Intelligent Information Management, **2011, 3, 43-48**
doi:10.4236/iim.2011.32005 Published Online March 2011 (http://www.SciRP.org/journal/iim)

Data Mining Technology across Academic Disciplines

Lesley Farmer[1], Alan Safer[2], Eric Chuk[3]
[1]*California State University, Long Beach, USA*
[2]*California State University, Long Beach, USA*
[3]*University of California at Los Angeles, Los Angeles, USA*
E-mail:{*lfarmer, asafer*}*@csulb.edu, echuk@ucla.edu*
Received December 3, 2010; *revised January* 7, 2011; *accepted January* 28, 2011

Abstract

University courses in data mining across the United States are taught primarily in departments of business, computer science/engineering, statistics, and library/information science. Faculty in each of these departments teach data mining with a unique emphasis, although there is considerable overlap relative to course offerings, terminology, technology, resources, and faculty publications. Content analysis research aims to describe in detail the range of data mining technology differences and overlap across academic disciplines.

Keywords: Data Mining, Statistics, Academics

1. Introduction

Data mining is essentially the process of uncovering meaningful new correlations, patterns and trends from large quantities of complex data using statistical and mathematical techniques. With the help of powerful computers, new applications of data mining have been developed recently and have expanded its areas of use. Data mining is now applied in such diverse fields as medicine, education, finance, marketing, meteorology, and national defense, along with many applications associated with the Internet.

Since the mid-1990s, many more university courses in data mining are being taught across the United States. The major departments teaching such courses are computer science/engineering, business, statistics, and library/information science. In each discipline, data mining is taught with a moderately different emphasis (see for example Olson and Shi, 2006; Duda, Hart, and Stork, 2000; Hastie, Tibshirani, and Friedman, 2009). In business, applications include: identifying credit card fraud, insider trading patterns, and defect analyses. In the sciences, applications include: Medicare fraud, astronomical variations, and disease risk. In statistics, new analytic approaches are being developed, such as fuzzy logic (Larose, 2005; Berry and Lindoff, 2004; Roiger and Geatz, 2003). In library and information sciences, both theoretical and technical approaches are used, often bridging this field and specific professions such as law, industry, and the health sciences.

As a result of these various applications, different software, textbooks, and techniques are being used. To clarify the differences and similarities in each discipline, this study will examine the major academic variations within the data mining field in relation to keywords, articles, books, courses offered, textbooks taught, and software used.

2. Method

2.1. Keywords Used to Identify Data Mining Courses across Disciplines

Data mining keywords from different disciplines were identified in 2009 by searching a compiled list of data mining courses for each of four academic disciplines: business, computer science/engineering, statistics, and library/information science. Graduate programs were exclusively searched in this regard since these courses are routinely taught at that level. The findings on computer science and engineering were combined since there was much overlap of courses in these disciplines. The count for statistics courses was obtained only in statistics departments. To find these courses, various accrediting societies and associations were consulted to identify universities offering programs in each of those disciplines. The university sites were then searched for courses relating to data mining. Browsing course catalogs and department-specific webpages for relevant course titles and descriptions resulted in finding key-

words.

The list of universities with business programs offering data mining courses was obtained from the Association to Advance Collegiate Schools of Business, https://www.aacsb.net/eweb/DynamicPage.aspx?Site = AACSB&WebKey=00E50DA9-8BB0-4A32-B7F7-0A92E98DF5C6. From that list of business schools, the keywords used were: business intelligence, decision support, and data mining.

The list of universities with computer science programs was obtained from the Accreditation Board for Engineering and Technology, http://www.abet.org/schoolallcac.asp.The keywords relating to computer science data mining courses were: machine learning, artificial intelligence, and data mining.

The list of engineering programs was taken from the Accreditation Board for Engineering and Technology, http://www.abet.org/ schoolalleac.asp. The keywords in engineering courses relating to data mining were: pattern recognition, artificial intelligence, and data mining.

The computer science and engineering programs were obtained from the same accrediting association but different keywords were utilized.

The list of statistics programs obtained was on the website of the American Statistical Association at http://www.sci.csueastbay.edu/~mwatnik/statlist. The keywords from statistics programs offering data mining courses were: neural network, decision tree, and data mining. The keyword count for statistics was obtained from schools with graduate programs in statistics, but math departments were excluded because of the extremely small number of data mining courses in mathematics outside of schools offering a graduate degree in statistics.

The list of library and information science programs was from the American Library Association, http://www.ala.org/ala/education careers/education/ accreditedprograms/index.cfm. The keywords associated with data mining courses were: informatics, information retrieval, information management, knowledge management, knowledge discovery in databases, competitive intelligence, bibliometrics, biometrics, bibliomining and data mining.

2.2. Top Resources and Publications by Discipline

Information about the most commonly used books and software by discipline was collected from course syllabi and instructors' replies to email in 2009. A reply specifying book(s), software, or both was counted as a response. There is no further breakdown of the percent who responded with each piece of information because the counts from the two sources, syllabi and emails, were not kept separately.

The average number of articles per year from 1990 to September 2009 was calculated for the disciplines: business, computer science/engineering, statistics, and library/information science. The phrase "data mining" in the abstracts of journals, books, and conference proceedings was used to search the business database *ABI Inform Complete*, the computer science/engineering database *Compendex*, the statistics database *Current Index to Statistics*, and the library/information science databases *Library Literature and Information Science* and *Library, Information Science & Technology Abstracts*. It should be noted that in the Current Index to Statistics, which is the main database for statistics, there was no specific identifier for abstracts (as in the other two databases), so title/keywords was the closest option. This is very likely the reason for a lower number of results found in the statistics search. If one looks at the number of articles divided by the number of departments in a particular discipline having data mining courses, one can compare articles/department across disciplines to compare publication productivity. The list of departments was obtained from the same list as keywords to identify data mining courses in the different disciplines.

3. Results

There were 75 business faculty surveyed by email and 48 responded (64%) providing information on data mining books or related software. Of the 235 computer science/engineering faculty surveyed who were teaching data mining courses, 127 responded (54%). For inquiries from statistics departments, 31 of the 44 surveyed responded (a 70% email response rate of either book or software or both). All library/information science programs had online information. Although the degree of response combined both texts and software, the text titles and the type of software were recorded separately.

3.1. Courses by Discipline

Once the data mining courses were identified by discipline, the number of departments offering them was determined. That 2009 data are reported in **Table 1**. The courses are listed by departments because of the marked variation in courses by department. Note that the computer science/engineering departments offer the most graduate data mining courses followed by business school offerings.

3.2. Keywords by Discipline

Keywords obtained from university catalog titles, course listings and descriptive words relating to data mining

courses from each of the four major disciplines are shown in **Table 2**. Keyword overlap between disciplines is surprisingly infrequent.

3.3. Software by Discipline

In the email responses from academicians in each discipline, numerous types of data mining software were reported. These are presented as proportions in **Figures 2, 3, 4, and 5.**

3.4. Books by Discipline

Data mining books vary by discipline largely because their focus and applications differ. Some of the leading books as identified in 2009 are listed below by discipline. Only the leading books are listed. The Russell and Norvig title was the most popular, used more than twice as often as the next most cited textbook, by Duda, *et al*.

3.4.1. Business

- Witten, I., and Frank, E. 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington, MA.
- Berry, M., and Linoff, G. 2004. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Wiley, New York.
- Olson, D., and Shi, Y. 2005. Introduction to

Table 1. Number of U. S. university departments offering data mining courses by discipline.

Discipline	# of depts. with data mining courses	% of total # of depts. in discipline offering data mining courses
Business	83	17.6%
Computer Science/Engineering	187	48.8%
Statistics	46	28.0%
Library/Information Science	15	30.0%

Business	**Computer Science/ Engineering**	**Statistics**	**Library/Info Science**
business intelligence	adaptive computation	association/ link analysis	automatic extracting
competitive advantage	artificial intelligence	clustering (K means, near-est neighbors)	bibliometrics
CRM	database/ data warehouse	decision trees	bibliomining
database mgmt. systems	intelligent agents	genetic algorithms	biometrics
database decision making	knowledge discovery in databases	machine learning	business intelligence
data warehouse	machine learning	model validation	competitive intelligence
decision support systems	multidimensionality(data cubes)	neural networks/ fuzzy logic	content mining
intelligent enterprise	neural networks/neurocomputing processing	nonparametric learning	database management
knowledge mgmt./discovery mgmt.	text mining	pattern recognition	database decision making
information systems		support vector machines	data warehouse
market-basket analysis		training/testing dataset	decision support
OLAP		unsupervised learning	fuzzy logic
quantitative methods			health informatics
			informatics
			information mgmt.
			information retrieval
			knowledge mgmt.
			knowledge disc/database
			quantitative methods
			text mining

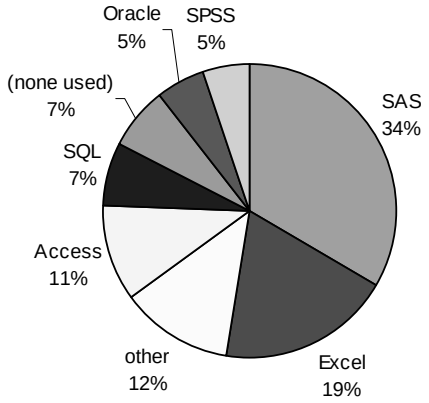

Figure 2. Business Data Mining Software by Brand Name (n = 57).

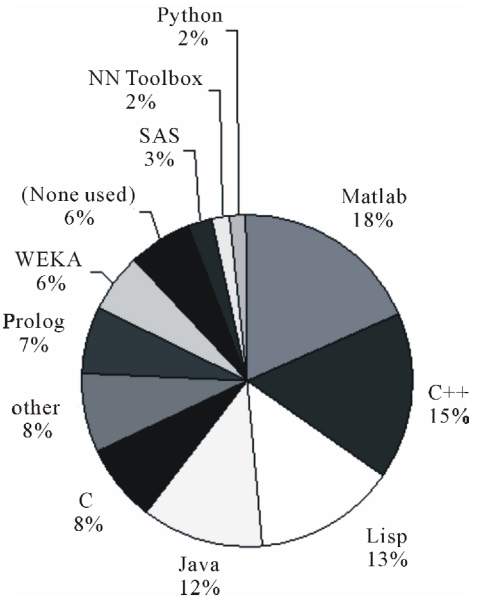

Figure 3. Computer Science/Engineering Data Mining Software by Brand Name (n = 118).

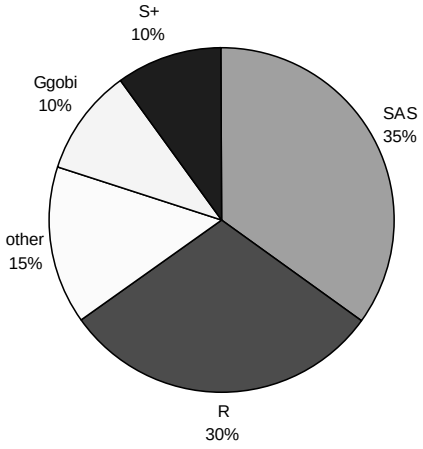

Figure 4. Statistics Data Mining Software by Brand Name (n = 18).

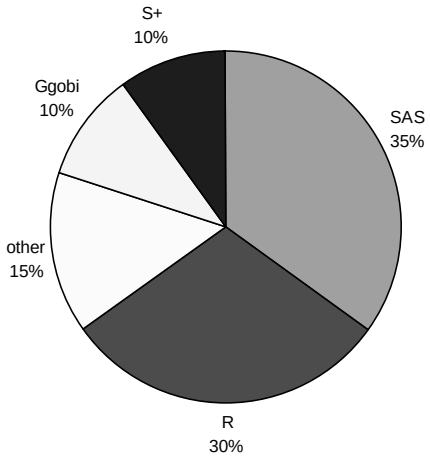

Figure 5. Library/Info Science Data Mining Software by Brand Name (n = 21).

Business Data Mining. McGraw-Hill, Columbus, OH.
- Marakas, G. 2002. Modern Data Warehousing, Mining, and Visualization. Prentice Hall, Upper Saddle River, NJ.
- Shmueli, G. *et. al*. 2006. Data Mining for Business Intelligence. Wiley-Interscience, Hoboken, NJ.

3.4.2. Computer Science/Engineering

- Russell S., and Norvig, P. 2009. Artificial Intelligence. Prentice Hall, Upper Saddle River, NJ.
- Duda, R., Hart, P., and Stork, D. 2000. Pattern Classification Wiley- Interscience, Hoboken, NJ.
- Mitchell, T. 1997. Machine Learning. McGraw-Hill, Columbus, OH.
- Luger, G. 2008. Artificial Intelligence. Addison Wesley, Boston.
- Haykin, S. 2008. Neural Networks and Machine Learning. Prentice Hall, Upper Saddle River, NJ.
- Hagan, M. *et. al*. 2002. Neural Network Design. Hagan Publishing, Boston.
- Bishop, C. 1996. Neural Networks for Pattern Recognition. Oxford, New York.
- Tan, P. *et. al*. 2006. Introduction to Data Mining. Addison Wesley, Boston.
- Han, J. *et. al*. 2005. Data Mining: Concepts and Techniques Morgan Kaufmann, Burlington, MA.

3.4.3. Statistics

- Hastie, T., Tibshirani, R., and Friedman, J. 2009. The Elements of Statistical Learning. Springer, New York.
- Larose, D. T. 2005. Discovering Knowledge in Data. Wiley, New York.
- Hand, D. *et. al*. 2001. Principles of Data Mining. MIT Press, Cambridge, MA.
- Tan, P. *et. al*. 2006. Introduction to Data Mining.

Addison Wesley, New York.
- Ripley, B. 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, UK.

3.4.4. Library and Information Science

- Han, J. *et. al*. 2005. Data Mining. Morgan Kaufmann, Burlington, MA.
- Witten, I., and Frank, E. 2005. Data Mining. Morgan Kaufman, Burlington, MA.
- Shortliffe, E., and Cimino, J., Eds. 2006. Biomedical Informatics. Springer, New York.

3.5. Data Mining Articles by Discipline

The average annual number of published data mining articles by discipline from 1990 through mid-2009 is listed in **Figure 6**. Note that the average per year increase over the decade in data mining articles in business journals was nearly two-fold, fifteen-fold in computer science/engineering journals, seven-fold in library/ information science articles, and there was little change in the number of statistics journals. As mentioned previously, there was no specific identifier for abstracts in the main database for statistics (as in the other databases), so title/keywords were used as the closest option. Again, this is the likely the reason for the lower number of results found within statistics.

3.6. Data Mining Articles per Department across Disciplines

For the 2005-2009 period, when the number of published articles is divided by the number of departments having data mining courses, the following rate pattern emerges: 5.3 articles per business department, 9.4 articles per computer science/engineering, 0.9 articles per statistics department, and 9.9 articles per library/information science department. Earlier calculations were not generated because it is not easily apparent how long data mining courses have been offered. From this perspective, computer science/engineering and library/information science faculty have been the most productive in publishing. Again, note that the number of articles in statistics is likely underrepresented because the main database in statistics does not include abstracts as do databases in the other fields.

4. Discussion

Course offerings dealing with data mining reflect its importance within each discipline. Business courses tended to incorporate data mining as a way to become more competitive financially. Computer science/engineering courses tend to focus on the technical and logical structure of data mining. Statistics courses emphasize data mining methodologies with an eye to applications in a variety of settings as well as comparing methods to more traditional parametric statistical techniques.

Library/information science reflects a broad range of perspectives: from logical architecture of data for mining to field-specific applications of data mining (especially health and business). Generally, courses blend theory and practice. Data mining is also considered a viable research methodology in library/information science, in which case it is were more likely to be offered at the doctorate level than at the master's. In no case is data mining a required course in library/information science, although Syracuse University and Wayne State offered specializations in data management, which included data mining as an elective.

Beyond the term data mining, each discipline generated unique associated terms. Business terms focused on decision-making, management, and competition. Computer science/engineering used more technology-related and intelligence-related terms. Statistics used more methodological terms. Library/information science terms had the greatest variation, from fuzzy logic to text mining, but most terms were associated with applications (e.g.,

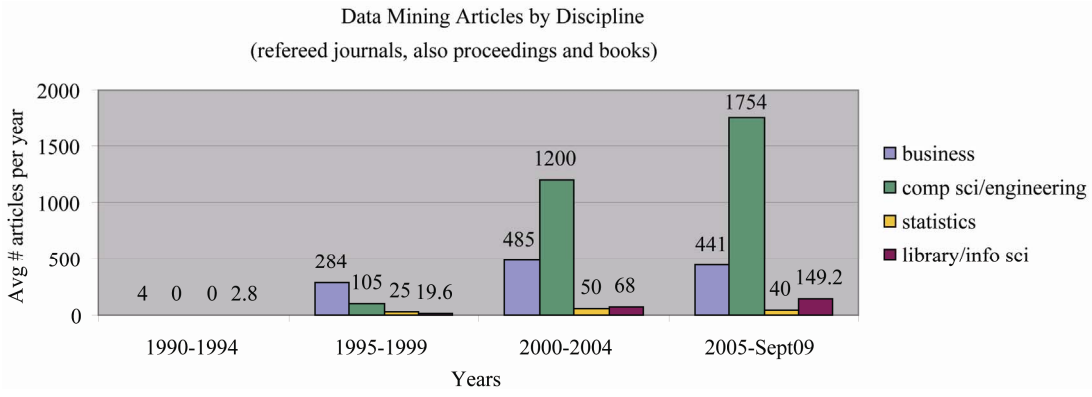

Figure 6. Average Annual Number of Data Mining Articles by Discipline from 1990-2009.

bibiometrics, health informatics, and information management). The greatest overlap existed between business and library/information science due to decision-making methodology and management issues.

Data mining software varied by discipline. SAS was the dominant software used in the business and statistics departments. Statistics had the most stable set of software brands. Matlab and C++ were the most frequently cited software in computer science/engineering courses for data mining. Computer programming languages, in general, were used by a majority of those courses. SPSS, SQL, and Excel were the dominant software used in library/information science courses. It appears that the choice of tools depended on the status of the databases to be utilized. One might assume that courses where computer programming software was used would address both database creation as well as data mining. Software also reflected the type of data needed, such as SPSS vs. RefEVAL or TextQuest. In addition, the choice of software might also reflect the technical sophistication within the academic community, with business using the least complicated software and computer science/engineering and statistics using the most complex products.

A good deal of overlap exists in textbook choices across disciplines--and in some cases within disciplines, especially for library/information science. Tan, *et al*.'s Introduction to Data Mining was used in computer science/engineering and in statistics, Han and Kamber's Data Mining was used in computer science/engineering and library/information science, and Witten and Frank's Data Mining was used in business and library/information science. Russell and Norvig's Artificial Intelligence was by far the most popular computer science/engineering textbook. Han and Kamber was the favorite title in library/information science, although Shortliffe and Cimino's Biomedical Informatics was the standard textbook for health informatics within library/information science. The picture that emerges shows little agreement on standard textbooks except in computer science/engineering. In specialized subsets of the field, such as biometrics, few titles may be available from which to choose. Instead, it appears that textbook choice depends on the specific course objectives and content focus, the academic "lens" determining the title to be used. It would be useful to survey faculty as to the basis for their textbook choice.

The number of articles over time varies by discipline. Business published the greatest number before the year 2000, but the rate leveled in the 21st century. By contrast, the library/information science article publication rate has shown a continuing rise, increasing a little over threefold from the late 1990s to the early 2000s and then a bit over twice as many in the past five years. Computer science articles rose dramatically (over tenfold) from the late 1990s to the early 2000s, and continued to rise by nearly 50% in the past five years.

A potential limitation in organizing data mining articles by discipline is that database aggregators may not have captured all relevant publications. It should be noted that another interpretation involving data mining articles by discipline is that the database aggregators may vary. In addition, deeper investigation into the quality of the articles would also shed light on the extent of scholarly contributions.

5. Conclusions

Data mining courses in the U. S. are available in various academic disciplines, and the overall field is rapidly expanding. Evidence presented in the figures and tables makes this abundantly clear. Detailed information concerning overlapping emphases in data mining disciplines has not been reported heretofore and deserves attention. Certain other academic areas include data mining courses and have associated texts and software. Nonetheless, the four disciplinary fields described in this review cover the major academic areas at this time. The emerging picture reveals a blend of theory and practice that reflects each academic discipline rather than a unified system. Hopefully, a productive merging of data mining approaches through increased cross-disciplinary research can develop and advance all these related fields. The rate of change in the data mining field is so rapid that the information is likely to be measurably different in the next ten to twenty years.

6. References

[1] M. Berry and G. Linoff, "Data Mining Techniques for Marking, Sales and Customer Support," 2nd Edition, Wiley, New York, 2004.

[2] R. Duda, P. Hart and D. Stork, "Pattern Classification," 2nd Edition. Wiley-Interscience, New York, 2000.

[3] T. Hastie, R. Tibshirani and J. Friedman, "*The Elements of Statistical Learning,*" 2nd Edition. Springer, New York, 2009. doi:10.1007/978-0-387-84858-7

[4] D. Larose, "Discovering Knowledge in Data," Wiley-Interscience, Hoboken, 2005.

[5] D. Olson and Y. Shi, "Introduction to Business Data Mining," McGraw-Hill, Columbus, OH, 2006.

[6] R. Roiger and M. Geatz, "Data Mining," Addison-Wesley, Boston, 2003.

