Extending the Undergraduate Computer Science Curriculum to Include Information Retrieval and Data Mining

D. GROSSMAN, N. GOHARIAN, O. FRIEDER
Information Retrieval Laboratory
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
United States

N. RAJU
Institute of Psychology
Illinois Institute of Technology
Chicago, IL 60616
United States

Abstract:

We describe our progress extending the undergraduate Computer Science (CS) curriculum to include a deep understanding of information retrieval (IR) and data mining (DM). Instead of simply understanding how to build applications using tools involving IR and DM, students *build* these tools and learn the relevant algorithms implemented in these tools. Some novel approaches exist in our work. We include a hands-on lab setting where students use the tools they have built to perform experiments that could ultimately extend the field. Hence, undergraduates have firsthand knowledge of performing research in Computer Science using a scientific method. Secondly, we have a rigorous set of evaluation criteria developed by our Psychology department that will evaluate how well students learn using our novel approaches. Ultimately, we believe these two courses warrant consideration into standards developed for the undergraduate CS curriculum.

1 Introduction

Our objective is to increase the "applied world" relevance of the undergraduate CS curriculum by including information retrieval and data mining. Both fields are linked through a need for understanding the fundamentals of algorithms, statistics, machine learning, and human-user interaction. Our group has done research in both fields and has developed a graduate level textbook on information retrieval [15]. Students should understand the fundamental algorithms in each area, be familiar with what commercial products exist, and have some experience using them. We believe it is inappropriate to simply teach students how to use existing commercial products without understanding the fundamental algorithms. With some experimentation, we believe we can find the right mix of theory and practice to offer a two semester sequence of information retrieval and data mining.

In our curriculum, students build systems that implement key data mining and information retrieval algorithms and learn how to apply these algorithms to solve real-world problems. Ultimately, we will work to migrate our efforts to other Computer Science departments. We have discussed this with representatives from several different schools in a DELOS Workshop of Excellence in Switzerland, and there was strong interest in teaching these courses to undergraduates.

At the end of the two-semester course sequence, students know the fundamental algorithms and existing state-of-the-art in web search engines, intranets, data mining, and customer relationship management. Moreover, students have two significant group projects to gain experience in a software development team project environment. These projects enable larger implementation achievements that further understanding of the algorithms, implementation trade-offs and software project management. In addition, such a group project is a critical experience that future employers and graduate schools look for. Future plans include evaluating our curriculum enhancements using the latest proven techniques from the field of Psychology.

We realize that our curriculum extensions will always have to be updated because the fields of Information Retrieval and Data Mining are changing at a high rate of speed. Few textbooks exist at either the graduate or undergraduate level. Algorithms are constantly being revised.

However, the topics are now mature enough that 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. Additional retrieval utilities such as thesauri, relevance feedback, semantic networks, passages, n-grams, etc. are also broadly used and will be for many years. An undergraduate who knows these algorithms is well prepared to stay current in the field by learning their advancements and refinements.

We overview both information retrieval and data mining and present our approach to curriculum development. We focus our discussion on information retrieval as we have developed the material and are teaching the information retrieval course in Spring 2002. We are in the preliminary phases of the course development for data mining.

2 Background

2.1 Information Retrieval and Data Mining

Information Retrieval (IR) is about finding relevant data in response to a user query. The data may be either structured (e.g., name, address, phone number) or unstructured (text, video, image, sound, geospatial). Most existing IR 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. A detailed survey of strategies and utilities is given in [15]. Techniques to improve search efficiency are presented in [10]. Another problem in information retrieval is, given a static query, route documents that answer the query on a daily basis. The idea is that a user who is always interested in particular topics would like to develop a profile of their interest where the profile is a sophisticated query. Once this is done, new data are routed to profiles that are relevant to the data.

Data Mining is about finding patterns in data that we do not already know about [4, 18]. The basic idea is that if we track a few million-customer purchases (often referred to as market baskets) we may find out what products are purchased with other products. For many years, an example was cited that diapers were often sold with beer. The idea was that dad decided that he was entitled to a few beers after being sent to the store to buy diapers. This has long been known to be a myth but it illustrates the point [24]. Many companies do not like to share real examples of success as these examples are used to gain competitive advantage. OLAP (On-line Analytical Processing) algorithms are often used to allow users to “surf” data and verify trends, which may have been identified by data mining algorithms [21]. In addition to business applications, data mining techniques are also used for fraud and intrusion and misuse detection [3, 11]. A good survey of fundamental data mining algorithms is given in [18].

2.2 Relationship of Data Mining to Information Retrieval

Data mining and information retrieval are related in part due to the similarity of the goals of each task – finding key information from large collections of data. Much work has focused on applying specific data mining algorithms to information retrieval problems [8]. For example, recent algorithms for text filtering apply weighted boosting techniques commonly used in a variety of data mining applications [23]. Genetic algorithms improve the efficiency of parallel information retrieval by allocating documents to various processors [12]. Finally, document clustering into related groups such as Legal, Medical, Sports, etc. is almost exactly the same problem as clustering related customer records into groups such as “likely to buy”, “likely not to buy”. Hence, document clustering algorithms and data mining clustering algorithms are often very similar [12]. Finally, for both data mining and information retrieval a key consideration is whether or not the input data will fit into main memory as data are often in the several gigabyte or even terabyte range. This means algorithms for compression and efficient disk access are crucial to both fields.

2.3 Existing Research at Information Retrieval Laboratory at IIT

Existing research at IIT has focused on developing new algorithms for information retrieval and data mining. This existing research is ripe for classroom integration. Much of our work has focussed on the integration of structured data with text [14] and has been implemented and deployed by both government and commercial organizations of varied sizes. Our recent work on intranet mediators takes data from a data warehouse and seamlessly integrates them with less structured data yielding a unified view of data regardless of where or how they are stored [20]. In addition to unified access, we have worked on using information fusion to improve effectiveness [6]. Both approaches are promising in the lab on test data. Undergraduates in the new curriculum will participate in measuring their effect on users, learning simultaneously about CS research and algorithms. Other projects of the lab include duplicate document detection that identifies and eliminates duplicate documents while searching for relevant documents [7]; sparse matrix information retrieval that is an alternative approach to store and query text [13], Arabic-English cross lingual information retrieval that is querying Arabic text using English query or vise versa [1], and a medical information system [17].

2.4 Analysis of state-of-art and need for curriculum development

Numerous web search engines exist. Some examples are: Alta Vista, AlltheWeb, DirectHit, Excite, FindWhat, Google, Infoseek, LookSmart, Thunderstone, Yahoo, Snap, and WebCrawler. Metasearch engines send a search to several search engines and collate the results. Commercial examples of these are: DogPile, Mamma, MetaCrawler, Profusion, and SavvySearch. Similarly, a variety of large-scale data mining products exist including IBM’s Data Miner, Oracle’s Darwin, SAS, SPSS’s Clementine and numerous smaller desktop products such as CART and PRW.

Both fields have quite a bit of research happening each year in addition to growth of new commercial products. Despite this, we are unaware of undergraduate courses in either information retrieval or data mining. At best, information retrieval algorithms such as inverted indexes or word counting algorithms are embedded in data structures courses in the form of a small assignment. Data mining is usually mentioned as a side note in an introductory database course. We propose that the time is ripe to give both of these burgeoning areas of computer science the attention they deserve. A student interviewing for a job related to search engines would be dismissed at if they mentioned they had done a word counting algorithm in data structures. Similarly, a student who wishes to work in the data-mining field has little to go on with only a day or so of lecture in a database systems course.

3 New Curriculum Development

While information retrieval and data mining have been overlooked in the Computer Science curriculum, both topics have attracted both widespread commercial and academic interest. Given this rare synergy between academia and practitioners, it seems fitting to offer these courses at the undergraduate level and fill a key gap in the existing CS curriculum. Our curriculum currently lacks any undergraduate specializations. By taking our database course and this new two-course sequence, students would effectively be showing a strong interest in database systems related issues. We believe this would fill a large void in our existing curriculum.

We have already developed and are teaching an information retrieval course this Spring. As mentioned earlier, the data mining course development is in its preliminary phases. Thus, the information retrieval syllabus given below is currently being taught, while the syllabus given for the data mining course may be modified further prior to teaching the course.

Data Mining

Week	Topic
1	Overview, what is data mining
2-3	Regression
4-5	Decision Trees
6-7	Clustering
8	Association Rules
9	On-Line Analytical Processing
10-11	Applications: Data Mining for Information Security
12-13	Application: Data Mining for Fraud Detection

Information Retrieval

Week	Topic
1	Overview, what is information retrieval
2-3	Inverted Indexes, compression of inverted indexes
4-6	Retrieval strategies: Vector space, Probabilistic, Inference Networks
7-8	Relevance Feedback
9	Thesauri, Semantic Networks: Wordnet
10-11	Clustering, n-grams, passages
12-13	Applications: Details of Web Search Engines

The semester consists of fifteen weeks. Only 13 weeks are planned as we opt to leave two weeks free for tests and student presentations. The first eleven weeks of each plan focus on algorithms while the last two weeks focus on commercial applications. In each course, a variety of tools would be made available to students, and they would use these tools at the end of the semester. The idea is for students to implement key algorithms during most of the semester and then to use commercial tools at the end of the semester.

Given that these courses have such high relevance to industry, we have found it relatively easy to recruit industry sponsors. Student projects are geared directly towards real-world requirements.

3.1 Incorporation of Research at IIT

Our goal here is not to frustrate students by asking them to do research for which they are ill prepared. Instead, we believe that enough tuning parameters and variations exist with current algorithms that undergraduates may well develop new means by which to improve model accuracy in data mining and effectiveness in information retrieval. In no case are we asking students to develop novel algorithms as we believe that is really a graduate level activity that often takes years to refine. Instead we suspect that undergraduates will benefit from the quest of “playing with” these algorithms and using them to improve their accuracy.

3.2 Experimental Methods and Procedures

We are tracking student’s progress with detailed evaluation forms during multiple points within the semester. Additionally, we will track the number of original research contributions obtained by undergraduates. We anticipate that the projects early in the semester will involve simple implementation of a known algorithm while later in the semester we will require students to “play” with an existing algorithm they have implemented and try to improve its effectiveness. At a minimum this will lead to new ideas for human-user interaction with these systems. Additionally, we expect some students will develop new modifications to improve effectiveness. We will develop clear learning objectives for students. Projects will be geared towards these objectives and tests will be given to ensure that students have met the objectives.

In teaching the course, we are conducting an ongoing formative evaluation of student learning on a weekly or biweekly basis. The goal of this formative evaluation is to track student learning, identify strengths and weaknesses of individual students and provide them with appropriate remedial help, and review content relevance and mode of instruction and revise content/instruction as needed. While formative evaluation is an integral part of the two-course sequence for all three years, it is especially crucial during the first year, and so, we are devoting substantial resources to this activity during the first year of the program. The goal here is to develop a sequence of courses that are not so difficult that students are frustrated but sufficiently challenging to motivate students to push themselves to learn more.

3.3 Evaluation of Results

An evaluation of the effectiveness of the proposed undergraduate curriculum in data mining and information retrieval is being carried out in three phases. Since the proposed new undergraduate courses in data mining and information retrieval are available for all eligible undergraduate students, the traditional control/experimental group paradigm is not very useful for evaluating the effectiveness of the proposed curriculum. Therefore, each group of undergraduates is being compared with itself with appropriately defined pre-and post-measures.

Phase 1 (Data Mining)

Information Gathering

A. On the first day of class for the data mining course (Version 1), all students will be asked to fill out a two-part questionnaire. The first part of the questionnaire will be designed to gather demographic information (race/ethnicity, gender, etc.) as well as information about previous course work, especially in computer science and mathematics.

B. The second part of the questionnaire, which will consist of items with categorical response options (Likert-type; [2]), will ask participants/students to indicate their current knowledge and understanding of the major topics (e.g., decision trees, association rules, neural networks, clustering, etc.) to be covered in the data mining course. In addition, each participant will be asked to describe his/her expectations for the course. Each participant will also have the option to provide written comments.

C. At the conclusion of the data-mining course, the second part of the above-described questionnaire, with appropriate modifications to items on expectations and some new items on course content and instruction, will be re-administered to the same students. A summative evaluation test [5], reflecting the content of the entire course will be developed and administered to all students. The summative test will consist of a mixture of multiple-choice, written-response, and performance-based type items to provide an accurate assessment of what the students know and are able to do.

Data Analysis

D. Information from Part A will be used to describe the group of undergraduate students enrolled in the data mining course. This information may also be used for additional statistical analyses to be described later.

E. Results from Parts B and C will form the basis for assessing the degree to which the students' expectations for the course are met and how much they have learned of the contents of the data mining course. Percent of students showing mastery on each of the skill areas/objectives of the course will be computed and an overall profile of mastery for the class as a whole will be developed for use in program evaluation, including course content and instruction. Appropriate univariate and/or multivariate statistical analyses [19, 25] will be performed to identify and articulate the significant outcomes about students' expectations and learning. It should be noted that the sample size for Phase 1 will be small (about 20 students), and hence, the proposed univariate/multivariate statistical tests will not have adequate power to detect significant differences [9]. Therefore, data about students' expectations will also be reported in effect size units [16] to facilitate the interpretation of outcome measures.

Phase 1 (Information Retrieval)

F. The information obtained about each student enrolled in this course is similar to the information to be gathered in Phase 1 for the data-mining course. The first part of the questionnaire was identical to the one described above. In the second part of the questionnaire, students were asked to indicate their knowledge and understanding of the major topics in information retrieval (e.g., vector space model, probabilistic model, Bayesian inference networks, neural networks, etc.) and their expectations for this course. While the format of the summative test will remain the same, the content of the test will be different, reflecting appropriately the relevant topics from information retrieval. The previously described statistical data analysis procedures will be adopted to provide information about how well the students' expectations are met and how much they have learned from the course on information retrieval.

G. In addition, the students that have completed the two courses will be monitored after graduation for a period of about six months. The purpose behind this monitoring is to gather information about time spent looking for a job, starting salary, position, current employer, and the positive effect the two-course sequence may have had on these matters. A specially designed questionnaire will be used for gathering the needed information in this monitoring phase. Whenever possible, graduates, who are not exposed to the two-course sequence, will also be monitored, and the resulting data will be analyzed for determining post-graduation benefits as a result of the students' exposure to the two-course sequence.

Phase 2 (Data Mining and Information Retrieval)

H. Implementation and data analysis in this phase will be similar to the ones described in Phase 1. The number of students for both courses in this phase will be approximately 40.

I. In addition, a separate evaluation of the two courses will be conducted using the data from both Phases 1 and 2. For this segment of the data analysis, the total sample will consist approximately of 60 students. This increased sample size will add significantly to the power of the proposed statistical analyses.

Phase 3 (Data Mining and Information Retrieval)

J. Implementation and data analysis in this phase will be similar to the ones described in Phase 1. As in Phase 2, the number of students for both courses is expected to be 40.

K. A separate evaluation of the two-course sequence will also be conducted with the questionnaire and summative test data from all three phases, using a sample of approximately 100 students. As in Phase 2, this increased sample size will contribute significantly to the power of the proposed statistical analyses.

The proposed three-phase evaluation plan is designed to evaluate how well the students enrolled in the two-course sequence will learn the major topics or skill areas in data mining and information retrieval. The students' expectations for these courses as well as how such expectations are realized will also be documented and evaluated. To the extent possible, information about post-graduation benefits (i.e., time spent on searching for a job and starting salary) resulting from the exposure to the two-course sequence will be monitored and analyzed for trends.

Finally, if sample sizes are deemed adequate, the questionnaire and test data from all three phases will be analyzed separately for males and females to identify different learning patterns, if any, and use that information to redesign the mode of presentation in order to maximally benefit all students. A similar analysis may also be undertaken for various racial/ethnic groups, provided the sample sizes are considered adequate for the proposed statistical analyses.

4 Current Status

We have laid out a clear plan to extend out undergraduate CS curriculum to include information retrieval and data mining. We have shown that our plan has the potential to extend the state-of-the-art as well as provide a new pedagogical opportunity for CS students. At present, we have already added a three credit course in Information Retrieval to our undergraduate curriculum. This course is being taught as an experimental course in Spring 2002. To prepare for it, software was built and tested in a graduate level course. Five draft book chapters were developed that are designed for use at the undergraduate level. These chapters were tested in the graduate course in Fall 2001 and all sessions were videotaped. An equipment grant for a fully-fledged teaching lab was obtained and set up for sole use of the class in Spring 2002.

Furthermore, a detailed evaluation form specific to the goals and objectives of the course is developed. The form was administered on the first day of class. Preliminary results from the first running of the course will be available May 2002 and will be presented in the conference. The students taking this course are expected to be able to continue their work in information retrieval research; thus, a course is also added to the graduate curriculum to cover advance information retrieval systems. In this course the students get actively involved in research areas of information retrieval.

5 Summary

We hypothesize there is strong pedagogical value in teaching these courses at the undergraduate level. Students will be able to tune algorithms and improve the state of the art. By "playing with" these algorithms students will enjoy themselves and learn the nuances of these algorithms. Historically, our few forays introducing data mining algorithms into our standard database course resulted in students thoroughly enjoying the assignments. The problem is that trying to stuff data mining into the already full database course simply dilutes both topics. A separate two-course sequence on data mining and information retrieval ensures that CS students are well prepared for a future in either industry or academia.

Acknowledgment

We wish to thank Dr. Catherine McCabe for her comments on and assistance in the preparation of this paper. This work was supported in part by the National Science Foundation under contract number EIA-0119469.

References

[1] M. Aljlayl and O. Frieder, "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.

[2] A. Anastasi. *Psychological Testing*. Macmillan Publishing Company, 1988.

[3] T. Bass. Intrusion Detection and Multisensor Data Fusion: Creating Cyberspace Situational Awareness, *Communications of the ACM*, 100-105, 1999.

[4] M. Berry and G. Linoff. *Data Mining Techniques*. Wiley Computer Publishing,1997.

[5] B.S. Bloom, J.T. Hastings, and D.F. Madaus. *Handbook of Formative and Summative Evaluation of Student Learning*. McGraw-Hill Publishing, 1971.

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

[7] A. Chowdhury, O. Frieder, D. Grossman, M. McCabe, "Collection Statistics for Fast Duplicate Document Detection," *to appear in ACM Transactions on Information Systems (TOIS).*

[8] H. Chen. Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms, Journal of the American Society for Information Science, 46(3), pp. 194—216.

[9] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Erlbaum Publishing, 1988.

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

[11] O. Frieder and D. Grossman, "Detection of Misuse of Authorized Access in an Information Retrieval System," Patent filing by the Illinois Institute of Technology. Patent pending.

[12] O. Frieder and H. Siegelmann, Document Allocation in Multiprocessor Information Retrieval Systems. *IEEE Transactions on Knowledge and Data Engineering,* 9(4), July/August 1997.

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

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

[15] D. Grossman and O. Frieder. *Information Retrieval: Algorithms and Heuristics. Kluwer Academic Publishers. Norwell, Mass. 1998.*

[16] L.V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, 1985.

[17] P. Jain, N. Goharian, G. Kora, R. Nadler, S. Kim, A. Weiser, "Computer-Assisted Medicine in the Treatment of Nephrolithiasis" *IEEE International Conference on Advanced Science and Technology (ICAST 2001),* 2001

[18] R. Kennedy, Y. Lee, B. Van Roy, C. Reed, and R. Lippman. *Solving Data Mining Problems through Pattern Recognition*. Prentice Hall, 1998.

[19] R.E. Kirk. *Experimental Design: Procedures for the Behavioral Sciences*, Brooks/Cole Publishing Co., 1995.

[20] M. Lee, S. Beitzel, E. Jensen, D. Grossman, and O. Frieder, "Intranet Mediators: A Prototype," *IEEE Second Int'l Conf. on Information Technology: Coding and Computing (ITCC),* Las Vegas, Nevada, April 2001.

[21] *OLAP Solutions: Building Multidimensional Information Systems*, John Wiley & Sons.

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

[23] R. Schapire and Y. Singer. Improved Boosting Algorithms Using Confidence-rated Predictions. *Machine Learning*, 37 , 296-336, 1999.

[24]. Personal communication with Jim Scoggins, project manager for the WalMart Data Warehouse.

[25] M.M. Tatsuoka. *Multivariate Analysis: Techniques for Educational and Psychological Research.* Macmillan Publishing, 1988.