December 2006

Improving Document Retrieval through a Browsable Interface: The Dimensional Document Store

Gregory Schymik
Arizona State University

Karen Corral
Arizona State University

David Schuff
Temple University

Robert St. Louis
Arizona State University

Improving Document Retrieval through a Browsable Interface: The Dimensional Document Store

Gregory Schymik
Arizona State University
Gregory.Schymik@asu.edu

Karen Corral
Arizona State University
Karen.Corral@asu.edu

David Schuff
Temple University
David.Schuff@temple.edu

Robert St. Louis
Arizona State University
stlouis@asu.edu

ABSTRACT

The corporate world must take advantage of the trillions of dollars worth of intangible assets (in the form of knowledge) it possesses. Specialized Knowledge Management Systems (KMSs) are being developed to leverage these assets, and a critical issue is finding the most effective way for users to retrieve needed information from organizational knowledge stores. Research has detailed the problems inherent in using keyword searches, as well as the use of visual knowledge hierarchies to overcome some of these problems. This paper describes an experiment that will extend that research by addressing whether it is easier to find information in a large knowledge artifact store through a "browseable" dimensional document store or by performing a full-text search.

KEYWORDS

Information retrieval, dimensional database, user interface design, information presentation

INTRODUCTION

The *February 2002 Intangible Management Value Survey* revealed that almost 70% of the total value of the 500 largest firms in the United States could be tied to the value of the firms' intangible assets. These assets represent approximately US$7.3 trillion (Stone, 2002). Unfortunately, many firms are not taking full advantage of these assets. They reside in file cabinets, on individual hard drives or servers, or on backup tapes. Even though these knowledge artifacts exist, the ability to retrieve the relevant artifacts is lacking. Since roughly 80% of organizational data is not stored in systems that allow for easy search and retrieval (Olsen, 2003), this represents a significant potential loss to organizations. It is imperative that organizations deploy systems to manage and retrieve this information.

Companies such as Google and Ask.com, are now selling enterprise search applications (Callaghan and McCright, 2002). However, these systems suffer from fundamental problems inherent in web search tools: the quality of their results is limited by the quality of the query terms input by the user. More specifically, they are limited by the users' ability to recall the query terms that will provide the best results.

Once users recognize the correct terms for a search, they are frequently faced with daunting result set sizes of thousands or even millions of records. This is because words can have many different meanings (Zipf, 1949). For example, a user wishing to find information on how to perform a regression analysis is not going to be helped by the large number of articles that have used regression as a methodology. Rather, that user needs only articles where the subject is regression. By categorizing knowledge artifacts into categories, or dimensions, the meaning of terms can be limited.

A system that helps users recall the correct terms and create the best query should improve the quality of the search results and reduce the amount of time spent searching. This should reduce the workload on the user and improve overall intention to use the search system. We propose that a "dimensional document store" will achieve this goal. Much like a dimensional data cube, a dimensional document store allows the user to construct queries based on browsing possible terms. Also like a dimensional data cube, a dimensional document store has a specific domain and thereby uses categories that are relevant to

the specific user of that document store. This paper presents the design of an experiment to test a dimensional search interface.

LITERATURE REVIEW

Commercial search engines can quickly perform full-text searches on millions of documents. Prior to this, searches were performed on only portions of documents (e.g., the title or the abstract). That limited users' ability to find relevant documents as they had to think of the exact term used within the title or the abstract. Full-text search products increase the likelihood that users will think of a word used in a document, however this has its own problems. The search term may be so common that the result set is overwhelming. For example, a Google search of "winter Olympics" returns over 73 million results.

The Act-R theory of memory storage states that memory is stored in chunks connected through links of differing strength (Andersen et al., 1998). Once a chunk has been activated in memory (by seeing the term that triggers that chunk), it is easier to follow links associated with that chunk. However, if a chunk is not activated, humans have limited ability to recall "from scratch." For example, if asked to recall the participant names for an AMCIS mini-track, most would struggle to recall more than a few names. However, if shown a list of information systems researchers, most would correctly identify the mini-track participants. Andersen et al. (1998) demonstrated empirically that subjects' ability to recall terms is much lower than their ability to recognize terms when presented with a list of both relevant and irrelevant terms.

LaBrie (2004) incorporated a mechanism to facilitate recognition into a search tool through a tree-view hierarchy interface. Subjects given the interface realized an "over 50% gain in retrieval accuracy over a traditional keyword search mechanism" when using a single search category (p.125). However, a hierarchy of all terms used in a library of knowledge artifacts can become prohibitively long, leaving users unable to take advantage of recognition. Therefore, the list needs to be constrained in a meaningful manner.

Classifying documents along multiple dimensions is one way of providing these constraints. This is because language is inherently ambiguous. Zipf (1949) demonstrated that the more times a term was used the greater the number of meanings it would have. This ambiguity can lead to irrelevant results in searches. For example, a search on "sailing" might retrieve the article containing this sentence: "This is not to suggest that the U.S. markets will be smooth sailing" (Browning, 2006). This would not happen if the document were classified by a category such as "subject." Documents can be classified along multiple attributes (such as date and author) in addition to subject. One could then browse those attributes, constraining their search along specific attribute values. The result is very similar to a data cube, where queries are performed by "browsing" attribute values for one or more dimensions (the aggregation of similar attribute values). Tseng and Chou (in press) presented a model for a dimensional document store. Although they implemented an OLAP interface, they did not compare its performance to existing search alternatives such as full-text search.

To that end, this research extends previous research by building a system that combines the concepts of hierarchies (to provide an intuitive structure for a set of related keywords) into a dimensional document store (to facilitate the browsing of those keyword sets). This paper describes an experiment to compare the performance of this search interface to an interface which employs full-text search.

PROTOTYPE DEVELOPMENT

To test the efficacy of a dimensional document store, we developed a prototype system. The questionnaires, search interface, and surveys were developed using Visual Basic for Applications within Microsoft Excel. The back end containing the dimensional information (the document metadata) and links to the documents (contained on a network server) is implemented in Microsoft SQL Server.

The interface consists of two parts: a search criteria entry form and form to review the search results. The user will locate documents by traversing keyword hierarchies within each dimension (see Figure 1). Table 1 shows the searchable dimensions defined for this document store.

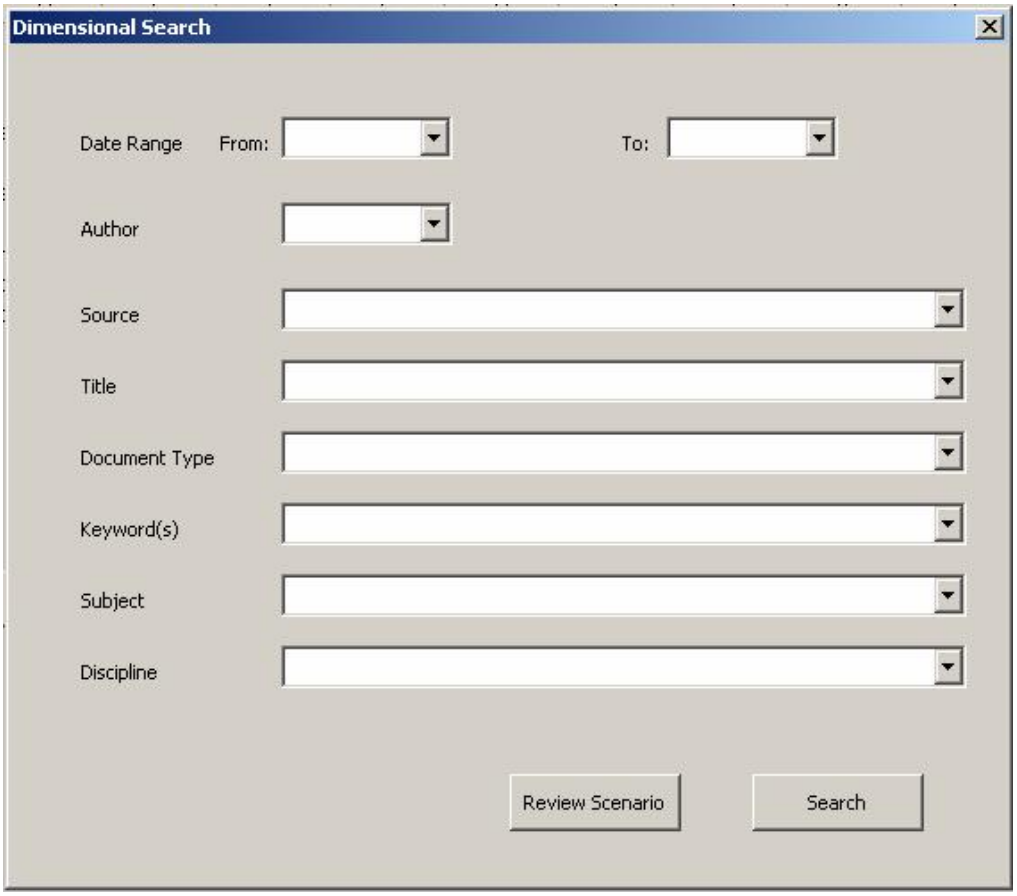

Figure 1: Input form for dimensional document store interface.

Date of publication	Document type (e.g., article, web page)
Author(s)	Keyword(s) (ACM Computing Classification System)
Source (e.g., journal name)	Subject (from Library of Congress)
Title	Discipline

Table 1: Attributes used to classify the documents

Users may search any combination of dimensions. The selection of the attribute values for each dimension used in the search is performed through the use of hierarchies. An example hierarchy is shown in Figure 2.

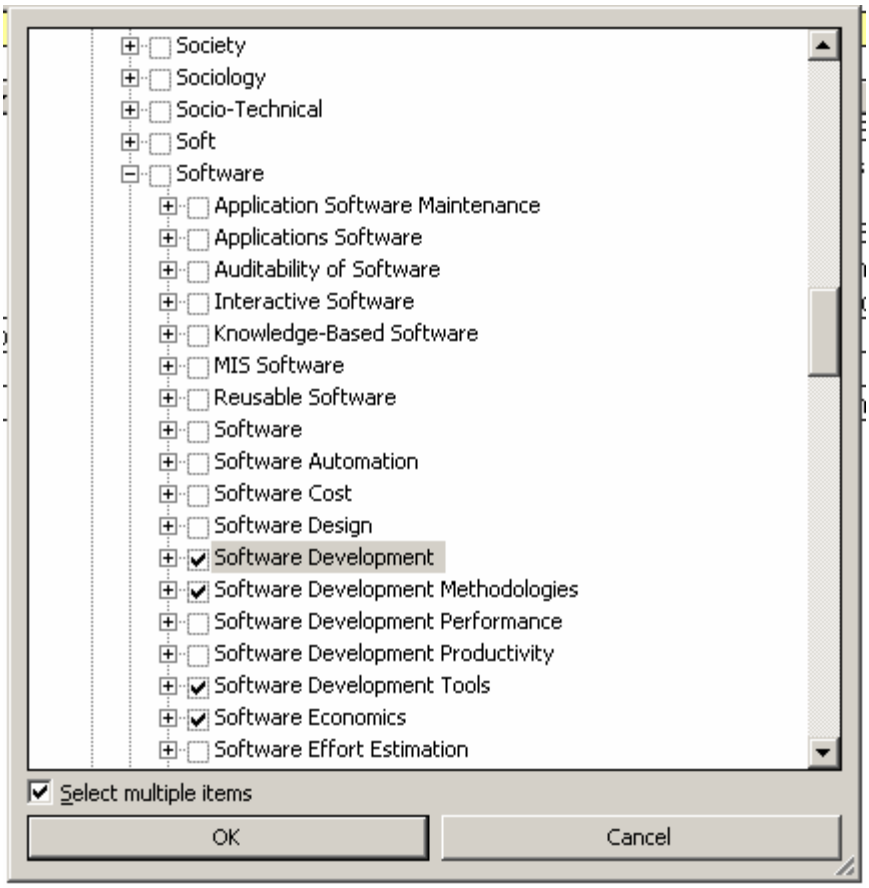

Figure 2: Visual Information Hierarchy Example

Once a search is completed, users can review the results (see Figure 3) and select which documents they wish to keep in their final results set.

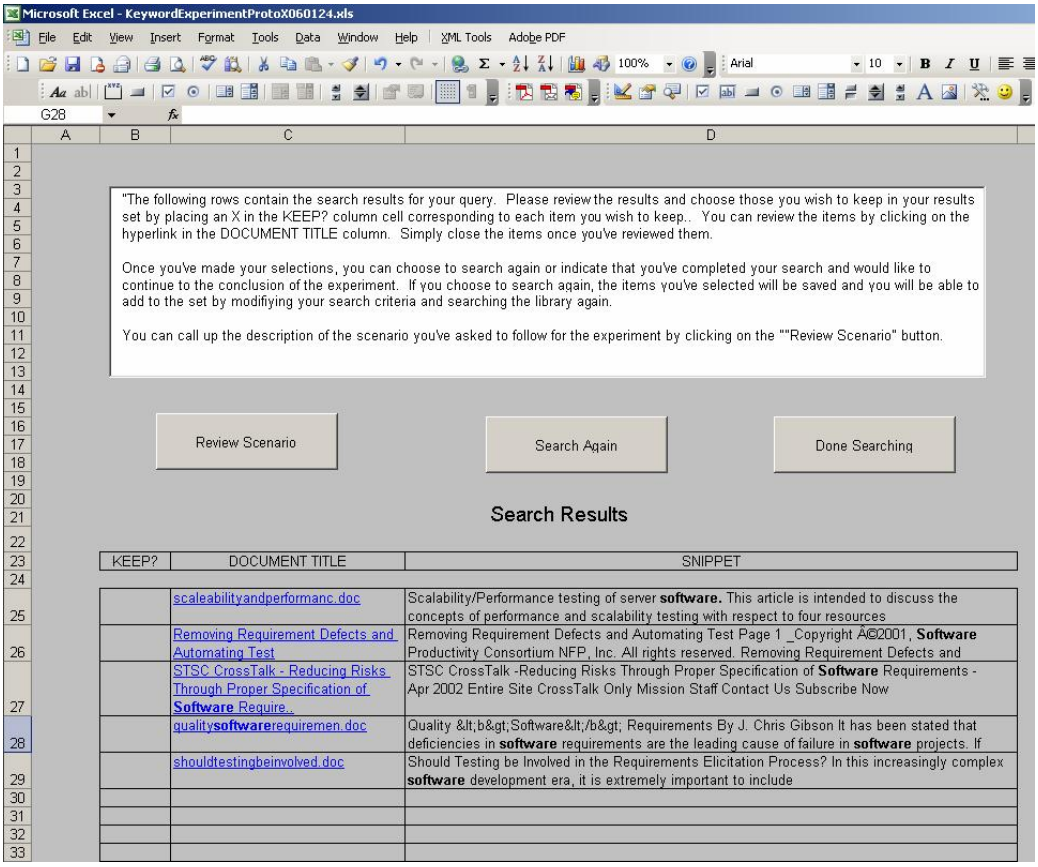

Figure 3: Form for reviewing search results.

After selecting the documents the user wishes to keep, they can either return to the keyword hierarchy form to continue searching or end their search.

HYPOTHESIS DEVELOPMENT

To measure the effectiveness of the prototype, we will compare its performance to a full-text search tool using several measures. One measure of search quality is the precision of the result set. Precision is the number of relevant documents returned divided by the total number of documents in the result set (Raghavan, Bollmann et al., 1989). The use of defined keyword lists organized into dimensions and hierarchies should result in reduced ambiguity, which should increase precision. This leads to the following hypothesis:

H1a: Searches performed using the dimensional search interface will yield more precise results than those performed using the full-text search interface.

A second measure of search quality is the accuracy of the result set. Accuracy is the number of relevant documents in the result set divided by the total number of relevant documents in the search space. The use of multiple dimensions should allow searchers to locate a relevant document from multiple perspectives. This should increase the likelihood that a relevant document is found and leads to the second hypothesis:

H2a: Searches performed using the dimensional search interface will yield more accurate results than those performed with the full-text search interface.

Although the dimensional document store interface should result in more accurate and more precise results, the effect on task time is unclear. The reduction in frustration from using the more intuitive interface, and the positive reinforcement from finding relevant articles, may cause individuals to increase the amount of time that is spent searching for relevant articles using a dimensional interface (instead of giving up in frustration). Because theory provides little guidance, we take an exploratory approach and state our time hypothesis in the null form.

H3: There will be no difference in the amount of time taken for searches performed using the dimensional search interface and the full-text search interface.

As with time, requiring less effort is desirable, but, theory provides little guidance regarding the effect on effort of using the dimensional interface. The effort saved by browsing keywords instead of recalling them may be offset by additional effort users expend because they believe they will be able to ultimately find what they are looking for. Therefore, we again take an exploratory approach and state the following hypothesis:

H4: There will be no difference in the amount of work for searches performed using the dimensional search interface and the full-text search interface.

Finally, intention to use a system is a critical measure of its success and a predictor of actual system usage (Venkatesh, 2000). We expect that the dimensional search engine will achieve better results. This should lead to the perception that the dimensional interface is more useful than its full-text alternative. Moreover, the intuitiveness of browsing hierarchical dimensions (instead of recalling search terms) should lead to the perception that the dimensional interface is easier to use. This leads to hypothesis 5:

H5a: Users will have a higher intention to use the dimensional interface compared to users of the full-text interface.

EXPERIMENTAL PROCEDURE

A controlled experiment will be run using student subjects to compare searching a document store using a dimensional interface with searching that same document store using a full-text search interface. Both groups will use a tool specifically developed for this study. Subjects will play the role of business analysts charged with improving their firms' requirements specification process. They will be asked to search a library of software process improvement literature for artifacts that will help them learn to prepare good requirements specifications. The library is limited to documents regarding software development taken from practitioner journals.

A separate search interface has been developed for each experimental condition. The treatment (the dimensional interface) is the interface described earlier in this paper. The control (the full-text search) is Excel-based with a wrapper around the "Google Desktop" search tool. As with the dimensional interface, the full-text interface consists of a search criteria entry form and a results reviewing form. The full-text search input form is shown in Figure 4. Users may search using single or multiple terms or phrases. They can also exclude documents containing terms from the result set. The Google desktop search tool is the search engine for this interface.

Demographic information will be collected from the subjects. The prototype systems will collect the time the user searches, the number of queries run, the documents returned in each search, and the documents selected for the final solution set. This data will be used to compute the accuracy and precision of the search results. Immediately after completing the task, subjects will complete the NASA/TLX mental effort questionnaire (Hart and Staveland, 1988) and an instrument to measure intention to use.

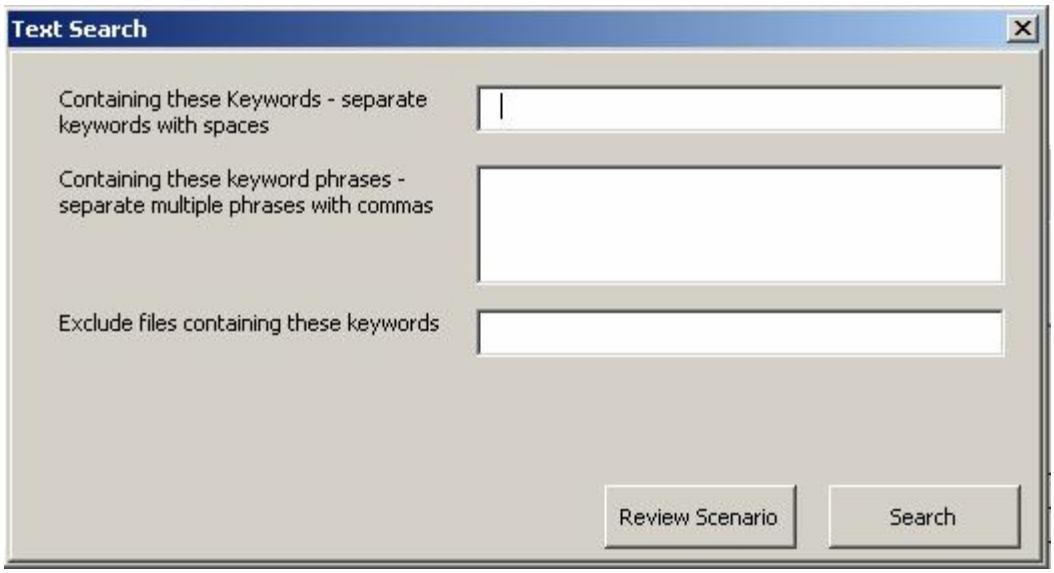

Figure 4: Form for entering search term in full-text search interface.

EXPECTED CONTRIBUTIONS

This paper outlines a study to develop and test a dimensional search interface as an alternative to full-text search. This experiment will demonstrate the value of leveraging recognition of search terms through dimensions for the retrieval of knowledge artifacts, allowing us to gain an understanding of the effort of this interface on the quality of search results and the effort required to achieve those results. The results of this study will be useful to both researchers and practitioners. For researchers, this study integrates literature in linguistics and cognition to present a rationale for the use of recognition-based document retrieval systems and empirically tests that claim through a rigorous, controlled experiment. For practitioners, this study serves as a "proof of concept" for applying a dimensional approach to how documents are stored and retrieved.

REFERENCES

1. Anderson, J.R., D. Bothell, et al. (1998). An Integrated Theory of List Memory. *Journal of Memory and Language* 38(4): 341-380.
2. Browning, E.S. (2006). Dow Industrials End Above 11000 As Stocks Extend New Year's Rally. *Wall Street Journal*, Jan 10. p. A.1.
3. Callaghan, D. & McCright, J.S. (2002). Taking a Swing at Enterprise Search. http://www.eweek.com/article2/0,1895,244925,00.asp
4. Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P.A. Handcock & N. Meshkati (eds.), *Human mental workload* (pp. 139-183). Amsterdam: North Holland.
5. Labrie, R.C. (2004). The Impact of Alternative Search Mechanisms on the Effectiveness of Knowledge Retrieval. Arizona State University.
6. Olsen, J.E. (2003). Data Quality: The accuracy dimension. Morgan Kaufmann: San Francisco, CA.
7. Raghavan, V., P. Bollmann, Jung, G.S. (1989). A critical investigation of recall and precision as measures of retrieval system performance *ACM Transactions on Information. Systems* 7(3): 205-229.
8. Stone, Peter (2002). Intangible Management: What is it, Really? http://www.refresher.com/!psintangible.html.
9. Tseng, F.S.C., Chou, A.Y.H. The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence. *Decision Support Systems, In Press, Corrected Proof.*
10. Venkatesh, V. (2000). Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. *Information Systems Research* 11(4): 342-365.
11. Zipf, G.K. (1949). Human behavior and the principle of least effort. Cambridge,MA: Addison-Wesley.