2009

Representational Indeterminancy and Enterprise Search: The Importance of Subject Indexes

Gregory Schymik
Arizona State University, gschymik@asu.edu

Representational Indeterminacy and Enterprise Search: The importance of subject indexes

Gregory Schymik
Department of Information Systems
W. P. Carey School of Business
Arizona State University
gschymik@asu.edu

ABSTRACT

The proposed research examines the impact that adding context – via the use of subject indexes – to a query has on search results. This design science research is motivated by the need for a solution to the well-documented failure of enterprise search. Preliminary experimental data is presented that indicates that the use of subject indexes to augment full-text search may indeed be a valid solution and thereby encourages continued investigation. Continuing the research, we propose an experiment where we simulate the search for randomly selected (single) documents in a collection. We will use the comparison of search results between full-text only and full-text plus subject metadata searches to evaluate search performance. The primary dependent variable used will be the rank of the searched-for document in the result set returned by the search engine for each search.

Keywords

Orderly distribution of meanings, design science, search futility points, dimensional search, full-text search, known-item search.

INTRODUCTION

Firms struggle to integrate knowledge management processes into their business processes. One reason for this struggle is the difficulty involved in transferring the knowledge possessed by a firm's knowledge workers to others in the firm. Enterprise search is a popular, but frequently unsuccessful, mechanism for transferring knowledge amongst knowledge workers inside individual firms. According to data presented during a recent Google webinar on the release of a new version of their enterprise search appliance, knowledge workers are wasting almost half of their time as a direct result of poor search capabilities (See Figure 1). They also spend another 25% of their time conducting what they define to be successful searches for information, leaving only about one quarter of a knowledge worker's time being spent on truly value added activity. Middle managers further noted that often times, the information they do find is wrong (KMWorld 2008). This data makes it no surprise that 86% of enterprise searchers are unsatisfied with their enterprise search capabilities (KMWorld 2008).

The success achieved by web search engines has led to the development of enterprise search tools that use those same, or very similar search engines to help users mine their corporate intranets and networks for unstructured information. These tools have, in most cases, failed to meet the needs of their users because full-text keyword-based searches are not the proper tool for workers in the enterprise context (Alavi and Leidner 2001;Gardner 2008), particularly for those workers who perform knowledge intensive tasks (Kontzer 2003).

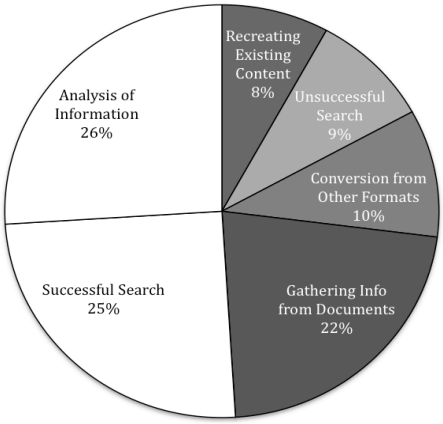


Figure 1. Impact of Ineffective Search on Knowledge Workers (adapted from KMWorld (2008))

The reasons behind the failure of full-text search appliances as enterprise search tools are myriad (Fagin, Kumar, McCurley, Novak, Sivakumar, Tomlin , and Williamson 2003;Raghavan 2001). Because internet search engines form the foundation of most enterprise search tools, one reason for failure is the differences between the requirements placed upon internet searches and enterprise searches. The most fundamental of these differences is the target of the search. When searching the internet, most searchers are looking for information about something. They are not exactly sure what they are looking for. Enterprise searchers, however, are most often looking for something specific (Fagin et al. 2003). They are searching for something they either know exists or suspect exists somewhere in the enterprise's vast store, or stores, of unstructured data.

We know that searchers do not care to search through too many pages of search results to find relevant material (Jansen and Spink 2006;Spink, Wolfram, Jansen , and Srarcevic 2001). In fact, research has shown that many searchers rarely look beyond the first or second page of search results returned by search engines (Jansen et al. 2006). Although very little research has investigated this phenomena in the enterprise context, some research exists that indicates that this same issue applies to searchers in the enterprise environment (Stenmark and Jadaan 2006). This indicates that effective enterprise search tools must return relevant documents in the first 10 to 20 results if the search is to be considered effective. Full-text, keyword-based searches are inherently biased against meeting this criterion.

Full-text searches tend to return large numbers of documents containing various combinations of the keywords submitted for the search. A typical keyword-based, full-text search engine does not take into account any sort of contextual information. It simply looks to match keywords, and/or combinations of keywords, to the words contained in the documents. Since words can take on many meanings, such searches tend to return documents that run the gamut between being highly relevant and totally irrelevant. Users in the enterprise context do not have the time to search through every artifact returned by these searches. Knowledge workers would not have time to do their assigned work if we expected them to read through every artifact returned in a typical search. The typically large result sets tend to cause searchers to abandon their search (Blair 2002b) in favor of spending time to recreate the document for which they were searching. Figure 1 indicates that approximately half of a knowledge worker's time is wasted due to poor search capabilities.

There must be a more efficient method for obtaining the knowledge searchers seek than reviewing every piece of information presented to them by a search engine. They need a method of searching their archives that dramatically reduces result set size while returning the relevant documents.

It is not hard to believe that the limitations of keyword-based, full-text searches could be costing corporations $billions. Google estimates that it could cost a company with 1000 employees approximately $21M annually (KMWorld 2008). Other cost estimates range from $9M to $33M annually per firm (EContent 2004;Ultraseek 2006). Simulations have estimated that a context-based, dimensional search could help to significantly reduce these costs over a fairly short period of time (Corral, Schuff, St. Louis , and Turetken 2007). This problem begs for a solution.

The notion that full-text, keyword-based search is not the solution is supported by Drabenstott (2004). She points out that she and other information retrieval researchers tend to avoid using cumbersome, full-text, keyword-based searches and rely

instead on subject, author, and bibliographic searches to find large sets of relevant documents. She calls for research into means by which these domain experts' tactics can be presented to end users of search tools (Drabenstott 2004).

If information retrieval researchers do not use keyword queries, why does everyone else so often rely on them? Zipf's Principle of Least Effort suggests this makes no sense (Zipf 1949). If there are easier methods to be applied to the search for information, why do we continue to use the seemingly more difficult keyword search?

The problem with keyword-based, full-text searches is that the nature of language works against the goal of returning only the artifacts relevant to the searcher's needs. Information science ascribes this problem to the representational indeterminacy of language – the fact that words can take on many meanings - and posits that contextual information needs to be added to the artifacts in order to mitigate the representational indeterminacy inherent in full-text searches (Blair 2006). From the information science perspective, it becomes a problem of description versus discrimination. As document collections get large, the complexities of language make it very difficult to define a set of query terms that will adequately describe the documents we search for yet sufficiently discriminate between relevant and irrelevant documents (Blair 2002a).

The problems caused by the indeterminacy of language are not limited to enterprise search. They also apply to internet search. Some claim that the solution to this and all of our information needs related problems exists in the form of the semantic web (Berners-Lee, Hendler , and Lassila 2001). The semantic web will result in information being available *in context:* the addition of semantics to web pages will enable agents to decipher the information available and solve problems associated with the ambiguity (i.e. – representational indeterminacy) of language that cause searches to return too many results. The semantic web is the ideal, but currently unrealized, solution to the problem. The costs are currently too high and challenges too broad for the semantic web to be realized but research indicates that it may soon be achievable (Hendler 2001;Sure, Hitzler, Eberhart , and Studer 2005).

The ideal of the semantic web may currently be out of our reach but the idea makes sense. Can we apply the idea of the semantic web in an environment where the costs do not outweigh the benefits? This research suggests that this can be done in the enterprise search environment.

The literature suggests that if the searcher can be provided a means by which they can reduce the representational indeterminacy of language inherent in full-text, keyword-based searching, they should have a more successful search experience. This suggestion leads to the research question on which this dissertation focuses:

Does reducing representational indeterminacy improve the effectiveness of document retrieval?

The remainder of this paper contains three sections. The first section presents a brief literature review of relevant knowledge management, behavioral science, information retrieval, and information sciences literature. The paper then presents the proposed methodology for the dissertation along with some preliminary results, and then ends with a brief conclusion to summarize the dissertation proposal.

LITERATURE REVIEW

Alavi and Leidner (2001) provide an often-cited foundation for knowledge management research (a recent check in Google Scholar noted 870 citing articles). Pointing out that little IT-based research had been done in the field (most research in the field to that point had arisen from strategic and organizational theory), they present a framework of knowledge management processes and discuss the roles IT might play in those processes. From the practical perspective, this work provides the fundamental motivation for this dissertation.

Focusing on knowledge retrieval, this research is motivated by two of the questions posed by Alavi and Leidner (2001):

1. How much context needs to be included in knowledge storing to ensure effective interpretation and application?
2. What retrieval mechanisms are most effective in enabling knowledge retrieval?

This research attempts to answer the first question by looking at the impact reducing representational indeterminacy through the addition of different amounts of contextual information could have on retrieval effectiveness. The second question is simultaneously addressed by noting that, if context does indeed need to be added to ensure effective interpretation, then a retrieval mechanism must be designed to use that contextual information to aid in knowledge retrieval.

The Principle of Least Effort suggests that people choose to apply solutions to problems that will minimize their effort required to solve both the problem they face and the problems they are likely to face in the future each according to their own

interpretation. In other words, the principle states that all of us are constantly driven by the urge to minimize effort in all that we do. (Zipf 1949).

In regards to human language, Zipf notes that words are tools that we use to convey meaning in order to achieve objectives. When applied literally, the Principle of Least Effort would suggest that speakers would prefer that a single word take on all possible meanings so that they would need to state only the one word to express their intended meaning while listeners would prefer that each word take on only one meaning so that they are not required to do any work deciphering the speaker's intended meaning. These two extremes are both unrealistic. Zipf, therefore, argues that two conflicting forces, The Forces of Unification (one word, all *m* meanings) and of Diversification (*m* words, one meaning for each word) act in concert to achieve a balanced, or orderly, distribution of meanings amongst words. He goes on to observe that the number of meanings a word takes on in a given collection of documents is roughly equivalent to the square root of the number of times the word appears in that set of documents. That is, if m_r represents the number of meanings of the *r*-th ranked (by frequency) word in a collection and F_r represents the frequency of occurrence (the number of times the word appears in the collection) of the *r*-th ranked word in the collection, then the following approximation applies:

$$m_r \approx \sqrt{F_r}$$

As document sets get large, the occurrence of a given word increases. The orderly distribution of meanings tells us that as the occurrence of a word increases, the word takes on more and more meanings. As a word takes on more and more meanings in a document collection, a keyword used in the search of that document collection will have many more meanings than the one meaning intended by the searcher. Therefore, as a document collection gets larger, the power of a given keyword to retrieve documents relevant to the searcher's inquiry is reduced. The search engine will return all of the documents with that keyword when the searcher desires only the documents containing the "version" of that keyword possessing their intended meaning and only their intended meaning. This leads to failed searches.

Blair attempts to explain the failure of searches from the perspective of the searcher and suggests that the searcher needs to avoid two futility points while searching or they will give up and call the search a failure (Blair 2002b). The *anticipated futility point* represents the largest number of documents through which a searcher is willing to begin searching and the *search futility point* is the total number of documents through which a searcher is willing to look for over the entire search. These two futility points represent the information retrieval domain's manifestation of the principle of least effort. An effective search mechanism needs to avoid hitting a searcher's utility points by reducing the size of the returned document sets and returning more relevant documents in those result sets.

Blair also focuses on the *determinacy of representation* describing the problems faced by large-scale document retrieval systems in terms of description and discrimination (Blair 2002a). Determinacy of representation is defined to be the measure of how precisely a document can be described in a given system. In systems biased towards description (usually a characteristic of a full-text search system), one can see that it is fairly easy to make a prediction about which words will be in a sought after document. However, given the fact that the number of meanings a word takes on increases with the square root of the number of times the word appears in a given collection (Zipf 1949), it is also fairly obvious that, for reasonably large collections (those containing more than a few hundred documents) it is nearly impossible to choose a set of keywords that will discriminate relevant from irrelevant documents.

Documents in full-text search systems tend to be over-described. Since an index of the entire collection of words in the document (excluding a stop-list of words) is used to represent the document, it is likely that some, or many, of the words in that description actually mis-represent the intellectual content of the document. These systems only describe the content of the documents stored within. Searches in such systems tend to exhibit an inability to achieve precision in their result sets. They very often return very large result sets that contain a large number of irrelevant documents.

In systems biased in favor of discrimination (such as a system allowing searches across only the title and author fields in a bibliographic record), it is easily seen that the descriptions of documents discriminate each document from each other document. These systems use only contextual information to represent documents. The problem for searchers in these systems is that it is unlikely that they can recall the precise terms necessary to return the documents for which they are searching unless the searcher can precisely recall the title and author of the document. Such systems tend to fail to achieve adequate recall in their result sets. They very often return small result sets that fail to include many of the relevant documents in the collection. This is most often the result of the fact that the connection between the contextual information found in the bibliographic record and the actual content of the document the record represents is rarely more than inferential.

Measure of Effectiveness	System Bias	
	Description	Discrimination
Recall	HIGH	LOW
Precision	LOW	HIGH

Table 1. Blair's Representational Determinacy

The problem of representational determinacy (or indeterminacy) is summarized in Table 1. In order to avoid hitting utility points, resulting in searchers abandoning their searches and organizations taking on the added costs of such failed searches, an effective knowledge retrieval system must achieve a proper balance between description and discrimination. The goal in such a system is to achieve highly determinate representations of the documents stored within so that search results will tend to exhibit both high recall and high precision.

The STAIRS study (Blair and Maron 1985), in an operational experiment in the legal environment, concluded that full-text searching of large collections is not a satisfactory solution. While evaluating a system intended to aid in discovery using full-text search of documents, searchers found only 20% of the relevant documents and only 48% of the highly-relevant documents in the collection while the goal of the system was to find 75% of the relevant and 100% of the highly relevant documents. This comprehensive study was the first to suggest that full-text search was an inadequate technology.

Wu and Li (2008) experimented with using a search interface that helped reduce representational indeterminacy by adding keyphrases automatically extracted from the documents in the collection to the document snippets in the search results. Unlike subject metadata taken from a controlled vocabulary, these keyphrases associated with each document were taken from the documents themselves so any keyphrase associated with the document must have appeared in that document. Subjects were asked to find four documents relevant to their search and the number of documents opened by each subject before they found the fourth was recorded. The results showed that reducing representational indeterminacy by adding the keyphrases resulted in a significant reduction in recall effort for the searchers using the new interface compared to those using a traditional interface (Wu and Li 2008).

Research into the impact metadata has on search results has produced varied results. Storey, Burton-Jones, Sugumaran, and Purao (2008) recently studied the impact adding contextual information to the query can have on internet search engine query results (Storey, Burton-Jones, Sugumaran , and Purao 2008). They developed a methodology (CONQUER CONtext-aware QUERy processing) of overcoming the problems associated with Zipf's ordinary distribution of meanings that applied the concept of word sense disambiguation to expand a searcher's query with the intention of specifying the intended meanings of the keywords used by the searcher. Where possible, the system attempted to automatically perform word sense disambiguation of the search terms. When the system could not determine a single word sense, it involved the searcher in choosing terms to help clarify the word sense. In a laboratory experiment using 261 subjects, they found that using the CONQUER system improved precision at 10 and 20 web pages returned over the basic internet search engine (Google, and AlltheWeb) results

Research done during the transition to online library catalogs led to the conclusion that adding subject metadata to the bibliographic record added value to the bibliographic records and improved search results, doubling relative recall by searching the subject terms instead of the titles (Voorbij 1998). In a related study, Gross and Taylor (2005) found that with the use of subject headings, keyword searches of bibliographic records (i.e. a full-text search of the bibliographic record but not the document text) in a university's online public access catalog (OPAC) system would return 40% fewer records. These initial studies were done before retrieval systems had the ability to search the text of the document.

Hemminger et al. (2007) compared full-text searching to the searching of title and abstract metadata in two online medical collections. They searched for gene names, which were usually acronyms such as COMT, and found that, on average, the documents returned by the metadata searches were more useful, as rated by expert reviewers, than were those returned by the full-text only searches. However, they also found that full-text search results could be improved to an equivalent level by simply weighting the frequency of occurrence of the acronym (keywords) more heavily in their document-ranking scheme. This led them to conclude: "…it may be time to make the transition to direct full-text searching as the standard" (Hemminger et al. 2007). The fact that prior research provides plenty of evidence to counter this conclusion demonstrates that the question is still an open and relevant research topic.

RESEARCH METHODOLOGY

This dissertation applies the design science methodology (Hevner, March, Park , and Ram 2004) to the investigation of the research question. Word limits prevent a detailed discussion regarding the design science contribution of this work. Figure 2 demonstrates that enterprise search failure meets the specified requirements of a "wicked problem" and Figure 4 details how the research method we propose (a single-item simulated search experiment defined later in this section) meets the requirements for a design science research contribution. Of course, this material on the research contribution is only speculative right now and may change once we get actual results from the experiments.

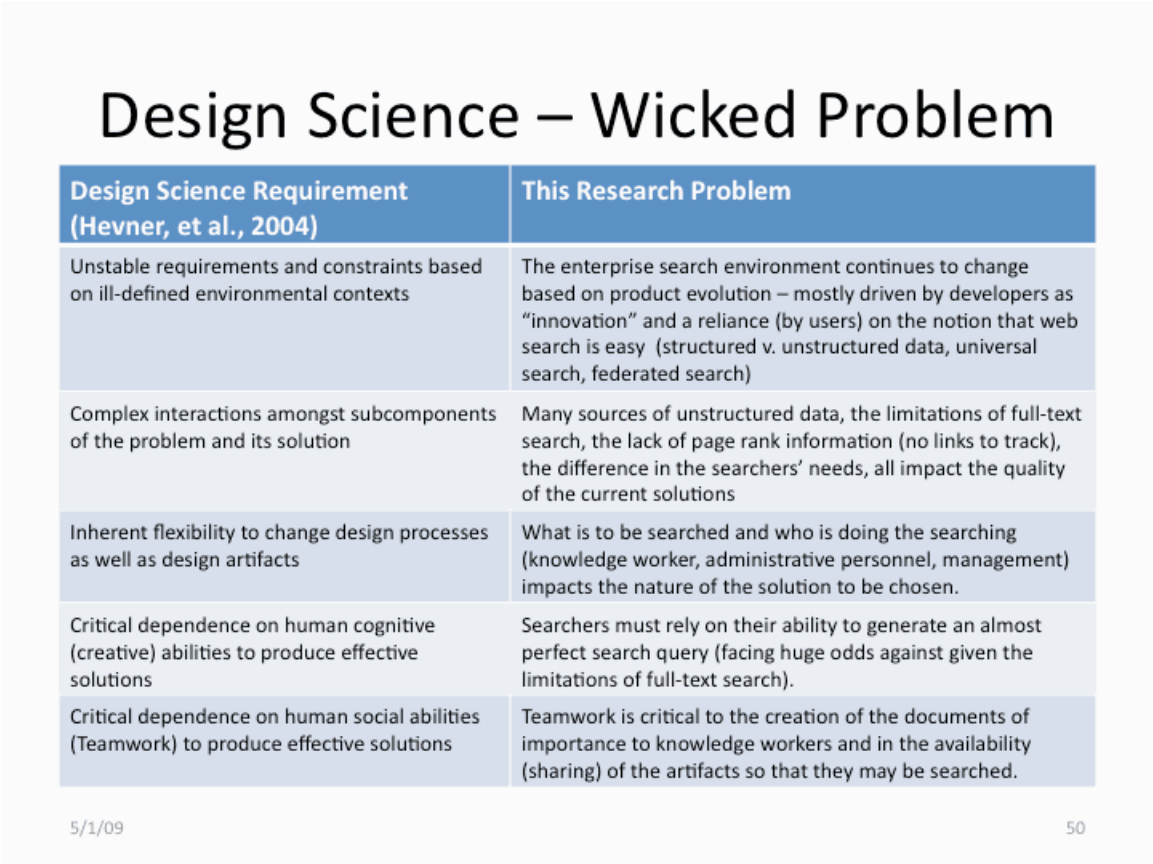


Figure 2 – Enterprise Search as a "Wicked Problem"

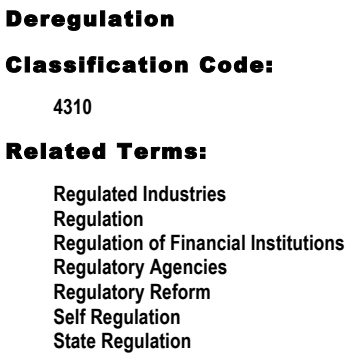
The objective of this research is to test the high-level hypothesis that adding contextual information to a full-text, keyword-based search will reduce the number of irrelevant documents returned to a searcher without negatively impacting the number of relevant documents returned. The literature tells us that a reasonable target for the total number of documents returned would be 20 or fewer; since the typical searcher, enterprise or web, rarely looks beyond the first two pages of results (Jansen et al. 2006; Stenmark et al. 2006), and the typical search engine returns ten results per page. The proposed research will test two hypotheses:

> H1: the number of documents returned will be reduced by the addition of subject metadata to a keyword-based, full-text search.
>
> H2: the precision of the search results will be improved by the addition of subject metadata to a keyword-based, full-text search.

An experiment has been developed and run to test H1 by running randomly generated queries on the ABI/Inform Global Edition Research Database, which is available through many university libraries.

The ABI/Inform research database was chosen as a proxy for enterprise document stores for three reasons. First, it is readily available to other researchers who might want to replicate our results. Second, it represents a large but bounded set of documents that are similar to a large organization's knowledge base of work products. Third, most of the documents in the collection have subject metadata defined, which is a requirement for the proposed research.

The subject thesaurus provided by ABI/Inform is the source for the query terms used in the experiment. The thesaurus provides a controlled vocabulary against which the documents in the collection are indexed. It is unique to this specific collection. Each term in the thesaurus is cross-referenced with associated terms in the thesaurus. Among others, links are provided to more restrictive terms associated with a smaller set of documents, less restrictive terms associated with a larger set of documents, and related terms. Subject matter experts index each document against the subject thesaurus. Articles in the collection are often indexed to several subject terms. Figure 3 is a screenshot of the thesaurus entry for the term deregulation.



Deregulation

Classification Code:

4310

Related Terms:

Regulated Industries
Regulation
Regulation of Financial Institutions
Regulatory Agencies
Regulatory Reform
Self Regulation
State Regulation

Figure 3: Thesaurus Entry Example

The experiment used 384 randomly generated pairs of query terms. The first term in each pair was randomly selected from the list of roughly 17,000 subject terms in the thesaurus. The second term was then randomly selected from the list of related terms for that particular subject term. We chose a related term for the second term in our queries because we believe this most closely approximates search behavior. As searchers work to refine their search query, they rarely would replace a keyword by a broader keyword. They may replace a keyword by a more restrictive keyword, but generally would not keep both keywords in the query. In most instances the second keyword in the query is a related term, either a synonym or a related dimension.

Subject terms in the thesaurus typically consist of more than one word. Examples of subject terms include "knowledge management," "plumbing fixtures," and "consumer attitudes." Figure 3 helps illustrate how the pair of terms was selected for the queries in Table 2. "Deregulation" was randomly selected from the thesaurus, and then "Regulatory Reform" was randomly selected from the seven related terms associated with Deregulation in the thesaurus.

Four queries were run for each of the 384 pairs of terms. The first query (KW1) is a full-text search of the collection using the first term in the pair. The second query (KW1 KW2) uses both terms as keywords in a full-text search. The third query (KW1 SU1) uses the first term in the pair as both a keyword in a full-text search of the collection and as a keyword in a search of the subject metadata field. Thus, the third query looks for the term in both the text of the articles and in the subject field of the metadata. The fourth query (KW1 SU1 SU2) adds the second term to the third query as an additional subject. That is, the fourth query searches the full-text of the documents for the first term in addition to searching the subject metadata for the first and second terms. The first two queries are control scenarios while the third and fourth are treatment scenarios in the comparison of full-text only searches versus combined full-text and subject metadata searches. An example of a set of queries created using this approach is given in Table 2.

These queries were submitted using ABI/Inform's standard search box interface. The system was set to search only those documents in the collection for which a full-text version was available. After each query was submitted, the number of articles returned by the search was recorded. A sample of the query terms and collected data appears in Table 3.

TEXT(Deregulation)
TEXT(Deregulation) AND TEXT(Regulatory Reform)
TEXT(Deregulation) AND SUBJECT(Deregulation)
TEXT(Deregulation) AND SUBJECT(Deregulation) AND SUBJECT(Regulatory Reform)

Table 2. An Example Set of Queries Generated for a Pair of Query Terms

Term 1	Term 2	KW1	KW1 KW2	KW1 SU1	KW1 SU1 SU2
Stock Exchanges	Capital markets	273536	20086	7458	212
Teaching Assistants	Teachers	1823	820	29	6
Employment Interviews	Hiring	817	422	105	15
Business Process Reengineering	Systems Management	14330	1140	1799	16

Table 3. Sample Query Terms and Data

The results of the experiment (Table 4) show a dramatic reduction in result set size between full-text only and full-text plus subject metadata searches. Table 4 shows that we can be 97.5% confident that an order-of-magnitude improvement will occur in at least 97.95% of the searches. We also can be 97.5% confident that a hundredfold improvement will occur in at least 87.71% of the searches. These are very strong results and support further investigation and testing of the hypotheses presented in this proposal.

Improvement	Lower Limit of C.I.	Point Estimate	Upper Limit of C.I.
10 fold	97.95%	98.96%	99.97%
20 fold	96.15%	97.66%	99.17%
100 fold	87.71%	90.62%	93.53%

Table 4. 95% Confidence Intervals for Proportion of Two-Subject Queries Reducing Result Set Sizes by 90%, 95%, and 99%

These findings support the earlier findings of Voorbij (1998) and Gross and Taylor (2005) that metadata has a positive impact on search results, and extends their findings beyond a bibliographic record search to a comparison of metadata search with the full-text search. In contrast, Hemminger et al. (2007) concluded that full-text only search should become the standard. However, they were searching for gene names, which tend to be acronyms (e.g., COMT). With acronyms, the number of appearances in the collection has little effect on the number of meanings the acronym takes on, especially in a focused collection such as a medical library. This is not the case for most searches and explains why their conclusion appears to contradict our initial results.

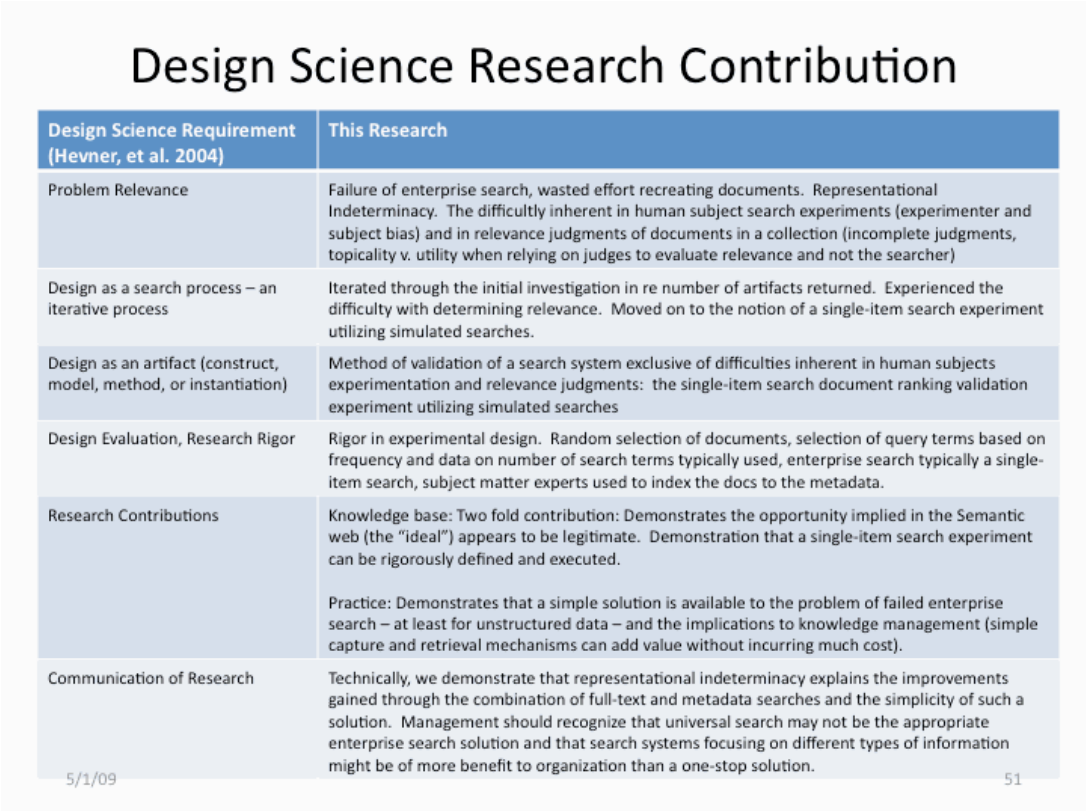
Although these results are impressive, they have very limited conclusive power without a test of H2: an examination of the relevance of the documents returned. Several options for performing such a test are currently being pursued and will be the focus of this dissertation.

Efforts are currently underway to gather real-world queries and query logs to enable a more accurate assessment of the relevance of search results. Two options are being pursued: getting queries from ABI/Inform query logs and building collection and controlled search environment that will provide us with our own query log of queries run by faculty and PhD

students in our IS department. Having real-world queries should allow for a more accurate judgment of the relevance of the search results.

Other options being considered are the running of either a laboratory experiment using the controlled library and search engine mentioned above or a field experiment using students and the ABI/Inform database. In each experiment, subjects will be assigned a topic (possibly from a list of freshman term paper topics) and be assigned to use either the full-text only or the full-text plus metadata interfaces. Once a subject has completed the search to their satisfaction, they will be asked to rank the relevance of the first 20 results their searches returned.

Yet another option exists that eliminates the problem involved in the above-mentioned experimental options: having relevance judgments being performed by humans – either subjects or outside judges. We are working on defining a method for running simulated single-item searches on the ABI/Inform collection. In such an experiment, we will randomly select a number of documents from the collection and then perform a set of simulated searches for each document recording the rank of the document in the results. The expectation is that searches run using the subject metadata search along with the full-text search will result in the document having a higher ranking (preferably showing up in the first two pages of results) than using the full-text only search. The challenge in this experiment will be in defining the query terms to be used so that we do not bias the experiment for or against our expected outcome. This single-item search experiment eliminates the need for relevance judgments to be performed on the documents in the collection and more accurately represents a real-world, enterprise search environment where searchers are usually looking for something specific (Fagin et al. 2003). This experimental method, if successful in answering the research question presented, would be the artifact of interest in the design science research contribution made by this dissertation (Figure 4).



Design Science Research Contribution

Design Science Requirement (Hevner, et al. 2004)	This Research
Problem Relevance	Failure of enterprise search, wasted effort recreating documents. Representational Indeterminacy. The difficultly inherent in human subject search experiments (experimenter and subject bias) and in relevance judgments of documents in a collection (incomplete judgments, topicality v. utility when relying on judges to evaluate relevance and not the searcher)
Design as a search process – an iterative process	Iterated through the initial investigation in re number of artifacts returned. Experienced the difficulty with determining relevance. Moved on to the notion of a single-item search experiment utilizing simulated searches.
Design as an artifact (construct, model, method, or instantiation)	Method of validation of a search system exclusive of difficulties inherent in human subjects experimentation and relevance judgments: the single-item search document ranking validation experiment utilizing simulated searches
Design Evaluation, Research Rigor	Rigor in experimental design. Random selection of documents, selection of query terms based on frequency and data on number of search terms typically used, enterprise search typically a single-item search, subject matter experts used to index the docs to the metadata.
Research Contributions	Knowledge base: Two fold contribution: Demonstrates the opportunity implied in the Semantic web (the "ideal") appears to be legitimate. Demonstration that a single-item search experiment can be rigorously defined and executed. Practice: Demonstrates that a simple solution is available to the problem of failed enterprise search – at least for unstructured data – and the implications to knowledge management (simple capture and retrieval mechanisms can add value without incurring much cost).
Communication of Research	Technically, we demonstrate that representational indeterminacy explains the improvements gained through the combination of full-text and metadata searches and the simplicity of such a solution. Management should recognize that universal search may not be the appropriate enterprise search solution and that search systems focusing on different types of information might be of more benefit to organization than a one-stop solution.

5/1/09 51

Figure 4 – Expected Design Science Research Contribution

Once the concern regarding relevance is addressed, a second question arises regarding the contribution of this research. That question concerns the cost/benefit tradeoff involved in indexing the documents in the enterprise document warehouse. Will the improvement gained via the addition of metadata search be worth the effort required to index the documents?

We will rely on a previously developed model (Corral et al. 2007), combining our experimental results with indexing cost data from ABI/Inform, to compare search costs with and without the metadata search to determine the potential cost savings. We expect to demonstrate that adding metadata search to full-text search can dramatically improve enterprise search without adding additional costs to the enterprise's operations.

CONCLUSION

We expect the proposed research to add to the literature in the area of context-based search in two key ways. First, we expect to demonstrate that the addition of a metadata search reduces the impact of representational indeterminacy (Blair 2002a; Jansen et al. 2006) inherent in the full-text search of a large collection and, therefore, reduces the likelihood that researchers will reach their futility points (Blair 2002b). This will extend the online-catalog-based research in the literature beyond the citation (or bibliographic record) and abstract to include the searching of the full text of the documents in our comparisons. Second, we expect to successfully demonstrate a unique experimental methodology (the random generation of queries from a controlled vocabulary) to a research question that has yet to be completely answered.

REFERENCES

1. Alavi, M., and Leidner, D.E. (2001) Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues, *Mis Quarterly,* 25, 1, 107-136.
2. Berners-Lee, T., Hendler, J., and Lassila, O. (2001) The Semantic Web, *Scientific American,* 284, 5, 34-43.
3. Blair, D.C. (2002a) The Challenge of Commercial Document Retrieval, Part I: Major Issues, and a Framework Based on Search Exhaustivity, Determinacy of Representation, and Document Collection Size, *Information Processing and Management,* 38, 2, 273-291.
4. Blair, D.C. (2002b) The Challenge of Commercial Document Retrieval, Part II: A Strategy for Document Searching Based on Identifiable Document Partitions, *Information Processing and Management,* 38, 2, 293-304.
5. Blair, D.C. (2006) The Data-Document Distinction Revisited, *The DATA BASE for Advances in Information Systems,* 37, 1, 77-96.
6. Blair, D.C., and Maron, M.E. (1985) An Evaluation of Retrieval Effectiveness for a FGull-Text Document-Retrieval System, *Communications of the ACM,* 28, 3, 289-299.
7. Corral, K., Schuff, D., St. Louis, R.D., and Turetken, O. (2007) "A Model for Estimating the Savings from Dimensional Versus Keyword Search," in: *Advanced Principles for Improvin Database Design, Systems Modeling, and Sotware Development,* K. Siau and J. Erickson (eds.), IGI Global, 2007.
8. Drabenstott, K.M. (2004) Information Retrieval Systems for End Users: Primetime Players that Just Don't Make the Grade, *Journal of Education for Library and Information Science,* 45, 2, 173-177.
9. EContent (2004) Getting Just What You Need, *EContent,* 27, 7/8, S12-S13.
10. Fagin, R., Kumar, R., McCurley, K.S., Novak, J., Sivakumar, D., Tomlin, J.A., and Williamson, D.P. (2003) Searching the Workplace Web, in *12th International Conference on International World Wide Web*, Budapest, Hungary, Association for Computing Machinery.
11. Gardner, W.D. "Most Users Are Unhappy With Enterprise Search," 2008, http://www.informationweek.com/news/software/database/showArticle.jhtml?articleID=207100963, Accessed 7/24/2008
12. Hendler, J. (2001) Agents and the Semantic Web, *IEEE Intelligent Systems,* 16, 2, 30-37.
13. Hevner, A.R., March, S.T., Park, J., and Ram, S. (2004) Design Science in Information Systems Research, *Mis Quarterly,* 28, 1, 75-105.
14. Jansen, B.J., and Spink, A. (2006) How are we Searching the World Wide Web? A Comparison of Nine Search Engine Transaction Logs, *Information Processing and Management,* 42, 1, 248-263.
15. KMWorld "Developing a Universal Search Strategy (Hint: Start with Usability)," 2008, http://www.kmworld.com/Webinars/90-Developing-a-Universal-Search-Strategy-(Hint-Start-with-Usability).htm, Accessed 1/16/2009
16. Kontzer, T. "Search On," in: *InformationWeek*, 2003, pp. 30-36.

17. Raghavan, P. (2001) Structured and Unstructured Search in Enertprises, *Bulletin of the IEEE Computer SOciety Technical Comittee on Data Engineering,* 24, 4, 15-18.
18. Spink, A., Wolfram, D., Jansen, M.B.J., and Srarcevic, T. (2001) Searcing the Web: The Public and Their Queries, *Journal of the American SOciety for Information Sience and Technology,* 52, 3, 226-234.
19. Stenmark, D., and Jadaan, T. (2006) Intranet Users' Information-Seeking Behavior: An Analysis of Longitudinal Search Log Data in *Proceedings of the American Society for Information Science and Technology*, Austin, TX.
20. Storey, V.C., Burton-Jones, A., Sugumaran, V., and Purao, S. (2008) CONQUER: A Methodology for Context-Aware Query Processing on the World Wide Web, *Information Systems Research,* 19, 1, 3-25.
21. Sure, Y., Hitzler, P., Eberhart, A., and Studer, R. (2005) The Semantic Web in One Day, *IEEE Intelligent Systems,* 20, 3, 85-87.
22. Ultraseek "Business Search vs. Consumer Search: Five Differences Your Company Can't Afford to Ignore," Autonomy, 2006.
23. Wu, Y.-f.B., and Li, Q. (2008) Document Keyphrases as Subject Metadata: Incorporating Document Key Concepts in Search Results, *Information Retrieval,* 11, 3, 229-249.
24. Zipf, G.K. (1949) Human Behavior and The Principle of Least Effort: An Introduction to Human Ecology Addison-Wesley Press, Inc., Cambridge, MA.