

Tutorial

Database Mining

Rakesh Agrawal

IBM Almaden Research Center
San Jose, CA 95120
ragrawal@almaden.ibm.com

Abstract

We view database mining as the efficient construction and verification of models of patterns embedded in large databases. Many of the database mining problems have been motivated by the practical decision support problems faced by most large retail organizations. In the Quest project at the IBM Almaden Research center, we have focussed on three classes of database mining problems involving classification, associations, and sequences. In this tutorial, I will draw upon my Quest experience to present my perspective of database mining, describe current work, and present some open problems.

References

- [1] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, "Database Mining: A Performance Perspective", *IEEE Transactions on Knowledge and Data Engineering*, Special Issue on Learning and Discovery in Knowledge-Based Databases, December 1993.
- [2] Rakesh Agrawal, Sakti Ghosh, Tomasz Imielinski, Bala Iyer, and Arun Swami, "An Interval Classifier for Database Mining Applications", *VLDB-92*, Vancouver, British Columbia, Canada, 1992, 560-573.
- [3] Rakesh Agrawal, Tomasz Imielinski and Arun Swami, "Mining Association Rules between Sets of Items in Large Databases", *SIGMOD-93*, Washington D.C., May 1993.
- [4] R. Agrawal, C. Faloutsos, A. Swami: "Efficient Similarity Search in Sequence Databases", *4th Int'l Conf. on Foundations of Data Organization and Algorithms*, Chicago, Oct. 1993.
- [5] T.M. Anwar, S.B. Navathe, and H.W. Beck, "Knowledge Mining in Databases: A Unified Approach Through Conceptual Clustering", Georgia Institute of Technology, Atlanta, Georgia, May 1992.
- [6] R.J. Brachman *et al.*, "Integrated Support for Data Archeology", *AAAI-93 Workshop on Knowledge Discovery in Databases*, July 1993.
- [7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, 1984.
- [8] Jason Catlett, "Megainduction: A Test Flight", *8th Int'l Conf. on Machine Learning*, Morgan Kaufman, June 1991.
- [9] Peter Cheeseman *et al.*, "AutoClass: A Bayesian Classification System", *5th Int'l Conf. on Machine Learning*, Morgan Kaufman, June 1988.
- [10] Greg Cooper and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data", *Machine Learning*, 1992.
- [11] Douglas H. Fisher, "Knowledge Acquisition Via Incremental Conceptual Clustering", *Machine Learning*, 2:2, 1987.
- [12] M. Holsheimer and A. Siebes, "Data Mining: The Search for Knowledge in Databases", Report CS-R9406, CWI, Netherlands.
- [13] J. Han, Y. Cai, and N. Cercone, "Knowledge Discovery in Databases: An Attribute-Oriented Approach", *VLDB-92*, Vancouver, British Columbia, Canada, 1992, 547-559.
- [14] Ravi Krishnamurthy and Tomasz Imielinski, "Practitioner Problems in Need of Database Research: Research Directions in Knowledge Discovery", *SIGMOD Record*, 20:3, Sept. 1991, 76-78.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association of Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

- [15] P. Langley, H. Simon, G. Bradshaw, and J. Zytkow, *Scientific Discovery: Computational Explorations of the Creative Process*, MIT Press, 1987.
- [16] David J. Lubinsky, "Discovery from Databases: A Review of AI and Statistical Techniques", *IJCAI-89 Workshop on Knowledge Discovery in Databases*, Detroit, August 1989, 204–218.
- [17] *Managing Database Marketing Technology for Success*, Direct Marketing Association, Inc., New York, 1992.
- [18] R.S. Michalski, L. Kerschberg, K.A. Kaufman, and J.S. Ribeiro, "Mining for Knowledge in Databases: The INLEN Architecture, Initial Implementation, and First Results", *Journal of Intelligent Information Systems*, **1**, 1992, 85–113".
- [19] Steve Muggleton and C. Feng, "Efficient Induction of Logic Programs", In *Inductive Logic Programming*, Steve Muggleton (Ed.), 1992.
- [20] Judea Pearl, *Probablstic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufman, 1992.
- [21] J. Ross Quinlan, "Induction of Decision Trees", *Machine Learning*, **1**, 1986, 81–106.
- [22] J. Ross Quinlan, "Learning Logical Definitions from Examples", *Machine Learning*, **5:3**, 1990.
- [23] J. Ross Quinlan, *C4.5: Programs for Machine Learning*, 1993.
- [24] G. Piatetsky-Shapiro and William J. Frawley, *Knowledge Discovery in Databases*, AAAI/MIT Press, 1991.
- [25] David Shepard Associates, *The New Direct Marketing*, Business One Irwin, Homewood, Illinois, 1990.
- [26] M. Stonebraker *et al.*, "The DBMS Research at Crossroads: The Vienna Update", Invited Talk, *VLDB-93*, Dublin, Ireland, Aug. 1993.
- [27] Shalom Tsur, "Data Dredging", *IEEE Data Engineering Bulletin*, **13**, 4, December 1990, 58–63.