I am most concerned with being efficient.
- 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.
- 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.
- When I select a document I almost always review the entire document.

- 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.
- My only method of sorting is to scan titles.
- When I search for information, I am most concerned with being thorough.
- When I search for information, I prefer to scrutinize (review entire content) the documents whose titles seem to meet my information need.
- My first/immediate method of sorting is to scan titles.
- I use titles to base my selection of documents.
- 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.
- When I select a document I rarely review the entire document.
- Scaled Agree/Disagree Questions (-3 to +3) When I search for documents:
 - I limit the depth of my exploration to scanning of titles of documents alone.
 - I scan titles and then skim selected documents based on the content of the titles.
 - I select documents based on titles, but I also randomly select documents for a broad exploration of the collection.
- When I select a document:
 - I prefer to limit my review to the first paragraph of the document.
 - I prefer to skim the entire document to get a general understanding of the content.
 - I prefer to scrutinize the entire document to get an in depth understanding of the content.

IR Task and Participant Instructions

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.

