Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2009 Proceedings

Americas Conference on Information Systems (AMCIS)

2009

Order of Magnitude Reductions in the Size of Enterprise Search Result Sets Through the Use of Subject Indexes

Gregory Schymik
Arizona State University, gschymik@asu.edu

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

Karen Corral
Boise State University, karencorral@boisestate.edu

Order of Magnitude Reductions in the Size of Enterprise Search Result Sets Through the Use of Subject Indexes

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

Robert St. Louis
Department of Information Systems
W. P. Carey School of Business
Arizona State University
st.louis@asu.edu

Karen Corral
Department of Information Technology and Supply Chain Management,
College of Business and Economics
Boise State University
karencorral@boisestate.edu

ABSTRACT

Keyword search has failed to adequately meet the needs of enterprise users. This is largely due to the indeterminate nature of languages. We argue a different approach needs to be taken, and draw on the success of previous library indexing concepts to propose a solution. We test our solution by performing search queries on a large research database. By incorporating readily available subject indexes into the search process, we obtain order of magnitude improvements in the performance of search queries. Our performance measure is the ratio of the number of documents returned without using subject indexes to the number of documents returned when subject indexes are used. We explain why the observed tenfold improvement in search performance on our research database can be expected to occur for searches on a wide variety of enterprise document stores.

Keywords

Enterprise search, Metadata, Representational indeterminacy, Orderly distribution of meanings, Full-text search

INTRODUCTION

Individuals in all sectors of the economy struggle to find the documents for which they search. From the perspective of the firm, these struggles equate to employees' time wasted on ineffective enterprise search. Researchers have tried to apply the principles learned in the successes achieved searching the internet to searching enterprise document stores. This requires a shifted focus from the open internet to a closed system incorporating the firm's network (intranet, servers, email systems, etc.) and a search engine. Depending on the scale, the search engine was either a single piece of software or a larger "enterprise search appliance" that combined search engine technology and a customized server designed to host the technology. Since these search engines were so successful searching the web, it was reasoned that they would be equally successful searching these enterprise knowledge bases. Unfortunately, they failed to produce the desired results (Gardner, 2008, Kontzer, 2003)

Hammer's (1990) seminal article on business process reengineering (BPR) titled "Don't Automate, Obliterate" argues to truly take advantage of information technology, processes must be completely redesigned and not simply automated. This concept works well with many existing business processes and arguably has been the biggest factor driving the massive productivity gains seen in our economy over the past two decades. However, rules often have exceptions, and enterprise search may be one of the exceptions to which Hammer's recommendation does not apply.

For enterprise search, Hammer's recommendation has been implemented by replacing libraries' author, subject, and title indexes with conveniently available and heavily marketed full-text search appliances. Such obliteration has not provided the

expected improvement in productivity. This article explains why full-text search alone cannot yield the results sought by enterprise searchers, and demonstrates the order of magnitude improvements that can be obtained through the incorporation of subject indexes into the search process.

The remainder of this paper is organized as follows. First we describe the failure of enterprise search, and our initial research into a potential solution. We then explain the information science behind our proposed solution and present the results of an experiment comparing different types of enterprise search. The paper ends with a discussion of the limitations of the research, and suggestions for future research.

ENTERPRISE SEARCH *DIS*-SATISFACTION

Search tools that use full-text, keyword-based search engines have failed to meet the needs of enterprise users (Alavi and Leidner, 2001, Gardner, 2008). This is particularly true for workers performing knowledge intensive tasks (Kontzer, 2003). The reasons behind these failures range from the technologies being used to the expectations and behaviors of those performing enterprise search (Fagin, Kumar, McCurley, Novak, Sivakumar, Tomlin and Williamson, 2003, Raghavan, 2001). With our ongoing research, we add another cause for the failure of enterprise search: the obliteration and reengineering of a process that should have been automated.

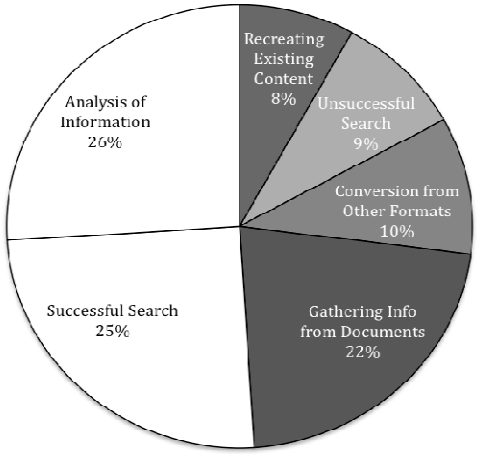

Figure 1. Impact of Ineffective Search on Knowledge Workers (adapted from KMWorld (2008))

In a recent webinar, Google presented data indicating that knowledge workers are wasting almost half of their time as a direct result of failed searches (see Figure 1). They also noted that middle managers spend approximately 25% of their time searching for information necessary to the successful completion of their jobs, and that the information they find is often wrong. Fully 86% of enterprise searchers report being unsatisfied with their firms' enterprise search capabilities (KMWorld, 2008). Google estimates that unsuccessful searches 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).

BUSINESS PROCESS REENGINEERING AND THE ENTERPRISE SEARCH PROCESS

Hammer (1990) argued that simply automating business processes through the application of IT to the problem is wasting the opportunity provided by IT. Processes should be reengineered so that they take advantage of the tremendous productivity gains usually thought to accompany the introduction of IT. For example, instead of reducing headcount in an accounts receivable department by 20% by automating the process, reduce it by 80% by rethinking the entire process (Hammer, 1990).

Before the advent of enterprise search appliances, reports were drafted using pen and paper by knowledge workers (well before the term existed), typed up by administrative personnel, and copies were kept in records storage facilities along with associated work-products. When such documents were needed, corporate librarians retrieved them. These documents were manually indexed using something similar to a standard library's author and subject indexes (extended with vocabularies appropriate to corporate needs), and retrieved when needed by browsing those indexes. The introduction of information technology has dramatically changed the process.

Indexes Replaced by Keyword Search

As information technology has been applied to libraries, the author and subject indexes, although still in existence, have been used less and less. The browsing of these indexes has been replaced by keyword searches of the indexes as stored in the bibliographic records associated with each document in the library. More recently, the ability to perform keyword based, full-text searches of libraries was added to the set of search tools. This has resulted in a successful search being dependent on the searcher's ability to recall specific keywords associated with the sought-after document.

With the advent of corporate networks and intranets and the emergence and popularity of internet search websites, organizations began adopting the internet search engine as the primary technology for corporate knowledge retrieval. Now, an organization can plug in a customized server programmed with crawlers, indexing tools, and a search engine (i.e. an enterprise search appliance), spend a few minutes configuring the device, and within days can have a search box on every desktop in the company that allows employees to search for any document on the corporate network and intranet. These enterprise search appliances seemed too convenient to pass up. We argue that by obliterating the more traditional approach to archive management, corporations have introduced tools destined to dissatisfy their users. The solution lies in taking a step backwards and applying our knowledge of language and the behavior of searchers to the automation of the traditional search process.

SEARCH BEHAVIOR, FUTILITY POINTS, AND THE ORDERLY DISTRIBUTION OF MEANINGS

Enterprise searchers are usually interested in finding a specific document that they know, or strongly suspect, exists (Fagin et al., 2003). When using a full-text, keyword-based search engine, the enterprise searcher faces sifting through possibly thousands of returned results to find the desired document. Usually searchers give up before putting in that effort: they reach their *search futility point* (Blair, 2002b). On reaching that point, they choose to recreate the document instead of continuing their search.

Blair defines two futility points with search. The *anticipated futility point* represents the largest number of documents through which a searcher is willing to begin searching, while the *search futility point* is the number of documents through which a searcher actually looks before giving up (assuming that search is an iterative process and that a searcher will make several attempts at retrieving relevant documents). These two futility points determine the number of irrelevant documents through which a searcher is willing to look (Blair, 2002b).

An effective enterprise search tool would improve on the typical full-text, keyword-based search appliance by returning a set of results to a query that greatly reduces, if not eliminates, the risk of exceeding either of the searcher's futility points. Moreover, it must do this without reducing, and preferably increasing, the likelihood that the target document will be in the result set.

A major problem with full-text, keyword-based searches is that a word can have many meanings. Zipf (1949) studied the characteristics of language, and found that the number of meanings a word takes on in a collection is approximately equal to the square root of the number of times the word appears in that collection (Zipf, 1949). Zipf calls this the orderly distribution of meanings. As a collection of documents gets larger, the number of occurrences of a given word will get larger and the number of meanings that word takes on will get larger. As a word takes on more meanings in a collection, a keyword used to search that collection will lose its ability to retrieve only those documents relevant to the searcher. An example is the word "bank": is the searcher looking for articles describing erosion of the Mississippi River's banks, articles on banking in the financial industry, or the degree of banking in the turns at the Indianapolis Motor Speedway? The search engine cannot make distinctions between such differences in meaning.

The complexity of language makes the construction of a near-perfect search query nearly impossible for any relatively large collection of documents (Zipf, 1949). To be successful, searchers must define a set of keywords to form a query that is capable of doing two things: 1) the query keywords must adequately describe the documents for which they are searching; and 2) the query keywords must be able to discriminate between the relevant and irrelevant documents in the collection (Blair, 2002a).

Search futility points and the orderly distribution of meanings naturally conflict. To avoid a search futility point, result sets need to be kept small. But as collections get large, the orderly distribution of meanings leads to larger result sets for a given set of search terms. For successful search, a means of resolving this conflict must be found.

ADDING CONTEXT TO THE SEARCH: FROM LIBRARY INDEXES TO MODERN METADATA

One solution to the conflict between search futility points and the orderly distribution of meanings is the addition of contextual information to the documents in the collection. If a searcher were able to utilize contextual information in a search, the searcher could reduce the number of possible meanings for a keyword. For the "bank" keyword example above, using contextual information to limit the search to documents that reference financial subject matter will eliminate most references to the Mississippi River or the Indianapolis Motor Speedway. By adding contextual information to the query we do not harm the keyword's ability to describe the documents being sought by the searcher, but we significantly increase the keyword's ability to discriminate between relevant and irrelevant documents.

Adding contextual information to a query should eliminate documents that contain the keywords sought but have different meanings than the one intended by the searcher. The result is an increase in the number of relevant documents returned as a percentage of the total number of documents returned (this is referred to as the precision of the search).

Eliminating documents that contain the desired keywords possessing unintended meanings for those keywords will result in smaller result sets; and smaller result sets should reduce the risk of searchers hitting their *anticipated futility point* and abandoning the search. Reducing this risk will reduce the amount of time wasted on failed enterprise searches.

Assuming that the contextual information stored with the documents is accurate, the discriminatory power of the search should be increased. Documents that meet the contextual criteria specified by the searcher are more likely to be relevant to the search than those not meeting the contextual criteria. Therefore, adding contextual information to the search will decrease the number of irrelevant documents without decreasing the number of relevant documents in the result set. This increases the likelihood of a successful search. Given what has been stated about the costs of failed enterprise search (EContent, 2004, KMWorld, 2008, Ultraseek, 2006), a system that can increase the likelihood that a search will be successful could potentially save corporations billions of US dollars.

The original indexing systems used by librarians arranged bibliographic records by author, title, and subject (Library of Congress, 2005). These systems were quite effective at guiding library patrons to the items for which they were searching. Since a full-text rendering of a document will contain the title and the author(s) of the document, we suggest that subject metadata be used as contextual information to augment the full-text, keyword-based search in the enterprise environment. Most electronic authoring tools (word processors, drawing/design tools, email systems, etc.) allow for metadata tags to be easily set by accessing the document properties, and most search engines allow for the searching of metadata through advanced search interfaces.

METADATA RESEARCH

Research into the impact metadata has on search results has produced varied 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 (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, Saelim, Sullivan, and Vision (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 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, Saelim, Sullivan and Vision, 2007).

Our research counters these findings by demonstrating how metadata reduces the impact of representational indeterminacy (Blair, 2002b, Jansen and Spink, 2006) and, therefore, reduces the likelihood that researchers will reach their futility points (Blair 2002b). If searchers, particularly in the enterprise context, are presented a smaller result set, they are more likely to take the time to review the results and not give up on the search.

STUDY

Our objective is to find a method of reducing the number of irrelevant documents returned to a searcher without negatively impacting the number of relevant documents returned. Research 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 and Jadaan, 2006), and the typical search engine returns ten results per page. We show that incorporating metadata into the search environment will achieve this. Our hypothesis is that the number of documents returned will be reduced by the addition of subject metadata to a keyword-based, full-text search. We test this hypothesis by running randomly generated queries on the ABI/Inform Global Edition Research Database, which is available through many university libraries.

We selected the ABI/Inform research database as a proxy for enterprise document stores for two 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. Although some documents are posted prior to being indexed, all but the most recently entered documents in the ABI/Inform database have subject metadata defined, which is a requirement for the solution we propose.

The subject thesaurus provided by ABI/Inform was 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 narrower terms (more restrictive terms associated with a smaller set of documents), broader terms (less restrictive terms associated with a larger set of documents), and related terms (a set of suggested similar terms). Figure 2 is a screenshot of the thesaurus entry for the term deregulation. Subject matter experts are used by ABI/Inform to index each document against the subject thesaurus. Articles in the collection are often indexed to several subject terms.

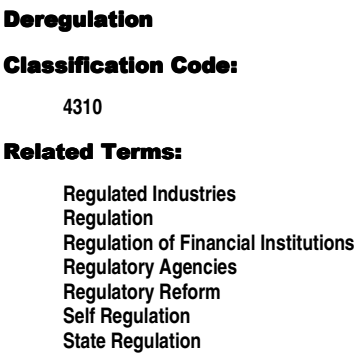

Figure 2: Thesaurus Entry Example

We randomly selected 384 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 consisted of more than one word. Examples of subject terms include "knowledge management," "plumbing fixtures," and "consumer attitudes." Figure 2 illustrates how the pair of terms was selected for the queries in Table 1. "Deregulation" was randomly selected from the thesaurus, and then "Regulatory Reform" was randomly selected from the list of 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 is a full-text search of the collection using the first term in the pair. The second query uses both terms as keywords in a full-text search. The third query 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 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 our 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 1.

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

Table 1. An Example Set of Queries Generated for a Pair of Query Terms

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 2.

Term 1	Term 2	KW1	KW1 KW2	KW1 SU1	KW1 SU1 SU2
Stock Exchanges	Capital markets	273,536	20,086	7,458	212
Teaching Assistants	Teachers	1,823	820	29	6
Employment Interviews	Hiring	817	422	105	15
Business Process Reengineering	Systems Management	14,330	1,140	1,799	16

Table 2. Sample Query Terms and Data

ANALYSIS

The descriptive statistics for the queries run using the 384 pairs of terms appear in Table 3. The means show a dramatic decrease in the number of documents returned when a search of subject metadata is added to the full-text searches. The average number of documents returned decreased by 97.5% (from 41,323 documents to 1,043 documents) when we required that the first search term appear in both the text of the document and the subject field of the metadata (KW1 SU1 -- keyword one in text of document and keyword one in subject metadata). When we added a second search term and required that it also appear in the subject field of the metadata (KW1 SU1 SU2 -- keyword one in text of document and keyword one in subject metadata and keyword two in subject metadata), the average number of documents returned decreased by 99.9% (from 41,323 documents to 43 documents).

Query Type	Mean	Std. Dev.	Median	Lower Quartile	Upper Quartile
KW1	41,323	146,772	3,045	666	21,598
KW1 KW2	2,814	9,780	194	33	1,022
KW1 SU1	1,043	4,601	56	10	426
KW1 SU1 SU2	43	219	2	0	11

Table 3. Descriptive Statistics – Number of Articles Returned

The average number of documents returned in the KW1 SU1 SU2 scenario was greater than 20. Averages, however, can be greatly influenced by outliers. A better measure of the effectiveness of using subject metadata is the percent of queries that returned 20 or fewer documents. We calculated 95% confidence intervals for the proportion of two-subject queries that returned fewer than 5, 10, and 20 results. Table 4 presents the point estimates and confidence intervals.

Document Count Limit	Lower Limit of C.I.	Point Estimate	Upper Limit of C.I.
5	62.3%	67%	71.7%
10	68.6%	74%	77.4%
20	79.2%	83%	86.8%

Table 4. 95% Confidence Intervals (C.I.) for Proportions of Two-Subject Queries Returning Fewer than 5, 10, or 20 Documents

These estimates tell us that we should expect to see 83% of the two subject searches (the KW1 SU1 SU2 scenario) meet our goal of returning 20 or fewer documents, and can be 97.5% confident that nearly 80% of the two-subject searches will return 20 or fewer documents. Similarly, we can be 97.5% confident that more than 62% of the KW1 SU1 SU2 queries will return 5 or fewer documents, and 97.5% confident that more than 68% of the KW1 SU1 SU2 queries will return 10 or fewer documents.

It also is interesting to observe the reduction in the number of documents that are returned in the KW1 SU1 SU2 scenario versus the KW1 scenario. Table 5 presents point estimates and confidence intervals for the proportion of instances in which the result set will be reduced by 90%, 95%, and 99% through the use of subject metadata (i.e., adding the restriction that KW1 and KW2 both must appear in the subject field of the metadata). When reading the title to this paper, most readers probably questioned whether an order-of-magnitude improvement in search results was possible. Table 5 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 87.81% of the searches. These are very strong results.

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 5. 95% Confidence Intervals (C.I.) for Proportion of Two-Subject Queries Reducing Result Set Sizes by 90%, 95%, and 99%

DISCUSSION

The reduction in result set size is extremely encouraging, and indicates that searching subject metadata along with the full-text of the document will make enterprise search a much more successful endeavor. If the size of the result set can be kept low, it is more likely that searchers will look through all of the documents returned and find the desired document.

Our findings support the earlier findings of Voorbij (1998) and Gross and Taylor (2005) that the addition of subject metadata search can improve search results, and extends their findings beyond a search of the bibliographic record to an evaluation of the impact the addition of metadata search has on 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 explains why their searches were not impacted by the indeterminacy of language and did not benefit from adding contextual information. This is not the case for most searches.

Our results show that we can be 97.5% confident that the use of metadata will reduce the average result set size to 20 or fewer documents more than 79% of the time. Our results also show that incorporating metadata into the search process is very likely (.975) to result in a tenfold improvement in search for 97.95% of searches. This is very strong evidence that the use of subject metadata should be incorporated into the search process.

CONCLUSION

We performed an experiment that measured the impact of adding subject metadata to keyword-based full-text searches. Our extremely encouraging results suggest that the traditional library process of indexing the contents of the library against a controlled vocabulary of subjects, authors, and titles might need to be rejuvenated in the context of enterprise search.

The use of subject indexes has largely been replaced by the use of enterprise search appliances built on full-text web search engines. The indeterminacy of language leads to very large result sets being returned by such search engines. We have demonstrated that incorporating the search of subject metadata into the search process dramatically reduces the size of the result set. In the case of enterprise search, we suggest that it might be better to automate, not obliterate, the traditional library search process.

LIMITATIONS AND FUTURE RESEARCH

Our research has several limitations. First, we assumed that ABI/Inform did a perfect job of subject indexing. If this assumption is true, then no relevant documents will be excluded by adding the requirement that the search term appear in the subject metadata. Hence we looked only at the number of documents returned by the searches. No attempt was made to evaluate the relevance of the documents returned. We acknowledge that our assumption of indexing perfection at ABI/Inform is not 100% accurate, and that this research needs to concern itself with the relevance of documents in result sets. We are currently planning a new experiment that will look at this problem from a perspective that does not require any assumptions related to document relevance.

Second, we assumed that there is little or no cost to gathering subject metadata. If the cost is low and its use reduces the size of the result set without excluding relevant documents, then clearly the use of subject metadata should be incorporated into the search process. However, if the cost is high, or if relevant documents are excluded, then it may not be desirable or possible to incorporate subject metadata into the search process. Further research needs to be done to determine the cost of collecting and codifying subject metadata. ABI/Inform should be an excellent source for investigating this. The fact that ABI/Inform currently is collecting and codifying subject metadata indicates that the cost is not prohibitive.

Finally, we used randomly generated query terms in this study. Although there is reason to believe that our process of generating search terms approximates the search behavior of persons tasked with doing enterprise search, we have not demonstrated this. We tried to be as objective as possible by selecting our set of terms from a thesaurus that was developed by subject matter experts, but we cannot guarantee that we mimicked the manner in which individuals perform searches. Future research needs to be conducted in this area. One possibility is to gain access to actual query logs. Another possibility is to conduct a controlled experiment.

Despite these limitations, the results from our simulated searches are very compelling. Prior to this research, there were no reasonable estimates for the improvement in search outcomes that might result from using subject metadata, and few, if any, researchers would have thought that an order of magnitude improvement was possible.

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. 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.
3. 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.
4. Library of Congress (2005) "Using the Library of Congress," http://www.loc.gov/rr/main/inforeas/card.html, Accessed 2/17/09.
5. EContent (2004) Getting Just What You Need, EContent, 27, 7/8, S12-S13.
6. Fagin, R., Kumar, R., McCurley, K.S., Novak, J., Sivakumar, D., Tomlin, J.A., and Williamson, D.P. (2003) Searching the Workplace Web, in Proceedigns of the 12th International Conference on International World Wide Web, Budapest, Hungary, Association for Computing Machinery.

7. Gardner, W.D. (2008) "Most Users Are Unhappy With Enterprise Search," http://www.informationweek.com/news/software/database/showArticle.jhtml?articleID=207100963, Accessed 7/24/2008.
8. Gross, T., and Taylor, A.G. (2005) What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results, College and Research Libraries, 66, 3, 212-230.
9. Hammer, M. (1990) Reengineering Work: Don't Automate, Obliterate, Harvard Business Review, 68, 4, 104-112.
10. Hemminger, B.M., Saelim, B., Sullivan, P.F., and Vision, T.J. (2007) Comparison of Full-Test Searching to Metadata Searching for Genes in Two Biomedical Literature Cohorts, Journal of the American Society for Information Science and Technology, 58, 14, 2341-2352.
11. 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.
12. KMWorld (2008) "Developing a Universal Search Strategy (Hint: Start with Usability)," http://www.kmworld.com/Webinars/90-Developing-a-Universal-Search-Strategy-(Hint-Start-with-Usability).htm, Accessed 1/16/2009.
13. Kontzer, T. (2003) "Search On," InformationWeek (923), pp 30-36.
14. Raghavan, P. (2001) Structured and Unstructured Search in Enterprises, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24, 4, 15-18.
15. 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, November 3-8, Austin, TX.
16. Ultraseek (2006) "Business Search vs. Consumer Search: Five Differences Your Company Can't Afford to Ignore" http://publications.autonomy.com/pdfs/Ultraseek/White%20Papers/mk0759_Business_v_Consumer_WP.pdf, Accessed 1/20/2008.
17. Voorbij, H. (1998) Title Keywords and Subject Descriptors: A comparison of Subject Search Entries of Books in the Humanities and Social Sciences, Journal of Documentation, 54, 4, 466-476.
18. Zipf, G.K. (1949) Human Behavior and The Principle of Least Effort: An Introduction to Human Ecology Addison-Wesley Press, Inc., Cambridge, MA.