Data Mining: How Popular Is It?

Sam Nataraj, (Email: s.nataraj@morehead-st.edu), Morehead State University
Jim Lee, Texas A&M University, Corpus Christi

ABSTRACT

Data Mining is a process used in the industry, to facilitate decision making. As the name implies, large volumes of data is mined or sifted, to find useful information for decision making. With the advent of E-business, Data Mining has become more important to practitioners. The purpose of this paper is to find out the importance of Data Mining by looking at the different application areas that have used data mining for decision making.

INTRODUCTION

Data mining is a term synonymous with *data dredging* or *fishing* and has been used to describe the process of trawling through data in the hope of identifying patterns. Data mining is only a small component to a larger study of extracting data from large databases. Over the year's databases have been growing at a phenomenal rate and will continue to grow.

What most businesses are interested in is extracting pertinent information to help get a leading edge in today's competitive environment. Many organizations are using data mining, which is the process of discovering and quantifying predictive relationships in data. The software industry has helped the spread of data mining by offering a variety of software for data mining. Some data mining software are beginning to target the casual business analyst who has a good knowledge of their business and some quantitative training.

Data mining has helped to improve the understanding, organization, and utilization of the data stored in large databases. Because of the rapid growth in data the job of data mining has become especially important in the areas of database, statistics, machine learning, and data visualization research. Data mining has also had a significant influence in scientific and business communities for tracking behavior of individuals and groups, processing medical information, and a number of other applications over several years. Data mining has been defined as the automatic discovery of patterns, associations, changes and anomalies in large data sets.

DATA MINING – THE PROCESS

David Cho and Amy Chou (1999) have summarized the data mining process as follows:

- Identify the problem domain and pursue the managerial support
- Select the data source for the project
- Clean and filter the data
- Determine the data-mining task
- Run the appropriate algorithm or combination of algorithm for pattern search
- Interpret and evaluate the output of the data mining
- Implement the results

DATA MINING USES LARGE DATABASES

Chen, et. al. (2000) listed several statistics that illustrate the huge size of databases involved in data mining. Here are a few:

- Wal-Mart makes over 20 million transactions daily.
- AT&T has 100 million customers and carries 200 million calls a day on its long distance network.
- Mobil Oil aims to store over 100 terabytes of data concerned with oil exploration.
- NASA Earth Observing System is projected to generate on the order of 50 gigabytes of data per hour.

NUMERICAL EXAMPLE OF DATA MINING

The following example taken from the publication by Two Crows Corporation explains the advantages of using the data mining approach in retail operations:

Total hardware store transactions: 1,000
Number which include "hammer": 50
Number which include "nails": 80
Number which include "lumber": 20
Number which include "hammer" and "nails": 15
Number which include "nails" and "lumber": 10
Number which include "hammer" and "lumber": 10
Number which include "hammer", "nails" and "lumber": 5

We can now calculate:

Support for "hammer and nails" = 1.5% (15/1000)
Support for "hammer, nails and lumber" = 0.5% (5/1000)
Confidence of "hammer -→ nails" = 30% (15/50)
Confidence of "nails -→ hammer" = 19% (15/80)
Confidence of "hammer and nails -→ lumber = 33% (5/15)
Confidence of "lumber -→ hammer and nails" = 25% (5/20)

Given the above example, the likelihood that a hammer buyer will also purchase nails (30%) is greater than the likelihood that someone buying nails will also purchase a hammer (19%). The prevalence of this hammer-and-nails association (i.e., the support is 1.5%) is high enough to suggest a meaningful rule.

DATA MINING USERS

Biggest users of data mining according to Chen, et. al. (2000) are:

- Credit card companies – American Express and Citibank
    - Approval of credit card applications
    - Make purchase authorization
    - Analyze cardholders' buying behavior
    - Detect fraud
- Retailers – Wal-Mart and Victoria's secret
    - Market Basket Analysis
    - Shopping Basket Analysis
    - Launch effective promotions

- Banks – Bank of America
    - Identify customers for loan campaigns
    - Enhance customer service

MINING DATA MINING

We did an extensive literature survey to find out how many journal articles are written in the field of data mining. We also wanted to establish the different business disciplines that use data mining. Table 1 summarizes our findings over the period 1996-2003. Following the pattern of e-commerce startups, the number of articles on data mining reached its peak in 2001 and declined dramatically in the past two years. About half of the articles are in science and technology instead of in business. This means that there is substantial room for research on business applications of data mining.

We also wanted to determine the different journals that publish articles dealing with data mining. We were surprised at the wide range of journals that publish data mining related articles. The list of the journals is summarized in Table 2. Table 3 summarizes the number of articles in each of the journals over an eight-year time span.

CONCLUSION

We investigated the data mining articles published in reputed journals from 1996 to 2003. We found that there were 54 different journals that have published data mining articles. Data mining as a technique has a wide variety of applications ranging from business to medicine. A list of all articles that we looked at is provided in the reference section. Our study is by no means exhaustive. We looked at a few databases. But, our study illustrates the uses of data mining and its importance in today's competitive business environment

Table 1: Classification Of Articles By Business Discipline

	2003	**2002**	**2001**	**2000**	**1999**	**1998**	**1997**	**1996**	
Finance	1	3	4	2	0	1	6	2	19
General Business	3	5	11	10	4	6	6	9	54
International Business	2	7	4	3	4	0	1	1	22
Management	4	2	5	4	5	5	6	4	35
Marketing	1	1	3	3	0	3	4	7	22
Medical	2	5	5	1	5	1	2	0	21
Technology/Science	12	8	11	15	20	15	6	3	90
	25	31	43	38	38	31	31	26	

Table 2: Index For Journals That Publish Articles Dealing With Data Mining

AB	American Banker
ACB	America's Community Banker
APAL	Annals of Pure & Applied Logic
AIM	Artificial Intelligence in Medicine
BI	Business Insurance
CRB	C.R. Biologies
CFOA	CFO Alert
CILS	Chemometrics & Intelligent Laboratory Systems
CSA	Chain Store Age
CBT	Community Banking Technology
CBC	Computational Biology & Chemistry
CSDA	Computational Statistics & Data Analysis
CN	Computer Networks
CPC	Computer Physics Communications
CPL	Chemical Physics Letters
CEA	Computers & Electronics in Agriculture
CIE	Computers & Industrial Engineering
CL	Computers in Libraries
C	Computing
DKE	Data & Knowledge Engineering
DSS	Decision Support Systems
EM	Ecological Modeling
EPSR	Electric Power Systems Research
EAAI	Engineering Applications of Artificial Intelligence
EJOR	European Journal of Operational Research
ESA	Expert Systems with Applications
FMMM	Fundamental & Molecular Mechanisms of Mutagenesis
FGCM	Future Generation Computer Systems
FSS	Fuzzy Sets & Systems
GF	Global Finance
ICTL	Information & Communications Technology Law
IC	Information & Computation
IM	Information & Management
IST	Information & Software Technology
IF	Information Fusion
IS	Information Sciences
Isy	Information Systems
IDA	Intelligent Data Analysis
IJAIT	International Journal on Artificial Intelligence Tools
JCS	Journal of Computer Security
JEM	Journal of Economic Methodology
JMB	Journal of Molecular Biology
JOCEC	Journal of Organizational Computing & Electronic Commerce
KBS	Knowledge Based Systems
MA	Management Accounting
MCM	Mathematical Computing Modeling
MB	Mortgage Banking
NU	National Underwriter
NN	Neural Networks
PC	Parallel Computing
PR	Pattern Recognition
PRL	Pattern Recognition Letters
IJPE	International Journal of Production Economics
T	Technometrics
V	Vaccine
WM	Waste Management

Table 3: Literature Survey Of Articles Dealing With Data Mining

	2003	2002	2001	2000	1999	1998	1997	1996	Total
AB							2		2
ACB								1	1
APAL	1								1
AIM			1		2	1			4
BI							1		1
CRB		1							1
CFOA							2		2
CILS		1			1				2
CSA								1	1
CBT							1		1
CBC	1								1
CSDA			1	2					3
CN				1	1				2
CPC			1						1
CPL					1				1
CEA				1					1
CIE				1		1	1		3
CL						1			1
C			1						1
DKE		1		1		1	1		4
DSS	2	1			2				5
EM	1								1
EPSR				1					1
EAAI				1					1
EJOR	2		1						3
ESA	1	3	3			1	1		9
FMMM	1								1
FGCM							1		1
FSS	1								1
GF			1						1
ICTL			1						1
IC					1				1
IM	1		1	2					4
IST					4				4
IF	1								1
IS					2	1			3
Isy	2			1		1			4
IDA					5				5
IJAIT			1						1
JCS				1					1
JEM			1						1
JMB						1			1
JOCEC				1					1
KBS		3				4			7
MA								1	1
MCM						1			1
MB								1	1
NU				1			1		2
NN	1	1							2
PC		1							1
PR			1						1
PRL	2								2
IJPE	1								1
T	1								1
V			1						1
WM				1					1
Total	19	12	15	15	19	13	9	3	

REFERENCES

1. Ananthanarayana, V.S., Narasimha, Murty M., Subramanian, D.K. Tree structure for efficient data mining using rough sets, *Pattern Recognition Letters*, Vol. 24, Iss. 6, Mar. 2003, pp 851-862.
2. Ananthanarayana, V.S., Narasimha, Murty M., Subramanian, D.K. An incremental data mining algorithm for compact realization of prototypes, *Pattern Recognition*, Vol. 34, 2001, pp 2249-2251.
3. Bose, I. and Mahapatra, R.K. Business data mining – a machine learning perspective, *Information and Management*, Vol. 39, 2001, 211-225.
4. Buydens, L.M.C., Reijmers, T.H., Beckers, M.L.M., and Wehrens, R. Molecular data mining: A challenge for chemometrics, *Chemometrics and Intelligent Laboratory Systems*, Vol. 49, 1999, pp 121-133.
5. Bykowski, A., Rigotti, C. DBC: A condensed representation of frequent patterns for efficient mining, *Information Systems*, Vol. 28, Iss. 8, Dec. 2003, pp 949-977.
6. Chen, C. and Lewis, B. A basic primer on data mining, *Information Systems Management*, Fall 2002, pp 56-60.
7. Chen, L., Sakaguchi, T., and Frolick, N. Data mining, methods, applications, and tools, *Information Systems Management*, Winter 2000, pp 65 – 70.
8. Chen, N., Zhu, D.D., and Wang, W. Intelligent materials processing by hyperspace data mining, *Engineering Applications of Artificial Intelligence*, Vol. 13, 2000, pp 527-532.
9. Chen, Q. and Mynett, A.E. Integration of data mining techniques and heuristic knowledge in fuzzy logic modeling of eurtrophication in Taihu Lake, *Ecological Modelling*, Vol. 162, Iss. 1-2, 1 April 2003, pp 55-67.
10. Cho, D.C. and Chou, A.Y. A manager's guide to data mining, *Information Systems Management*, Fall 1999, pp 33-41.
11. Clifton, C. Using sample size to limit exposure to data mining, *Journal of Computer Security*, Vol. 8, 2000, pp 281-307.
12. Cook, S. Observations on the practice of data mining: Comments on the JEM symposium, *Journal of Economic Methodology*, Vol. 8, No. 3, 2001, pp 415-419.
13. Coppola, M. and Vanneschi, M. High-performance data mining with skeleton-based structured parallel programming, *Parallel Computing*, Vol. 28, 2002, pp 793-813.
14. Daskalaki, S., Kopanas, I., Goudara, M., and Avouris, N. Data mining for decision support on customer insolvency in telecommunications business, *Europeon Journal of Operational Research*, Vol. 145, Iss. 2, Mar. 2003, pp 239-255.
15. Du, Y., Liang, Y. Data mining for seeking accurate quantitative relationship between molecular structure and GC retention indices of alkanes by projection pursuit, *Computational Biology and Chemistry*, Vol. 27, Iss. 3, July 2003, pp 339-353.
16. Duan, G., Smith, V., and Weaver, D.F. An ab initio and data mining study on aromatic-amide interactions, *Chemical Physics Letters*, Vol. 310, 1999, pp 323-332.
17. Feelders, A., Daniels, H., and Holsheimer, M. Methodological and practical aspects of data mining, *Information and Management*, Vol. 37, 2000, pp 271-281.
18. Fu, Z. Dimensionality optimization by heuristic greedy learning vs. genetic algorithms in knowledge discovery and data mining, *Intelligent Data Analysis*, Vol. 3, 1999, pp 211-225.
19. Giraud-Carrier, C. and Povel, O. Characterising data mining software, *Intelligent Data Analysis*, Vol. 7, 2003, pp 181-192.
20. Guan, T. and Wong, K. KPS: A web information mining algorithm, *Computer Networks*, Vol. 31, 1999, pp 1495-1507.
21. Hegland, M., Clarke, W., and Kahn, M. Mining the MACHO dataset, *Computer Physics Communications*, Vol. 142, 2001, pp 22-28.
22. Ho, T., Nguyen, T., Nguyen, D., and Kawasaki, S. Visualization support for user-centered model selection in knowledge discovery and data mining, *International Journal on Artificial Intelligence Tools*, Vol. 10, No. 4, 2001, 691-713.
23. Hoshi, T., Sasaki, T., Tsutsui, H., Watanabe, T., and Tagawa, F. A daily harvest prediction model of cherry tomatoes by mining from past averaging data and using topological case-based modeling, *Computers and Electronics in Agriculture*, Vol. 29, 2000, pp 149-160.

24. Hui, S.C. and Jha, G. Data mining for customer service support, *Information and Management*, Vol. 38, 2000, pp 1-13.
25. Inselberg, A. Visualization and data mining of high-dimensional data, *Chemometrics and Intelligent Laboratory Systems*, Vol. 60, 2002, pp 147-159.
26. Ishibuchi, H. and Yamamoto, T. Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining, *Fuzzy Sets and Systems*, Vol. 141, Iss. 1, 1 Jan. 2004, pp 59-88.
27. Jourdan, L., Dhaenens, C., Talbi, E.G., and Gallina, S. A data mining approach to discover genetic and environmental factors involved in multifactorial diseases, *Knowledge Based Systems*, Vol. 15, 2002, pp 235-242.
28. Kim, S.H. An architecture for advanced services in cyberspace through data mining: A framework with case studies in finance and engineering, *Journal of Organizational Computing and Electronic Commerce*, Vol. 10, No. 4, 2000, pp 257-270.
29. Kohlenback, U., Olivia, P. Proof mining in L1 – approximation, *Annals of Pure and Applied Logic*, Vol. 121, Issue 1, 2003, pp 1-38.
30. Koonce, D.A. and Tsai, S.C. Using data mining to find patterns in genetic algorithm solutions to a job shop schedule, *Computers and Industrial Engineering*, Vol. 38, 2000, pp 361-374.
31. Lavrac, N. Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine*, Vol. 19, 1999, pp 3-23.
32. Le, S., Liu, W., and Maizel, J.V. A data mining approach to discover unusual folding regions in genome sequences, *Knowledge Based Systems*, Vol. 15, 2002, pp 243-250.
33. Lee, A.J.T., Wang, Y. Efficient data mining for calling path patterns in GSM networks, *Information Systems*, Vol. 28, Iss. 8, Dec. 2003, pp 929-948.
34. Lee, K.C., Kim, J.S., Chung, N.H., and Kwo,n S.J. Fuzzy cognitive map approach to web-mining inference amplification, *Expert Systems with Applications*, Vol. 22, 2002, pp 197-211.
35. Li, S. A web-aware interoperable data mining system, *Expert Systems with Applications*, Vol. 22, 2002, pp 135-146.
36. Lin, X., Li, Y., and Tsang, C. Applying on-line bitmap indexing to reduse counting costs in mining association rules, *Information Sciences*, Vol. 120, 1999, pp 197-208.
37. Liu, L., Bhattacharyya, S., Sclove, S.L., Chen, R., and Lattyak, W.J. Data mining on time series: An illustration using fast-food restaurant franchise data, *Computational Statistics and Data Analysis*, Vol. 37, 2001, pp 455-476.
38. Luan, J. Data mining and its applications in higher education, *New Directions for Institutional Research*, No. 113, Spring 2002, pp 17-36.
39. Manganaris, S., Christensen, M., Zerkle, D., and Hermiz, K. A data mining analysis of RTID alarms, *Computer Networks*, Vol. 34, 2000, 571-577.
40. Melab, N. Data mining: A key contribution to e-business, *Information and Communications Technology Law*, Vol. 10, No. 3, 2001.
41. Menczer, F. Complementing search engines with online web mining agents, *Decision Support Systems*, Vol. 35, Iss. 2, May 2003, pp 195-212.
42. Moshkovich, H.M., Mechitov, A.I., and Olson, D.L. Rule induction in data mining: Effect of ordinal scales, *Expert Systems with Applications*, Vol. 22, 2002, pp 303-311.
43. Ng, M.K. and Huang, Z. Data mining massive time series astronomical data: Challenges, problems and solutions, *Information and Software Technology*, Vol. 41, 1999, pp 545-556.
44. Ngan, P.S., Wong, M.L., Lam, W., Leung, K.S., and Cheng, J.C.Y. Medical data mining using evolutionary computation, *Artificial Intelligence in Medicine*, Vol. 16, 1999, pp 73-96.
45. Niu, M.T., Erwin, D.E., and Braun, M.M. Data mining in the US vaccine adverse event reporting system (VAERS): Early detection of intussusception and other events after rotavirus vaccination, *Vaccine*, Vol. 19, 2001, pp 4627-4634.
46. Orre, R., Lansner, A., Bate, A., and Lindquist, M. Bayesian neural networks with confidence estimations applied to data mining, *Computational Statistics and Data Analysis*, Vol. 34, 2000, pp 473-493.
47. Parsons, O. and Carpenter, G.A. ARTMAP neural networks for information fusion and data mining: Map production and target recognition methodologies, *Neural Networks*, article accepted and in press, available online.

48. Piramuthu, S. Evaluating feature selection methods for learning in data mining applications, *European Journal of Operational Research*, article accepted and in press, available online.
49. Pudi, V. and Haritsa, J.R. Quantifying the utility of the past in mining large databases, *Information Systems*, Vol. 25, No 5, 2000, pp 323-343.
50. Rombel, A. CRM shifts to data mining to keep customers, *Global Finance*, Oct. 2001, Vol. 15, Iss. 11, pp 97-99.
51. Rosenkranz, H.S. A data mining approach for the elucidation of the action of putative etiological agents: Application to the non-genotoxic carcinogenicity of genistein, *Mutation Research*, Vol. 526, Iss. 1-2, May 2003, pp 85-92.
52. Roussinov, D. and Zhao, J.L. Automatic discovery of similarity relationships through Web mining, *Decision Support Systems*, Vol. 35, Iss. 1, 2003, pp 149-166.
53. Schikora, P.F., Godfrey, M.R. Efficacy of end-user neural network and data mining software for predicting complex system performance, *International Journal of Production Economics*, Vol. 84, Iss. 3, 11 June 2003, pp 231-253.
54. Scott, P.D. and Wilkins, E. Evaluating data mining procedures: Techniques for generating artificial data sets, *Information and Software Technology*, Vol. 41, 1999, pp 579-587.
55. Shen, L., Shen, H., and Cheng, L. New algorithms for efficient mining of association rules, *Information Sciences*, Vol. 118, 1999, pp 251-268.
56. Sforna, M. Data mining in a power company customer database, *Electric Power Systems Research*, Vol. 55, 2000, pp 201-209.
57. Song, H.S., Kim, J.K., and Kim, S.H. Mining the change of customer behavior in an internet shopping mall, *Expert Systems with Applications*, Vol. 21, 2001, pp 157-168.
58. Tsechansky, M.S., Pliskin, N., Rabinowitz, G., and Porath, A. Mining relational patterns from multiple relational tables, *Decision Support Systems*, Vol. 27, 1999, pp 177-195.
59. Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery*, 3rd edition, 1999.
60. Valafar, H. and Valafar, F. Data mining and knowledge discovery in proton nuclear magnetic resonance spectra using frequency to information transformation, *Knowledge Based Systems*, Vol. 15, 2002, pp 251-259.
61. Wang, S. Nonlinear pattern hypothesis generation for data mining, *Data and Knowledge Engineering*, Vol. 40, 2002, pp 273-283.
62. Wu, C.H. Data mining applied to material acquisition budget allocation for libraries: Design and development, *Expert Systems with Applications*, Vol. 25, Iss. 3, Oct. 2003, pp 401-411.
63. Yuan, B., Wang, X.Z., and Morris, T. Software analyzer design using data mining technology for toxicity prediction of aqueous effluents, *Waste Management*, Vol. 20, 2000, pp 677-686.
64. Zhang, S. Aggregation and maintenance for database mining, *Intelligent Data Analysis*, Vol. 3, 1999, pp 475-490.
65. Zhong, N., Dong, J., and Ohsuga, S. Meningitis data mining by cooperatively using GDT-RS and RSBR, *Pattern Recognition Letters*, Vol. 24, Iss. 6, Mar. 2003, pp 887-894.