Migrating Information Retrieval from the Graduate to the Undergraduate Curriculum

Nazli Goharian
David Grossman
Ophir Frieder
Information Retrieval Laboratory
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
{goharian, grossman,frieder}@ir.iit.edu

Nambury Raju
Institute of Psychology
Illinois Institute of Technology
Chicago, IL 60616
raju@iit.edu

ABSTACT

As reliance on the web in general and web search engines in particular expands, student interest in the domain of information retrieval is increasing. In response to this increased interest, at the Illinois Institute of Technology, we recently extended our computer science undergraduate curriculum to include an introductory course in information retrieval. Instead of simply understanding how to build applications using information retrieval tools, our students *build* these tools and learn the relevant algorithms implemented in these tools. Our syllabus includes a hands-on lab setting where students use the tools they build to perform experiments that could ultimately extend the field. Also as part of our course, we administered pre- and post-questionnaires to assess how much knowledge (of the topics included in the course) the students thought they had prior to the start of the course, their expectations for the course, and how well these expectations were met. We present results that indicate a statistically significant improvement in the students' self-reported knowledge in information retrieval. Student final exam scores substantiate these results.

Keywords: Undergraduate Curriculum, Information Retrieval Education

1. INTRODUCTION

As the dependence on search technology both for the Internet and various intranet applications continues to grow, more and more computer science programs are introducing information retrieval courses into their curriculum. Although many computer science programs are introducing information retrieval courses into their graduate offerings, relatively few programs offer an undergraduate course in information retrieval. Some of those that do, for example Johns Hopkins University and University of Tennessee at Knoxville, offer their courses to combined graduate and undergraduate audiences. At the Illinois Institute of Technology (IIT), as at the University of Texas and at the University of Science and Technology at Hong Kong, we offer a course dedicated to undergraduate students. We also offer a more advanced course at the graduate level that focuses on research with the described undergraduate course as its prerequisite.

We describe our newly introduced undergraduate information retrieval course that is based on what previously was our graduate (more theoretical) class. In developing an undergraduate (more applied) version, we believe that it is still inappropriate to simply teach students how to use existing commercial products without first providing them with an understanding of the fundamental algorithms. That is, we believe that a reasonable mix of theory and practice is needed to offer an undergraduate course in information retrieval.

There is the added benefit that this course will assist with our research efforts. In fact, a co-authored paper with an undergraduate student who took this course was recently published (Stein, 2003). Issues such as testing human-computer interfaces and developing more robust relevance assessments can be addressed by assignments that are well suited to undergraduate student contributions. Many issues in human-user interaction are often ignored in information retrieval. The students at the undergraduate level are eager to serve as test users and provide significant feedback as to the pros and cons of different user interfaces. Additionally, test collections often lack detailed user feedback. Information retrieval researchers often rely on collections with merely binary judgments of document relevance; a document is relevant or not relevant. Students at the undergraduate level can provide multi-tiered relevance judgments such as "this document is a little bit relevant" that may well serve to improve our understanding of existing effectiveness tools. It should be noted that asking students to invent something is appealing but it can be a tremendously frustrating experience for most students. Hence, the students are asked to evaluate the relevance of the retrieved results manually based on their perception as the user and compare that to the relevance judgment provided by the IR system. This is an area where all students are able to contribute their impressions, and we are much more likely to see a successful contribution on the part of the undergraduate student.

We realize that our curriculum extensions always need to be updated because the field of information retrieval is changing rapidly. Our textbook (Grossman, 1998), published in 1998, lists five fundamental retrieval strategies. Since then, at least two new strategies were introduced. Similarly, the number of web search engines has expanded; we are aware of over twenty actively used web search engines and five meta-search engines. In spite of this rapid change, however, the information retrieval field is sufficiently mature, as exemplified by recent related surveys (Kobayashi, 2000; Meng, 2002), that the field fundamentals can be taught. The original retrieval strategies such as the vector space model and the probabilistic model are widely used and will be with us for a long time. Understanding these retrieval strategies provides the foundation for graduate studies or commercial employment on new strategies. The retrieval utilities such as thesauri, relevance feedback, semantic networks, passages, n-grams, etc. are also broadly used and will be for many years. Future work will build on knowledge of these fundamentals. Hence, an undergraduate who knows these algorithms is well prepared to stay current in and potentially advance the field.

2. BACKGROUND

Initially, we briefly overview the field of information retrieval focusing in Section 2.2 on the efforts ongoing in our Information Retrieval Laboratory. An analysis of the state-of-art and a discussion on the need for curriculum enhancement conclude this background section.

2.1 Information Retrieval

A goal of Information Retrieval (IR) is to find relevant data in response to a user query. The data can be either structured (e.g., relational database attribute values) or unstructured (e.g., text, video, image, sound, and geospatial). Most existing information retrieval algorithms focus on text and work to improve the response of a user query to a large document collection. Strategies exist to rank documents for a given query, and utilities exist to improve on a given strategy. Examples of a strategy include the following: vector space model (Salton, 1975), probabilistic model (Robertson, 1976), (Extended) Boolean (Fox, 1983), fuzzy set (Salton, 1989), Bayesian inference networks (Turtle, 1991), Latent Semantic Indexing (Deerwester, 1990), and neural networks (Syu, 1994). Utilities improve any strategy, i.e. information retrieval effectiveness, and cover topics such as relevance feedback (Harman, 1992), thesauri (Gauch, 1996), clustering (Salton, 1989), semantic networks such as WordNet (Beckwith, 1990), and n-grams (Damashek, 1995). Texts describing strategies and utilities include (Salton, 1989; Grossman, 1998; Baeza-Yates, 1999; Grossman, 2003). Techniques to improve search efficiency are presented in (Witten, 1999; Frieder, 2000).

2.2 Existing Research at Information Retrieval Laboratory at IIT

Existing research at IIT has focused on developing new algorithms for information retrieval. Much of our work has focused on the integration of structured data with text (Grossman, 1997; Lundquist, 1999) and was implemented and deployed by both government and commercial organizations of varied sizes. The idea of our recent work on intranet mediators, where data from a data warehouse is seamlessly integrated with less structured data, is that a user really wants a unified view of data regardless of where or how it is stored (Grosman, 2002). At IIT, a mediator runs on the IIT web site to answer English-language questions posed predominantly by students and identifies answers found in sources available on campus (http://mediator.iit.edu). In addition to unified access, we have worked on using information fusion, namely, the combining of a variety of techniques, to improve effectiveness (Chowdhury, 2001; Beitzel, 2003).

Other projects in our lab include duplicate document detection that identifies and eliminates duplicate documents while searching for relevant documents (Chowdhury, 2002), sparse matrix information retrieval as an alternative approach to store and query text (Goharian, 2001; Goharian, 2003), Arabic-English cross lingual information retrieval that queries Arabic text using English queries or vise versa (Aljlayl, 2001; Aljlayl, 2002), a medical information system (Goharian, 2002), a parallel clustering algorithms to improve efficiency (Jensen, 2002; Ruocco, 1997), a text extraction algorithm to remove the unstructured data from web documents to improve the accuracy (Ma, 2003), and a misuse detection system for information retrieval system to detect any potential misuse by an authorized user (Ma, 2003). Much of the technology and experimentation can be transferred into the classroom.

2.3 Analysis of the State-of-the-Art and Need for Curriculum Development

Numerous web search engines exist including: Alta Vista, AlltheWeb, Teoma, FindWhat, Google, Infoseek, LookSmart, Thunderstone, Yahoo, and WebCrawler. Meta-search engines send a search to numerous search engines and collate the results. Commercial examples of these are: DogPile, Mamma, MetaCrawler, Profusion, and Search.

Research in information retrieval continues to grow rapidly with numerous papers published each year primarily in ACM SIGIR, ACM CIKM, NIST TREC as well as papers in traditional database conferences like SIGMOD and VLDB. Many of the efforts described within those forums are sufficiently mature to be taught to undergraduates. At IIT, we are both teaching our undergraduate students information retrieval concepts and also involving them in our research.

3. COURSE DESIGN

We recently developed and taught the described undergraduate information retrieval course. The topic syllabus is presented in Table 1. The semester consists of fifteen weeks. The first eleven weeks of the course focus on algorithms, followed by a week focusing on commercial applications. We allocated the last two weeks to recent special topics in information retrieval and student presentations. Our goal is to choose fundamental categories of algorithms within information retrieval so that the high-level structure of the course does not change. We strive to teach the latest developments after building a solid foundation covering essential topics that have not changed in spite of over 40 years of information retrieval research. The class projects, i.e, the hands-on lab experiments, center around students implementing key algorithms during most of the semester. Ultimately, commercial tools are investigated at the end of the semester.

The core goals of the course that students should be able to do are:

- Explain fundamental information retrieval storage methods (inverted index and signature files).
- Explain fundamental retrieval models, such as Boolean model, vector space model, and probabilistic model.
- Explain fundamental retrieval utilities such as stemming, relevance feedback, n-gram, clustering, thesauri, and parsing, and token recognition.
- Design and implement a search engine prototype using storage methods, retrieval models and utilities listed above.
- Apply the research ideas into their experiments in building a search engine prototype.

These goals are achieved in the course topics and assignments, i.e, the hands-on lab experiments. A week-by-week overview follows.

Table 1: Information Retrieval Syllabus

Week	Topic
1	Overview of information retrieval
2-3	Architecture of search engine, IR utilities: token recognition, stemming
4	Index structures: inverted indexes, signature files
5-6	Retrieval strategies: Boolean, fuzzy sets, vector space, probabilistic
7	Exam
8	Relevance feedback
9	Compression of inverted indexes and efficiency techniques
10-11	IR utilities: thesauri, semantic networks(WordNet), clustering, n-grams
12	Applications: details of web search engines
13-14	Special topics and student presentations
15	Exam

3.1 Overview of IR (Week 1)

The field of information retrieval concerns the storage and retrieval of large volumes of text, video, image, and sound. We discuss the need to maintain a balance between efficiency and effectiveness. For example, a sorting algorithm does not have a notion of a successful conclusion whereby the data are sorted as best as can be done at the current time. With information retrieval, a search that obtains the perfect answer is not possible; hence, we must do the best we can. This is a foreign concept to computer science students, and must be clearly motivated and explained. Traditional measures of effectiveness such as precision, recall, and average precision are then described in detail. Students are given an introductory assignment that requires them to become familiar with a pedagogical search engine that we built at IIT called *SimpleIR.* They are asked to install it, index some documents, and run a few queries. This is designed to get them familiar with the tool, as they will be using it throughout the entire semester as their experimental shell.

3.2 Architecture of Search Engine, Token Recognition, Stemming (Weeks 2 and 3)

The architecture of a search engine in general, and *SimpleIR* specifically, is explained and discussed. Algorithms that parse text, identify the term dictionary and then build the inverted file are discussed. An inverted index is the primary storage structure used for large volumes of text. Phrase recognition and stemming algorithms are discussed. Such algorithms are applied in information retrieval systems to increase the effectiveness of the search. Once this is done, students are given an assignment that requires them to modify the parser of *SimpleIR* so as to recognize special token cases, correctly parse and recognize phrases, and to properly stem the identified terms using well known stemming algorithms such as *Porter* (Porter, 1980) and *Lovins* (Lovins, 1968).

The students are instructed to perform many experiments to see the effects of phrase recognition and stemming in the size and building time of the inverted index.

3.3 Index Structures (Week 4)

An inverted index is the conventional index structure of information retrieval systems. Different approaches to build inverted index structures, including the approach applied in *SimpleIR*, are explained. In the later part of the project the students modify the existing memory-based approach to build an inverted index to a disk-based approach. As an example of other index structures, algorithms for signature files are explained.

3.4 Retrieval Strategies (Weeks 5 and 6)

Despite numerous research activities in information retrieval, we believe that the field can be partitioned into those strategies that simply provide a basic ranking of documents in response to a query and those utilities that are designed to improve ranking. In weeks 5 and 6, we focus on basic retrieval strategies. Since the vector space model, the probabilistic model, and the Boolean model are the most widely used, these particular models are emphasized. Should newer strategies become the strategies of choice, these can easily be inserted in this section of the course. *SimpleIR* is built with a simplistic implementation of the vector space model. Students are asked to modify *SimpleIR* to improve the vector space model. Potential improvements include incorporating more sophisticated similarity measures or the addition of other retrieval strategies, e.g., probabilistic, to the software.

3.5 IR Utility: Relevance Feedback (Week 8)

The notion of using feedback from an initial retrieval to improve effectiveness has been around since the 1960's. If we had to pick the single most enduring utility, this is certainly one of the key candidates. Existing web search engines often have a "more like this" button that incorporates a relevance feedback algorithm. The idea is to do an initial retrieval and then identify new query terms from those found in relevant documents that were retrieved. The selection of relevant can be either automated as with *pseudo-relevance feedback* algorithms or manual as with conventional relevance feedback algorithms. *SimpleIR* intentionally does not include a feedback algorithm because students are asked to incorporate one into the system.

3.6 Efficiency Techniques (Week 9)

During this week's session, compression algorithms used for reducing the size of the inverted index are discussed in detail. Once these are covered, students are given an assignment that requires them to modify *SimpleIR* applying index compression algorithms to compress the inverted index. Other efficiency techniques such as the use of index or query thresholds are discussed. Additional efficiency techniques will likewise be taught once they are developed.

3.7 Additional Utilities (Weeks 10 and 11)

Using a thesaurus is another example of reliance on an information retrieval utility to improve the retrieval effectiveness. We have found that students find this section quite interesting due to its potentially counter-intuitive nature. It would seem obvious that a thesaurus would assist with retrieval, as it would appear that the use of synonyms would reduce problems with ambiguities in language. The idea is that a search that includes the word *dog* should be improved by automatically extending the search to include *canine.* Although this is intuitive, it turns out that automatically incorporating a thesaurus is non-trivial and has yet to be shown as a definitive improvement in search. In spite of the lack of definitive proof of improvement, algorithms exist to automatically construct thesauri and to use them. These algorithms are described in this section of the course, and the students are instructed to experiment with these algorithms evaluating their effect on retrieval.

Another utility is called semantic networks, the most common of which is WordNet (Word 2003). WordNet is a network of words built for the purpose of improving search engines. Words are linked not just by the synonym relationship (as with a thesaurus) but also with many other relationships such as antonyms, hypernyms (is-a), hyponyms (part-of), etc. Students are asked to incorporate WordNet into SimpleIR. Additionally, in class, as part of this section, several clustering, n-grams based, and passage based algorithms are described.

3.8 Web Search Engines (Week 12)

This section brings the algorithms of the course into focus with a discussion of how web search engines are designed. Searching terabytes of text is non-trivial, and we discuss how this is done. In addition to the search algorithms, we also include a section on web crawling, search agent design, index updating, etc. We conclude this section with a general discussion on distributed and parallel search algorithms.

3.9 Special Topics and Students Paper Presentations (Weeks 13 and 14)

Students are asked to read and present research papers in information retrieval from a pre-selected list of the papers that covers both the information retrieval classic papers and papers that are on a current special topic in information retrieval. Students work in pairs rather than individually to foster discussions and research interaction among themselves.

3.10 Exams

Each exam is planned to measure the students' knowledge on the topics of Table 1, which cover the areas listed in Tables 2 and 3. The first exam covers the topics of weeks 1 thru 6 of Table 1, corresponding to the content areas of retrieval models (vector space model, Boolean, probabilistic), storage models (inverted index, signature files), and some of retrieval utilities (parsing, stemming), listed in Tables 2 and 3. The second exam although covers some of the topics covered in the first exam, however the main focus is on the topics of weeks 8 thru 11, which corresponds to the remaining retrieval utilities (n-grams, relevance feedback, thesauri, clustering), and efficiency issues (compression, index and query thesholding), listed in Table 3.

4. LESSONS LEARNED FROM TEACHING THE COURSE

By teaching the course, we learned that the prerequisite student knowledge must include algorithms, data structures, and strong programming skills. Those students that did not have sufficient algorithm and data structures backgrounds tended to focus on purely the development of the required software rather than on the experimentation with the resources. That is, although quite often these students did get the program to execute, the effort involved in creating the software eliminated the opportunity to "tinker" with the solution as a form of optimization. Since in the information retrieval domain there typically does not exit a perfect answer, tinkering with the approximate answers is a key component in learning, especially when dealing with the utilities. Clearly, those students with insufficient programming background generally did not complete the assignments, and hence, gained relatively little from the course, if they indeed completed the course. A problem that we encountered regarding the assignments was that the projects, by design, were incremental in their scope. As such, not completing the assignment in time implied that the students needed to play "catch-up" which likewise implied certain failure. In future offerings of the course, we plan to provide the students with the option of completed modules. Thus, failure to complete one module will not necessarily penalize the student in the succeeding assignments.

5. EVALUATION OF RESULTS

To measure the effectiveness of our proposed teaching approach, we conducted evaluations of student learning of the information retrieval course content in both semesters we taught the course, i.e., Spring and Fall 2002. In each of the two semesters, two questionnaires were given to the students, one at the beginning of the semester (Pre-Questionnaire) and the other at the end of the semester (Post-Questionnaire). Students were asked to indicate their knowledge and understanding of the course topics in information retrieval and their expectations for this course. Standard statistical data analysis procedures were adapted to assess student pre- and post-course knowledge and expectation.

Table 2: Summary Statistics for Common Items in the Pre- and Post-Questionnaires (Sample size: 15 students - Spring 2002)

Content Area	Pre M	Pre SD	Post M	Post SD	Post –Pre
Retrieval Models	1.37	0.35	4.16	0.64	2.79**
Vector Space Model	1.40	0.51	4.47	0.52	3.07**
Probabilistic Model	1.40	0.63	4.07	0.80	2.67**
Boolean Model	1.33	0.62	3.93	0.88	2.60**
Storage Methods	1.49	0.75	4.11	0.73	2.62**
Inverted Index	1.67	1.05	4.60	0.63	2.93**
Signature Files	1.33	0.62	3.73	1.03	2.40**
Compression of Inverted Index	1.47	0.92	4.00	0.85	2.53**
Retrieval Utilities	1.57	0.50	4.17	0.65	2.60**
Stemming	1.67	0.72	4.53	0.64	2.87**
Relevance Feedback	1.40	0.74	4.60	0.63	3.20**
N-Gram	1.27	0.59	3.87	0.92	2.60**
Clustering	1.67	1.05	3.87	0.99	2.20**
Thesauri	1.33	0.49	3.47	1.06	2.13**
Parsing or Token Recognition	2.07	1.03	4.67	0.49	2.60**
Applications					
Commercial Search Engines	3.07	1.03	3.73	0.8 8	0.66*
* p < .05 & ** < .01					

Table 3: Summary Statistics for Common Items in the Pre- and Post-Questionnaires (Sample size: 12 students – Fall 2002)

Content Area	Pre M	Pre SD	Post M	Post SD	Post –Pre
Retrieval Models	2.14	0.79	4.22	0.89	2.08**
Vector Space Model	2.00	0.85	4.50	0.85	2.50**
Probabilistic Model	2.17	0.83	3.92	0.90	1.75**
Boolean Model	2.25	0.97	4.25	1.14	2.00*
Storage Methods	1.67	0.71	3.89	0.83	2.22**
Inverted Index	2.00	1.35	4.42	1.16	2.42**
Signature Files	1.58	0.79	3.50	1.00	1.92**
Compression of Inverted Index	1.42	0.67	3.75	0.97	2.33**
Retrieval Utilities	1.93	0.64	3.96	0.92	2.03**
Stemming	1.67	0.98	4.08	1.08	2.41**
Relevance Feedback	1.92	1.08	4.25	1.14	2.33**
N-Gram	1.67	0.98	3.92	1.08	2.25**
Clustering	2.08	1.16	3.75	0.87	1.30**
Thesauri	1.50	0.67	3.42	1.16	1.92**
Parsing or Token Recognition	2.75	1.29	4.33	1.23	1.58*
Efficiency Techniques	1.46	0.62	3.46	1.44	2.00**
Top Docs (Posting List Threshold)	1.50	0.67	3.58	1.50	2.08**
Query Threshold	1.42	0.67	3.33	1.50	1.91**
Applications					
Commercial Search Engines	2.92	0.79	3.58	0.90	0.66*
* p < .05 & ** < .01					

Twenty-nine students initially enrolled in the Information Retrieval course in the spring of 2002, and 25 of these students completed the course. In Fall 2002, eighteen students initially enrolled in course and 14 of those students completed the course. The pre/post-questionnaires were designed to gather some demographic and background information about each student (student ID, age, gender, race/ethnicity, year in school, number of CS courses taken, level of proficiency in English, and the current level of programming proficiency); to seek information about a student's current level of knowledge on retrieval models, storage methods, retrieval utilities, and applications; and to provide the student an opportunity to express his/her expectations/comment for the course.

Twenty-two students completed the pre-questionnaire and 20 students completed the post-questionnaire, resulting in data from 15 students for both questionnaires in Spring 2002. The number of students that filled out both questionnaires in Fall 2002 was 12. Responding to the questionnaires was optional but requested. A 5-point scale (1 = No Knowledge, 3 = Some Knowledge, and 5 = Very Knowledgeable) was used for all items.

In Tables 2 and 3, we present the pre- and post-questionnaires summary of findings for Spring and Fall 2002, respectively. As shown, most students had little prior knowledge in the core content areas of our information retrieval course. The only area that the students appeared knowledgeable in prior to taking the class was "commercial search engines", where the mean score was 3.07 and 2.92, respectively from within a one to five point scale. Low mean scores (less than 2.00 in Spring and 2.25 in Fall) on the remaining content areas meant that this was a course where the potential for gaining new knowledge was substantial for most students. For completeness, the mean and standard deviation scores as well as the post- and pre- mean differences for all the individual area components are presented.

To formally assess the self-perceived gain in course-content knowledge, the Pre-Q and Post-Q data were statistically analyzed with a t-test (Winer, 1991). Only students with both pre and post-questionnaire data were included in the analysis. As shown, the results for the retrieval models, the storage methods, and the retrieval utilities are significant at a 99% confidence interval, while the results for the applications area are significant at a 95% confidence interval for both semesters the class was taught. As an aside, it should be noted that the overall final exam scores (with a mean of 3.36 in Spring and 3.21 in Fall on a 4-point scale) likewise demonstrate that indeed the students did actually learn the material and not only perceived that they learned it. Note that the exams were designed to cover the content areas that appeared in the pre- and post- questionnaires.

6. PARTITIONING GRADUATE AND UNDERGRADUATE MATERIAL

In Section 2.1, we provided a brief overview of the field of Information Retrieval. A key decision with our new undergraduate course was what to include at the undergraduate level given that we had taught for some time at the graduate level.

It turns out that we were able to include almost all of the strategies and utilities discussed at the graduate level. The exceptions were those that required more mathematical maturity than most undergraduates would have. Two interesting models, the Bayesian Inference Network (Turtle 1991) and Latent Semantic Indexing (Deerwester, 1990), require a fair amount of mathematical sophistication. Thus, we decided to introduce them strictly at the graduate level course. Given that we also teach the material at the undergraduate level more slowly, this is a fortuitous reduction in material. These models are fairly popular, but major search engines and commercial products do not include them; so we do not feel that students suffer tremendously from their departure from the curriculum.

Our advanced information retrieval course covers both Bayesian Inference Network and Latent Semantic Indexing that are omitted from the undergraduate level, and the remainder of the graduate course is taught as a seminar in which students read recently published research papers and present their key topics to the class. We have found that the background provided at the undergraduate level enables students to easily digest new material from the state-of-the-art. Clearly, this is crucial as with most computer science topics, improvements to the state-of-the-art are constantly happening.

7. SUMMARY

We presented our effort to extend our undergraduate computer science curriculum to include a course on information retrieval. The first year of the NSF-funded project involved the development and implementation of an undergraduate course in information retrieval. It was offered during the spring of 2002, where 29 students enrolled and 25 students completed the course; and in the fall of 2002 that 18 students enrolled and 14 students completed the course. As part of the course, students in both semesters were administered a pre-questionnaire at the beginning of the course and a post-questionnaire at the end of the course. In addition to obtaining some demographic and background information about the students, the questionnaires were designed to assess how much the students thought they knew about various course topics at the beginning of the course and how much knowledge they thought they had gained as a result of taking this course. A statistical analysis of the questionnaire data from both semesters showed that students, on average, thought that they had gained a good deal of knowledge regarding the topics presented in this course. Final exam scores indicated that indeed the students did learn the course material. The course materials and further details can be found at http://ir.iit.edu/~nazli/cs429.

8. ACKNOWLEDGEMENT

This work is supported in part by the National Science Foundation under contract # 0119469.

9. REFERENCES

Aljlayl, M. and O. Frieder (2001), "Effective Arabic-English Cross-Language Information Retrieval via Machine Readable Dictionaries and Machine Translation," ACM Tenth Conference on Information and Knowledge Management, Atlanta, Georgia, November 2001.

Aljlayl, M. and O. Frieder (2002), "On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach," ACM Eleventh Conference on Information and Knowledge Management, Washington, DC, November 2002.

Baeza-Yates, R. and B. Ribeiro-Neto (1999), Modern Information Retrieval, Addison Wesley Longman, 1999.

Beckwith, R. and G. Miller (1990), "Implementing a Lexical Network", International Journal of Lexicography, 3(4), pp. 302-312, 1990.

Beitzel, S., E. Jensen, A. Chowdhury, O. Frieder, D. Grossman, and N. Goharian (2003), "Disproving the Fusion Hypothesis: An Analysis of Data Fusion via Effective Information Retrieval Strategies," ACM Eighteenth Symposium on Applied Computing (SAC), Melbourne, Florida, March 2003.

Chowdhury, A., O. Frieder, D. Grossman, and M. C. McCabe (2001), "Analyses of Multiple-Evidence Combinations for Retrieval Strategies," ACM Twentieth SIGIR, New Orleans, Louisiana, September 2001.

Chowdhury, A., O. Frieder, D. Grossman, M. McCabe (2002), "Collection Statistics for Fast Duplicate Document Detection," ACM Transactions on Information Systems, 20(2), April 2002.

Damashek, M. (1995), "Gauging similarity via N-grams: Language independent categorization of text", Science, 267 (5199), 1995.

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Hashman (1990), Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), pp. 391-407, 1990.

Fox, E. (1983), “Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types, PhD thesis, Cornell University.

Frieder, O., D. Grossman, A. Chowdhury, and G. Frieder (2000), "Efficiency Considerations in Very Large Information Retrieval Servers," Journal of Digital Information, 1(5), April 2000.

Gauch, S., and J. Wang (1996), “Corpus Analysis for TREC-5 Query Expansion”, Proceedings of the Fifth Text REtrieval Conference (TREC-5), 1996, pp. 537-546.

Goharian, N., T. El-Ghazawi, and D. Grossman (2001), “Enterprise Text Processing: A Sparse Matrix Approach” IEEE International Conference on Information Techniques on: Coding & Computing (ITCC 2001), 2001.

Goharian, N., P. Jain, G. Kora, A. Jain (2002), “A Web-Based Medical Diagnosis and Treatment System for Urinary Tract Infections”, The 2002 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS'02), June 2002.

Goharian, N., A. Jain, Q. Sun (2003), "Comparative Analysis of Sparse Matrix Algorithms for Information Retrieval", Journal of Systemics, Cybernetics and Informatics, 2003.

Grossman, D., O. Frieder, D. O. Holmes, and D. C. Roberts (1997), “Integrating Structured Data and Text: A Relational Approach“, Journal of the American Society of Information Science, 48(2), Feb. 1997.

Grossman, D. and O. Frieder (1998), Information Retrieval: Algorithms and Heuristics, Kluwer Academic Publishers. Norwell, Mass. 1998.

Grossman, D., S. Beitzel, E. Jensen, and O. Frieder (2002), “IIT Intranet Mediator: Bringing Data Together on a Corporate Intranet,” IEEE IT PRO, January/February 2002.

Grossman, D., and O. Frieder (2003), Information Retrieval: Algorithms and Heuristics, Second Edition, Kluwer Academic Publishers. Norwell, Mass. 2003.

Harman, D. (1992), “Relevance Feedback Revisited”, Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, June 1992.

Jensen, E., S. Beitzel, A. Pilotto, N. Goharian, O. Frieder (2002), “Parallelizing the Buckshot Algorithm for Efficient Document Clustering”, ACM Conference on Information and Knowledge Management (CIKM), November 2002.

Kobayashi, M., and K. Takeda (2000), “Information Retrieval on the Web,” ACM Computing Surveys, 32(2), June 2000.

Lovins, J. B. (1968), “Development of a Stemming Algorithm” Mechanical Translation and Computational Linguistics, 11:22-31, 1968.

Lundquist, C., O. Frieder, D. Holmes, and D. Grossman (1999), “A Parallel Relational Database Management System Approach to Relevance Feedback in Information Retrieval”, Journal of the American Society of Information Science, 50(5), April 1999.

Ma, L., N. Goharian, A. Chowdhury, M. Chung (2003), “Extracting Unstructured Data From Template Generated Web Documents”, ACM 12th Conference on Information and Knowledge Management (CIKM), November 2003.

Ma, L., R. Cathey, N. Goharian, D. Grossman (2003), “Misuse Detection for Information Retrieval Systems”, ACM 12th Conference on Information and Knowledge Management (CIKM), November 2003.

Meng, W., C. Yu, and M. F. Liu (2002), “Building Efficient and Effective Metasearch Engines,” ACM Computing Surveys, 34(1), March 2002.

Porter (1980), Porter Stemming Algorithm, Program, 14(3), July 1980.

Robertson, S. E., and K. Sparck Jones (1976), “Relevance weighting of search terms”, Journal of the American Society of Information Science. May-June 1976.

Ruocco, A., and O. Frieder (1997), **“**Clustering and Classification of Large Document Bases in a Parallel Environment”, Journal of the American Society of Information Science, 48(10), October 1997.

Salton, G., A. Wong, and C.S. Yang (1975), “A Vector Space Model for Automatic Indexing”, Communications of the ACM, p. 613-620, 1975.

Salton, G. (1989), Automatic Text Processing, Addision-Wesley, 1989.

Stein, S., and N. Goharian (2003), “On the Mapping of Index Compression Techniques on CSR Information Retrieval", IEEE International Conference on Information Techniques on: Coding & Computing (ITCC 2003), Las Vegas, Nevada 2003.

Syu, I., and S. Lang (1994), “A Competition-Based Connectionist Model for Information Retrieval”, Proceedings of the 1994 Conference on Neural Networks.

Turtle, H. (1991), “Inference Networks for Document Retrieval”, Ph.D. Thesis, University of Massachusetts, Amherst, 1991.

Winer, B. J., D.R. Brown, and K.M. Michels (1991), Statistical Principles in Experimental Design, New York, NY: McGraw-Hill.

Witten, I. H., A. Moffat, and T. C. Bell (1999), Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition, Morgan Kaufmann Publishing, 1999.

Word03 http://www.cogsci.princeton.edu/~wn/.