

73. J. S. Ribeiro, K. A. Kaufman, and L. Kerschberg. Knowledge discovery from multiple databases. In *Proc. First Int. Conf. on Knowledge Discovery and Data Mining*, pp. 240–245, Montreal, Canada, Aug. 1995.
74. A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In *Proc. 1995 Int. Conf. Very Large Data Bases*, pp. 432–443, Zürich, Switzerland, Sept. 1995.
75. P. G. Selfridge, D. Srivastava, and L. O. Wilson. IDEA: Interactive data exploration and analysis. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data*, Montreal, Canada, June 1996.
76. J. Shafer, R. Agrawal, and M. Mehta. Fast serial and parallel classification of very large data bases. *IBM Research Report*, 1996.
77. N. Shan, W. Ziarko, H. Hamilton, and N. Cercone. Using rough sets as tools for knowledge discovery. In *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD'95)*, pp. 263–268, Montreal, Canada, Aug. 1995.
78. J.W. Shavlik and T.G. Dietterich. *Readings in Machine Learning*. Morgan Kaufmann, 1990.
79. W. Shen, K. Ong, B. Mitbander, and C. Zaniolo. Metaqueries for data mining. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 375–398. AAAI/MIT Press, 1996.
80. A. Siebes. Data surveying: Foundations of interestingness in knowledge discovery. In *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD'95)*, pp. 269–274, Montreal, Canada, Aug. 1995.
81. A. Silberschatz, M. Stonebraker, and J. D. Ullman. Database research: Achievements and opportunities into the 21st century. *SIGMOD Record*, 25:52–63, March 1996.
82. A. Silberschatz and A. Tuzhilin. On subjective measure of interestingness in knowledge discovery. In *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD'95)*, pp. 275–281, Montreal, Canada, Aug. 1995.
83. E. Simoudis, B. Livezey, and R. Kerber. Using Recon for data cleaning. In *Proc. First Int. Conf. on Knowledge Discovery and Data Mining*, pp. 258–262, Montreal, Canada, Aug. 1995.
84. R. Srikant and R. Agrawal. Mining generalized association rules. In *Proc. 1995 Int. Conf. Very Large Data Bases*, pp. 407–419, Zürich, Switzerland, Sept. 1995.
85. R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data*, Montreal, Canada, June 1996.
86. K. Wang and J. Tan. Incremental discovery of sequential patterns. In *Proc. 1996 SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96)*, Montreal, Canada, June 1996.
87. S. M. Weiss and C. A. Kulikowski. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufman, 1991.
88. J. Widom. Research problems in data warehousing. In *Proc. 4th Int. Conf. on Information and Knowledge Management*, pp. 25–30, Baltimore, Maryland, Nov. 1995.
89. L. Winstone, W. Wang, and J. Han. Multiple-level data classification in large databases. *submitted for publication*, March 1996.
90. O. R. Zaïane and J. Han. Resource and knowledge discovery in global information systems: A preliminary design and experiment. In *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD'95)*, pp. 331–336, Montreal, Canada, Aug. 1995.
91. T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data*, Montreal, Canada, June 1996.
92. W. Ziarko. *Rough Sets, Fuzzy Sets and Knowledge Discovery*. Springer-Verlag, 1994.

54. Q. Li and D. McLeod. Conceptual database evolution through learning in object databases. *IEEE Trans. Knowledge and Data Engineering*, 6:205–224, 1994.
55. H. Lu, R. Setiono, and H. Liu. Neurorule: A connectionist approach to data mining. In *Proc. 21st Int. Conf. Very Large Data Bases*, pp. 478–489, Zürich, Switzerland, Sept. 1995.
56. W. Lu, J. Han, and B. C. Ooi. Knowledge discovery in large spatial databases. In *Far East Workshop on Geographic Information Systems*, pp. 275–289, Singapore, June 1993.
57. H. Mannila, H. Toivonen, and A. I. Verkamo. Discovering frequent episodes in sequences. In *Proc. First Int. Conf. on Knowledge Discovery and Data Mining*, pp. 210–215, Montreal, Canada, Aug. 1995.
58. C.J. Matheus, G. Piatetsky-Shapiro, and D. McNeil. Selecting and reporting what is interesting: The KEFIR application to healthcare data. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 495–516. AAAI/MIT Press, 1996.
59. M. Mehta, R. Agrawal, and J. Rissanen. SLIQ: A fast scalable classifier for data mining. In *Proc. 1996 Int. Conference on Extending Database Technology (EDBT'96)*, Avignon, France, March 1996.
60. R. S. Michalski. A theory and methodology of inductive learning. In Michalski et al., editor, *Machine Learning: An Artificial Intelligence Approach, Vol. 1*, pp. 83–134. Morgan Kaufmann, 1983.
61. 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. *J. Int. Info. Systems*, 1:85–114, 1992.
62. T. M. Mitchell. Generalization as search. *Artificial Intelligence*, 18:203–226, 1982.
63. S. Muggleton and L. D. Raedt. Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19:629–680, 1994.
64. R. Ng. Spatial data mining: Discovering knowledge of clusters from maps. In *Proc. 1996 SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96)*, Montreal, Canada, June 1996.
65. R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. In *Proc. 1994 Int. Conf. Very Large Data Bases*, pp. 144–155, Santiago, Chile, September 1994.
66. J.S. Park, M.S. Chen, and P.S. Yu. An effective hash-based algorithm for mining association rules. In *Proc. 1995 ACM-SIGMOD Int. Conf. Management of Data*, pp. 175–186, San Jose, CA, May 1995.
67. J.S. Park, M.S. Chen, and P.S. Yu. Efficient parallel mining for association rules. In *Proc. 4th Int. Conf. on Information and Knowledge Management*, pp. 31–36, Baltimore, Maryland, Nov. 1995.
68. G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pp. 229–238. AAAI/MIT Press, 1991.
69. G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. From data mining to knowledge discovery: An overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 1–35. AAAI/MIT Press, 1996.
70. G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.
71. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
72. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

36. J. Han and Y. Fu. Exploration of the power of attribute-oriented induction in data mining. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 399–421. AAAI/MIT Press, 1996.
37. J. Han, Y. Fu, and R. Ng. Cooperative query answering using multiple-layered databases. In *Proc. 2nd Int. Conf. Cooperative Information Systems*, pp. 47–58, Toronto, Canada, May 1994.
38. J. Han, Y. Fu, W. Wang, J. Chiang, K. Koperski, and O. R. Zaiane. DBMiner: Interactive mining of multiple-level knowledge in relational databases. In *Proc. 1996 ACM-SIGMOD Int'l Conf. on Management of Data (SIGMOD'96)*, Montreal, Canada, June 1996.
39. J. Han, Y. Fu, W. Wang, K. Koperski, and O. R. Zaiane. DMQL: A data mining query language for relational databases. In *Proc. 1996 SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96)*, Montreal, Canada, June 1996.
40. J. Han, Y. Huang, N. Cercone, and Y. Fu. Intelligent query answering by knowledge discovery techniques. In *IEEE Trans. Knowledge and Data Engineering (to appear)*, 1996.
41. J. Han, S. Nishio, and H. Kawano. Knowledge discovery in object-oriented and active databases. In F. Fuchi and T. Yokoi, editors, *Knowledge Building and Knowledge Sharing*, pp. 221–230. Ohmsha, Ltd. and IOS Press, 1994.
42. V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data*, Montreal, Canada, June 1996.
43. M. Holsheimer and A. Siebes. Data mining: The search for knowledge in databases. *CWI Report CS-R9406*, Amsterdam, The Netherlands, 1994.
44. T. Imielinski and A. Virmani. DataMine – interactive rule discovery system. In *Proc. 1995 ACM-SIGMOD Int. Conf. Management of Data*, p. 472, San Jose, CA, May 1995.
45. J. Elder IV and D. Pregibon. A statistical perspective on knowledge discovery in databases. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 83–115. AAAI/MIT Press, 1996.
46. D. Keim, H. Kriegel, and T. Seidl. Supporting data mining of large databases by visual feedback queries. In *Proc. 10th of Int. Conf. on Data Engineering*, pp. 302–313, Houston, TX, Feb. 1994.
47. J. Kivinen and H. Mannila. The power of sampling in knowledge discovery. In *Proc. 13th ACM Symp. Principles of Database Systems*, pp. 77–85, Minneapolis, MN, May 1994.
48. M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. In *Proc. 3rd Int'l Conf. on Information and Knowledge Management*, pp. 401–408, Gaithersburg, Maryland, Nov. 1994.
49. W. Klösgen. Explora: a multipattern and multistrategy discovery assistant. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 249–271. AAAI/MIT Press, 1996.
50. W. Klösgen and J. Zytlow. Knowledge discovery in database terminology. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 573–592. AAAI/MIT Press, 1996.
51. K. Koperski, J. Adhikary, and J. Han. Knowledge discovery in spatial databases: Progress and challenges. In *Proc. 1996 SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96)*, Montreal, Canada, June 1996.
52. K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proc. 4th Int'l Symp. on Large Spatial Databases (SSD'95)*, pp. 47–66, Portland, Maine, Aug. 1995.
53. C-S. Li, P.S. Yu, and V. Castelli. Hierarchyscan: A hierarchical similarity search algorithm for databases of long sequences. In *Proc. 1996 Int'l Conf. on Data Engineering*, New Orleans, Louisiana, Feb. 1996.

18. C. Clifton and D. Marks. Security and privacy implications of data mining. In *Proc. 1996 SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96)*, Montreal, Canada, June 1996.
19. S. Dao and B. Perry. Applying a data miner to heterogeneous schema integration. In *Proc. First Int. Conf. on Knowledge Discovery and Data Mining*, pp. 63–68, Montreal, Canada, Aug. 1995.
20. S. Dzeroski. Inductive logic programming and knowledge discovery. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 117–152. AAAI/MIT Press, 1996.
21. M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. In *Proc. 4th Int. Symp. on Large Spatial Databases (SSD'95)*, pp. 67–82, Portland, Maine, August 1995.
22. C. Faloutsos and K.-I. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proc. 1995 ACM-SIGMOD Int. Conf. Management of Data*, pp. 163–174, San Jose, CA, May 1995.
23. U. M. Fayyad, S. G. Djorgovski, and N. Weir. Automating the analysis and cataloging of sky surveys. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 471–493. AAAI/MIT Press, 1996.
24. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.
25. R. Feldman and I. Dagan. Knowledge discovery in textual databases (KDT). In *Proc. First Int. Conf. on Knowledge Discovery and Data Mining*, pp. 112–117, Montreal, Canada, Aug. 1995.
26. D. Fisher. Improving inference through conceptual clustering. In *Proc. 1987 AAAI Conf.*, pp. 461–465, Seattle, Washington, July 1987.
27. D. Fisher. Optimization and simplification of hierarchical clusterings. In *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD'95)*, pp. 118–123, Montreal, Canada, Aug. 1995.
28. Y. Fu and J. Han. Meta-rule-guided mining of association rules in relational databases. In *Proc. 1st Int'l Workshop on Integration of Knowledge Discovery with Deductive and Object-Oriented Databases (KDOOD'95)*, pp. 39–46, Singapore, Dec. 1995.
29. T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data*, Montreal, Canada, June 1996.
30. C. Glymour. Available technology for discovering casual models, building Bayes nets, and selecting predictors: the TETRAD II program. In *Proc. First Int. Conf. on Knowledge Discovery and Data Mining*, pp. 130–135, Montreal, Canada, Aug. 1995.
31. A. Gupta, V. Harinarayan, and D. Quass. Aggregate-query processing in data warehousing environment. In *Proc. 21st Int. Conf. Very Large Data Bases*, pp. 358–369, Zürich, Switzerland, Sept. 1995.
32. J. Han. Mining knowledge at multiple concept levels. In *Proc. 4th Int. Conf. on Information and Knowledge Management*, pp. 19–24, Baltimore, Maryland, Nov. 1995.
33. J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. *IEEE Trans. Knowledge and Data Engineering*, 5:29–40, 1993.
34. J. Han and Y. Fu. Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases. In *Proc. AAAI'94 Workshop on Knowledge Discovery in Databases (KDD'94)*, pp. 157–168, Seattle, WA, July 1994.
35. J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proc. 1995 Int. Conf. Very Large Data Bases*, pp. 420–431, Zürich, Switzerland, Sept. 1995.

Data Mining Techniques: A List of References

1. R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. *IBM Research Report*, 1996.
2. R. Agrawal, T. Imielinski, and A. Swami. Database mining: A performance perspective. *IEEE Trans. Knowledge and Data Engineering*, 5:914–925, 1993.
3. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data*, pp. 207–216, Washington, D.C., May 1993.
4. R. Agrawal, K.-I. Lin, H.S. Sawhney, and K. Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In *Proc. 21st Int. Conf. Very Large Data Bases*, pp. 490–501, Zürich, Switzerland, Sept. 1995.
5. R. Agrawal and G. Psaila. Active data mining. In *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD'95)*, pp. 3–8, Montreal, Canada, Aug. 1995.
6. R. Agrawal, G. Psaila, E. L. Wimmers, and M. Zait. Querying shapes of histories. In *Proc. 21st Int. Conf. Very Large Data Bases*, pp. 502–514, Zürich, Switzerland, Sept. 1995.
7. R. Agrawal and J. C. Shafer. Parallel mining of association rules: Design, implementation, and experience. *IBM Research Report*, 1996.
8. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 1994 Int. Conf. Very Large Data Bases*, pp. 487–499, Santiago, Chile, September 1994.
9. R. Agrawal and R. Srikant. Mining sequential patterns. In *Proc. 1995 Int. Conf. Data Engineering*, pp. 3–14, Taipei, Taiwan, March 1995.
10. C. Bettini, X. S. Wang, and S. Jajodia. Testing complex temporal relationships involving multiple granularities and its application to data mining. In *Proc. 15th ACM Symp. Principles of Database Systems*, Montreal, Canada, June 1996.
11. A. Borgida and R. J. Brachman. Loading data into description reasoners. In *Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data*, pp. 217–226, Washington, D.C., May 1993.
12. R. Brachman and T. Anand. The process of knowledge discovery in databases: A human-centered approach. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 37–58. AAAI/MIT Press, 1996.
13. P. Cheeseman and J. Stutz. Bayesian classification (AutoClass): Theory and results. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 153–180. AAAI/MIT Press, 1996.
14. M. S. Chen, J. S. Park, and P.S. Yu. Efficient data mining for path traversal patterns in distributed systems. In *Proc. 1996 Int'l Conf. on Distributed Computing Systems*, May 1996.
15. D.W. Cheung, A. W.-C. Fu, and J. Han. Knowledge discovery in databases: A rule-based attribute-oriented approach. In *Proc. 1994 Int'l Symp. on Methodologies for Intelligent Systems*, pp. 164–173, Charlotte, North Carolina, October 1994.
16. D.W. Cheung, J. Han, V. Ng, A. Fu, and Y. Fu. A fast distributed algorithm for mining association rules. *submitted for publication*, February 1996.
17. D.W. Cheung, J. Han, V. Ng, and C.Y. Wong. Maintenance of discovered association rules in large databases: An incremental updating technique. In *Proc. 1996 Int'l Conf. on Data Engineering*, New Orleans, Louisiana, Feb. 1996.

Future Research

- **Theoretical foundation of KDD and data mining.**
- **Implementation:**
 - A set of well-tuned, standard data mining operators.
 - Data and knowledge visualization tools.
 - Integration of multiple data mining strategies.
- **Knowledge discovery in advanced database systems:**
 - Data mining in heterogeneous DBs, legacy DBs, multimedia DBs, and WWW information-base.
- **New data mining methodologies & Applications:**
 - Statistical tools, probabilistic reasoning, neural nets, etc.
 - Database content browsing, query optimization, multi-resolution model, data warehousing tools, etc.
- **Challenge: Data mining as a threat to information security.**

Conclusions

- **Data mining:** A rich, promising, young field with broad applications and many challenging research issues.
- **Recent progress:** Database-oriented, efficient data mining methods in relational and transaction DBs.
- **Tasks:** Characterization, association, classification, clustering, sequence and pattern analysis, prediction, and many other tasks.
- **Domains:** Data mining in extended-relational, transaction, object-oriented, spatial, temporal, document, multimedia, heterogeneous, and legacy databases, and WWW.
- **Technology integration:**
 - Database, data mining, & data warehousing technologies.
 - Other fields: machine learning, statistics, neural network, information theory, knowledge representation, etc.

Major Knowledge Discovery Methods in KDD Systems

- Database-oriented approach: Quest, DBMiner, etc.
- OLAP-based (data warehousing) approach: MetaCube (Stanford), Redbrick Warehouse, etc.
- Machine learning: AQ15, ID3/C4.5, Cobweb, etc.
- Knowledge representation & reasoning: e.g., IMACS.
- Rough sets, fuzzy sets: Datalogic/R, 49er, etc.
- Statistical approaches, e.g., KnowledgeSeeker.
- Neural network approach, e.g., NeuroRule (Lu et al.'95).
- Inductive logic programming: Muggleton & Raedt'94, etc.
- Deductive DB integration: KnowlegeMiner (Shen et al.'96).
- Visualization approach: Keim et al.'94, etc.
- Multi-strategy mining: INLEN, KDW+, Explora, etc.

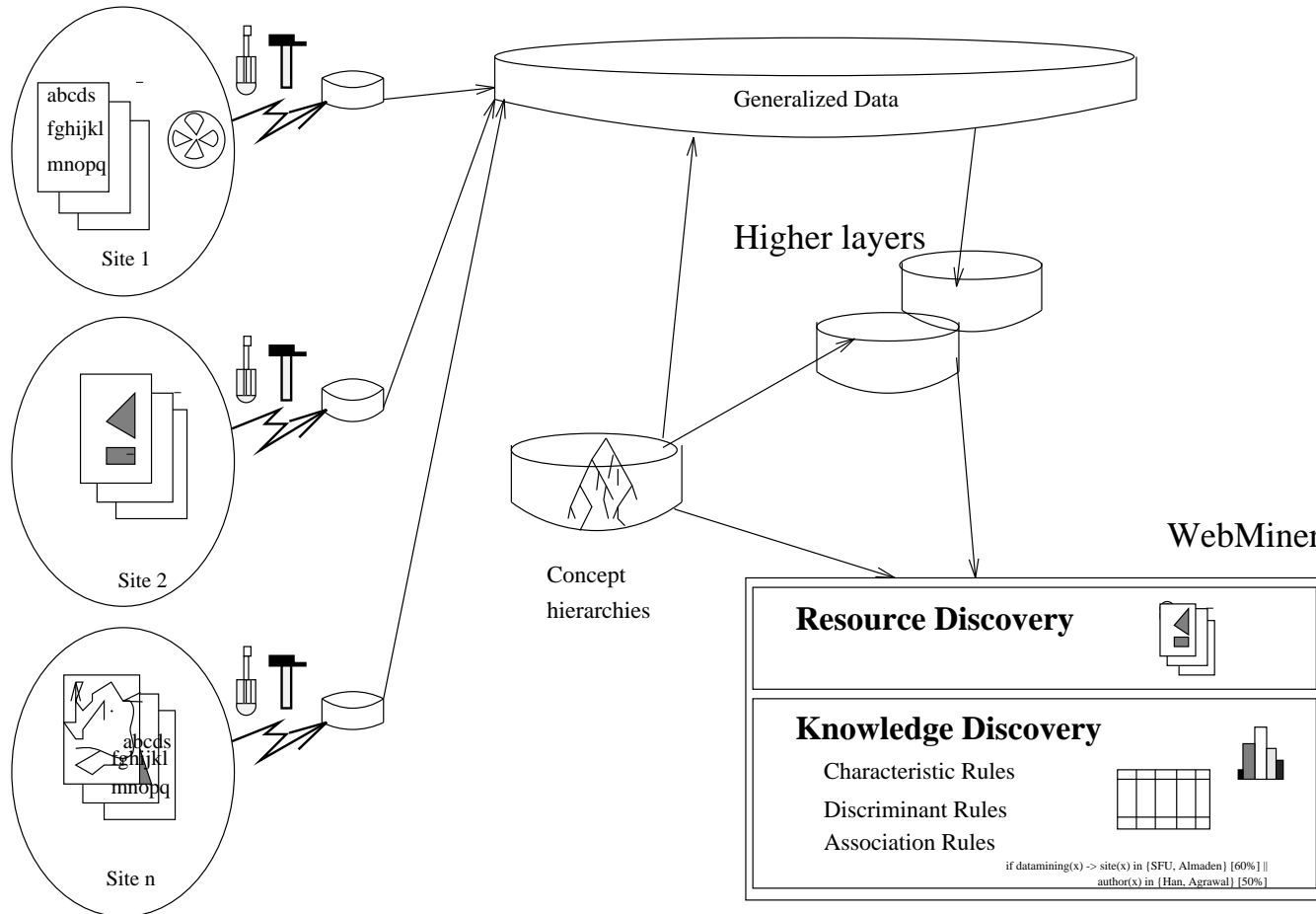
Several Data Mining Systems and/or Prototypes

- **Quest (IBM Almaden: Agrawal, et al.):** large DB-oriented association, classification, sequential patterns, similar sequences, etc.
- **DBMiner (SFU: Han, et al.):** Interactive, multi-level characterization, classif., association & prediction.
- **KDW+ (GTE: Piatetsky-Shapiro, et al.):** multi-strategy, strong rules, statistical approaches, etc.
- **Explora (GMD: Klösgen):** multi-pattern, multi-strategy discovery assistant.
- **SKICAT (JPL: Fayyad, et al.):** Large-scale sky survey.
- **IMACS (AT & T: Brachman et al.):** KR & KB construction.
- **INLEN (George Mason: Michalski, Kerschberg, et al.):** Integration of multiple learning strategies.
- **Many others, including many data warehousing systems.**

Data Mining Applications: Cooperative Query Answering

- Numerous data mining applications: Querying database knowledge, multi-level data browsing, prediction, market analysis, database design, query optimization, etc.
- Intelligent query answering by data mining techniques.
 - New primitives: Concept hierarchies, multi-layered databases, generalized relations/cubes, data mining tools.
 - Data mining power: Multi-level summaries & statistics, neighborhood info, ‘roll-up’ & ‘drill-down’ facilities.
- “What kind of houses can be bought with \$500K in L.A.?”
 - A “probe” query: Progressive information focusing.
 - Generalized answers with summary statistics.
 - Relaxation of query condition: e.g., price range.
 - Answering with “extra” info.: e.g., width extension.

Internet Information Mining: A WebMiner Proposal



Knowledge Discovery in Global Information Systems

- **Internet:** stores huge, fast growing volumes, unstructured, multimedia, heterogeneous information for diverse users.
- **A multiple layered database approach for resource and knowledge discovery in the global information-base.**
 - **Layer-0:** An unstructured, massive, primitive, diverse global information-base.
 - **Layer-1:** A relatively structured, descriptor-like, massive, distributed database by data analysis, transformation and generalization techniques.
 - **Higher-layers:** Further generalization to form progressively smaller, better structured, and less remote databases for efficient browsing, retrieval, and information discovery.

Schema Integration and Schema Evolution

- **Should data influence schema evolution? A schema should reflect the current status of data in the database.**
 - **Objects with many common attributes and properties should be grouped together.**
 - **Frequently accessed and complex structured data should have detailed classification.**
 - **Different classes should be maximally distinguished from each other.**
- **The role of data mining in schema construction.**
 - **Finds general characteristics, regularity, distribution and evolution of data in a database.**
 - **Provides multi-views for different users, but the physical one should be chosen for efficient accessing.**

Knowledge Discovery in Multi-Databases

- **A major challenge in multi-DBs: semantic heterogeneity.**
 - Multi-databases: low level heterogeneity but high-level regularity (e.g., school grading systems).
 - The role of generalization-based knowledge discovery: raise concept levels and ease semantic heterogeneity.
- **Knowledge discovery and query transformation.**
 - Not only an exchangeable “export schema” but also a common high level “language” (“vocabulary”).
 - Each local database system provides two-way transformation between low and high level data.
 - The transformation contributes to high level knowledge exchange (for KDD) and query/result interpretation (for interoperability).

Intelligent Reactions to Dynamic Environments

- **Regularity extraction.**
 - Use data sampling and KDD techniques to discover current status rules, stable rules, and evolution rules.
 - Store regular data summaries at a high level.
- **High-level active rule specification and triggering.**
- **Integration of active DB and KDD techniques.**
 - Use active DB techniques to activate a KDD process based on the importance (e.g., critical conditions) and freshness (compared with the similar situation in KBs).
 - Active dedicated sampling & knowledge discovery process at critical points.
 - Progressively refined knowledge discovery process.

Knowledge Mining in Active Databases

- **Active database technology:** automatic and prompt reaction of changes and intelligent control of dynamic environment.
- **In a dynamic environment,** data are generated rapidly, continuously, and in huge volumes.
 - **Data sampling technique:** Sample interesting pieces of information dynamically and systematically.
- **Level gap:** Data are presented at low, primitive levels.
 - It is desired to analyze the system and express the control primitives at a relatively high level.
- **Knowledge discovery technique:** Bridging the level gap.
 - Efficient and effective data generalization to discover useful knowledge or regularity at high levels.

Integration of KDD and Deductive Database Techniques

- **A knowledge-intensive data model.**
 - Data consists of primitive level data, high-level data, and meta-data.
 - Knowledge consists of expert-defined rules, discovered rules, rules defining concept hierarchies, etc.
- **Data mining by integration with deductive DB techniques.**
 - Data collection by applying deduction rules,
 - Induction using rule-specified concept hierarchy.
 - Inductive logic programming: facts \Rightarrow rules.
 - Metarule-guided data mining: Meta-rule (meta-query), such as ' $A \wedge B \rightarrow C$ ', serves as a desired pattern.
- **Knowledge integration: knowledge-base construction.**
 - Integration of discovered rules with deduction rules.

Spatial-Dominant Spatial Data Characterization

- A spatial data mining query:
 - discover characteristic rule
 - from temperature-map
 - where province = "B.C." and period = "summer"
 - and year = 1990
 - in relevance to region and temperature.
- Steps (spatial-dominant spatial data mining):
 1. Nonspatial & spatial data collection: Query processing.
 2. Spatial generalization on “region” according to certain spatial hierarchy.
 3. Nonspatial pointers are collected in the generalized and merged spatial object entries.
 4. Nonspatial generalization on “temperature”: weighted average and concept tree ascension.

Nonspatial-Dominant Spatial Data Characterization

- **A spatial data mining query:**

characterize region

from precipitation-map

where province = "B.C." and period = "spring"

and year = 1990

related to precipitation and region.

- **Steps (nonspatial-dominant spatial data characterization):**

1. Nonspatial generalization on “precipitation” (e.g., averaging and concept tree ascension to “wet”).
2. Collect spatial object pointers into generalized and merged nonspatial data entries.
3. Spatial generalization on “region” by region merging and approximation (ignoring small areas or irregular distributions).

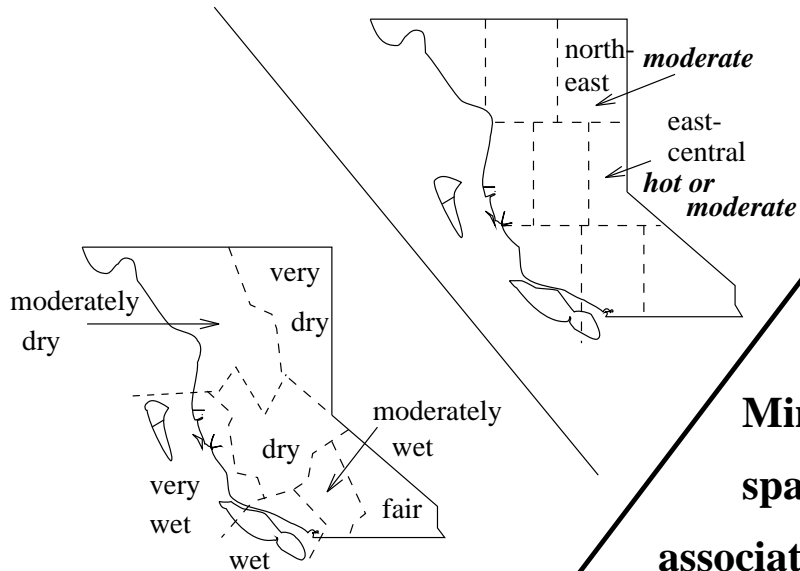
Mining Different Kinds of Knowledge in Spatial DBs

- **Several spatial data mining tasks:**
 - Spatial data characterization, classification, etc.
 - Spatial clustering analysis.
 - Spatial data association.
 - Spatial pattern analysis.
- **Spatial concept hierarchies: thematic vs. spatial.**
 - Thematic hierarchy: e.g., agriculture (food (grain (corn, rice, ...), vegetable, fruit), others(...)).
 - Spatial hierarchy, based on
 - ◇ Spatial data structures (MBR, quad-tree & R-tree).
 - ◇ Spatial related semantics (geo-region classification).
 - ◇ Clustering analysis (e.g., neighborhood or adjacent_to).
- **Primitive spatial data mining techniques:**

Nonspatial dominant vs. spatial dominant vs. interleaved.

Spatial Data Mining: A General Picture

Generalization based data mining



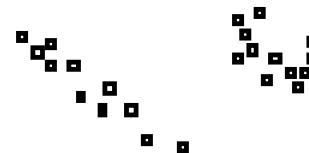
Mining using clustering techniques



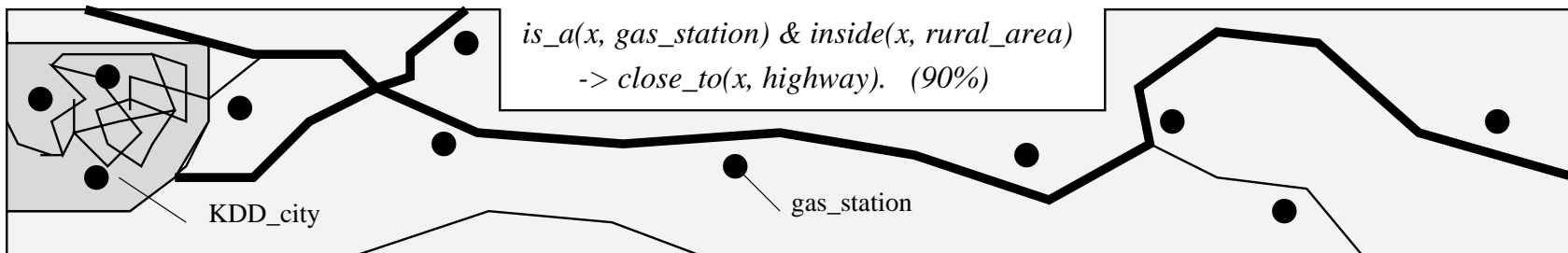
Expensive housing units in this cluster are mainly condominiums



Expensive single houses are located in these three clusters



Mining spatial association rules



Generalization of Complex Objects in OODBs

- **Structured values, tuples, trees, nested structures, etc.**

Hobby: $\{tennis, hockey, violin\} \Rightarrow \{sports(2), music(1)\}$.

Education: “ $((B.Sc. \dots), \dots, (Ph.D. in Computer Science, UCLA, Aug., 1987))$ ” $\Rightarrow (Ph.D. in CS, UCLA, 1987)$.

- ◇ gen. each attr., ◇ aggregation (sum, avg, ...), ◇ retain important attr., ◇ structure flattening, ◇ overview, ◇ typing, ◇ multi-direction \Rightarrow heterogeneous type.
- **Spatial data: clustering, region merging, using spatial data structures, spatial computation, approximation (ignoring scattered regions) & aggregation.**
- **Text, document, & hypertext: title, author, date, abstract, keywords, table of contents, content summary, etc.**
- **Graphics & images: annotation, indexing, object color, type, and size, aggregation, approximation, etc.**

Generalization-Based Data Mining in Obj.-Oriented DBs

- **Generalization on object identifiers, complex structured data, class hierarchies, class composition hierarchies, methods, spatial and multimedia data, and active data.**
 - **Object Identifier:** Gen. each OID to the lowest subclass it belongs to, and then “climb-up” the class hierarchy.
 - **Complex structured data, spatial data, multimedia data:** See the next slide.
 - **Inherited and derived data:** Treated as stored data.
 - **Methods:** Derive behavioral data by method application.
 - **Class composition hierarchy:** Generalize closely-related components.
 - **Active data:** See active data mining.
- **Then perform attribute-oriented induction, classification, etc.**

Knowledge Discovery in Advanced Database Systems

- Knowledge mining in object-oriented databases.
- Spatial data mining.
- Temporal data mining (e.g., mining data evolution regularities, and time-related pattern analysis).
- Knowledge discovery in multimedia databases (open?).
- Data mining in active databases & dynamic environments.
- Integration of data mining & deductive DB techniques.
- Knowledge discovery in multi-(heterogeneous) databases.
- Schema integration & evolution by data mining techniques.
- Resource and knowledge discovery in global information systems (WWW).

Data/Knowledge Visualization in Data Mining

- **Visualization of characteristic and discriminant rules:** tables & cubes + bar/pie charts, curves, surfaces, etc.
- **Visualization of association rules:** Nodes for large 1-itemset, lines for large 2-items sets, arrows for implication strength, etc.
- **Clustering analysis:** viewing clustering results.
- **Deviation analysis:** emphasizing deviated points.
- **Interactive data mining by data visualization:**
 - Interact with visualized result using visualization and data mining tools.
 - Visual impression of large data mining results by arranging and coloring data items as pixels (Keim et al.'94).

Deviation Analysis in Large Databases

- Major trend and characteristics vs. deviations.
 - E.g., mutual funds which perform much better (or worse) than average.
- Data deviation analysis: Discover and describe the set(s) of data which deviate from major trend/characteristics.
- A method for trend deviation analysis in large databases.
 - A-O induction on time for mining trends at multiple time scales.
 - Generalization or removal of less relevant attributes.
 - Find major and minor trends by data clustering and data distribution analysis.
 - Smoothing, similarity matching, and trend analysis.

Techniques for Mining Time-Related Knowledge

- **Time-related selection and grouping as preprocessing.**
 - E.g., select those with sales up over 20% in 1995.
- **Generalization on static data and/or along time hierarchies.**
 - E.g., aggregation on day, month, quarter, year.
- **Extension of the related, existing mining techniques.**
 - Generalization, association, classification, clustering, etc.
 - Roll-up/drill-down and progressive deepening.
- **Pattern-directed similarity matching (Agrawal et al.'95).**
 - Amplitude scaling and offset transformation.
 - Atomic sequence construction, and window stitching.
 - Approximate and fuzzy matching, & curve smoothing.
- **Visualization in trend and deviation analysis.**

Knowledge Mining with Time-Related Data

- Time-related data and temporal databases (or attributes).
- Characterize data evolution regularities:
 - Describe the firms whose sales increased 20% in 1995.
- Discriminate classes with different evolution behaviors.
 - Compare up-turn firms with the down-turn ones.
- Classify and cluster time-related data.
- Mining sequential patterns and temporal associations.
 - First buy PC then CD-ROM within 6 months.
 - Find those firms which go up together.
- Trend and deviation analysis:
 - How the stocks fluctuate in the past 12 months?
- Mining similar shapes/patterns in temporal data.
 - Shape/pattern specification and analysis: up, Up, UP.

Distance-Based Clustering Analysis in Large Databases

- **Statistical approaches:** scan data frequently, iterative optimization, hierarchical clustering, etc.
- **CLARANS (Ng & Han'94):** randomized search (sampling) + PAM (a distance-based clustering algorithm).
- **Ester et al.'95:** Improves CLARANS' efficiency by clustering only a sample of the data set drawn from the corresponding R*-tree data pages.
- **BIRCH (Zhang et al.'96):** Balanced iterative reducing and clustering using hierarchies.
 - Focus on densely occupied portions of the data space.
 - Measurement reflects the “natural” closeness of points.
 - A height-balanced tree (CF-tree) is used for clustering.
 - I/O complexity is a little more than one scan of data.

Data Clustering Analysis

- **Data clustering (“unsupervised learning”):** Cluster objects into classes, based on their features, which maximize intraclass similarity and minimize interclass similarity.
- **Probability-based vs. distance-based clustering analysis.**
- **Typical probability-based clustering analysis algorithms.**
 - **COBWEB (Fisher’87):** Incremental concept formation.
 - * **Category utility measurement** (probability of each concept’s occurrence).
 - * **Top-down, incremental, hierarchical organization of concepts.**
 - **CLASSIT (Gennari’89):** extend it to real-valued data.
 - **Integration of attribute-oriented induction & COBWEB.**
 - * **Clustering on the generalized data** (extracted by the attribute-oriented induction).

Predictive Modeling in Large Databases

- **Predictive modeling:** Predict certain data values based on similar groups of data.
 - E.g., the amount of research grants that one may obtain.
- **Limited prediction power:** on ranges or a few categories.
 - Can we predicate your ID based on others?
- **Determine the major factors which influence the prediction.**
 - **Analysis of data relevance/correlation:** Statistical techniques (e.g., χ^2 -test), decision-tree construction, expert judgement, etc.
- **Select relevant attributes, generalize less relevant data, high-level concept matching, and predict when sufficient evidence exists.**
- **Prediction will usually be a distribution.**

Generalization-Based Decision-Tree Induction

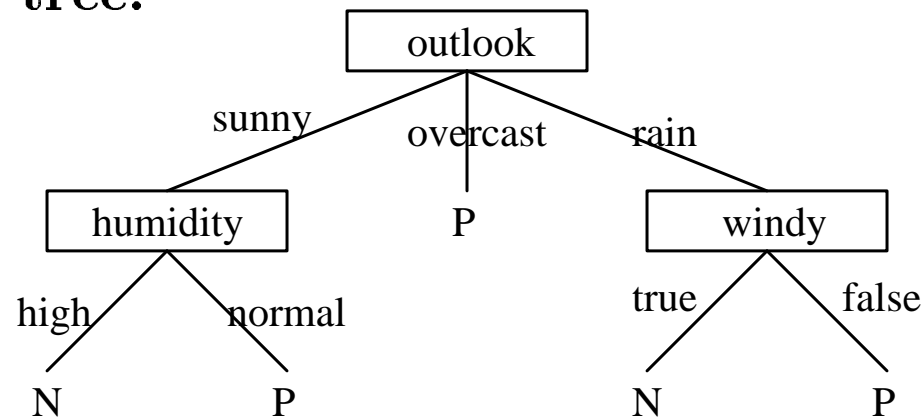
- **Integration of generalization with decision-tree induction.**
- **Classification at primitive concept levels, e.g., precise temperature, humidity, outlook, etc.**
 - **Weakness:** low-level concepts, scattered classes, bushy classification-trees, semantic interpretation problems.
- **Classification at high or medium concept levels:**
 - may lead to imprecise classification.
- **Medium level generalization & adjustment:**
 - **Generalize to intermediate concept level(s).**
 - **Merge and split concept levels for better class representation and classification accuracy.**
 - **Efficiency:** Analysis performed in compressed, generalized relations.

Scalable, Database-Oriented Classification Methods

- Scalability of decision-tree classification algorithms.
- Previous approaches:
 - Incremental tree construction (Quinlan'86): total cost is high.
 - Data sampling and discretizing continuous attributes (Cattlet'91): still in main memory.
 - Data partition and merge of parallel partition (Chan and Stolfo'91): reduced classification accuracy.
- SLIQ & SPRINT (Mehta et al.'96, Shafer et al.'96): disk-based
 - Decision-tree construction algorithms.
 - Techniques: Pre-sorting, breadth-first tree-growing, and tree-pruning.

A Decision-Tree Based Classification Method

- A decision tree:



- **ID-3 and its extended version C4.5 (Quinlan'93):** A top-down decision tree generation algorithm.
 - At start, all the training examples are at the root.
 - Partition examples recursively based on selected attributes.
 - **Attribute selection:** Maximizing an information gain measure, i.e., favoring the partitioning which makes the majority of examples belong to a single class.

Data Classification: Another Data Mining Task

- **Data categorization based on a set of training objects.**
 - **Applications:** credit approval, target marketing, medical diagnosis, treatment effectiveness analysis, etc.
 - **Example:** classify a set of diseases and provide the symptoms which describe each class or subclass.
- **The classification task:** Based on the features present in the class-labeled training data, develop a description or model for each class. It is used for
 - classification of future test data,
 - better understanding of each class, and
 - prediction of certain properties and behaviors.
- **Data classification methods:** Decision-trees (e.g., ID3, C4.5), statistics, neural networks, rough sets, etc.

Meta-Rule Guided Mining of Association Rules

- Use a meta-rule pattern, e.g., “ $P(x, y) \rightarrow Q(x, y, z)$ ”, to confine search space and help discover interesting rules.
- Meta-query guided mining (Shen et al.’96):
A meta-predicate may match relation predicates, deductive predicates (defined in LDL), attributes, etc.
- DBMiner: Restricted matching of potential predicates.

find rules in the form of

$$major(s : student, x) \wedge Q(s, y) \rightarrow R(s, z)$$

related to major, gpa, status, birth_place, address
from student

where birth_place = “Canada”

- Multiple-level rules can be discovered with such guidance.

$$major(s, \text{“Science”}) \wedge gpa(s, \text{“Excellent”}) \rightarrow status(s, \text{“Graduate”}) \quad (60\%)$$

$$major(s, \text{“Physics”}) \wedge gpa(s, \text{“3.8_4.0”}) \rightarrow status(s, \text{“M.Sc”}) \quad (76\%)$$

Mining Association Rules in Relational Databases

- **Two techniques: transaction DB vs. multi-dimens. DB.**
- **Transaction DB: Folding flat relations into nested ones.**
 - **Example.** A “course_taken” data relation,
$$\text{course_taken} = (\text{student_id}, \text{course}, \text{semester}, \text{grade})$$

can be folded into a nested relation with the schema,
$$\begin{aligned} \text{course_taken} &= (\text{student_id}, \text{course_history}) \\ \text{course_history} &= (\text{course}, \text{semester}, \text{grade}). \end{aligned}$$
- **Multi-dimens. DB method:**
 - **Multi-D cube contains support and confidence info.**
- **Integration of both: Finding relationships involving repeated and distinct attributes.**
 - **E.g., 90% senior CS students are likely to take at least three CS courses at 300-level or up in each semester.**

Interestingness Measurements for Association Rules

- Two popular measurements: support and confidence.
 - The longer (itemset), the fewer (support).
 - * Use Apriori + filtering. First minimum `min_support`, then larger thresholds for shorter itemsets.
 - The lower (level), the fewer (support).
 - * Use different support thresholds at different levels.
 - Comparable confidence: “ $A \rightarrow B$ ” vs. “ $B \rightarrow A$ ”.
- Use taxonomy information for pruning redundant rules (Srikant and Agrawal’95).
 - A rule is “redundant” if its support and confidence are close to their expected values based on an ancestor of the rule.
 - Example. “ $milk \rightarrow cereal$ ” vs. “ $skim\ milk \rightarrow cereal$ ”.
 - More effective than that based on statistical significance.

Efficient Mining of Multi-Level Association Rules

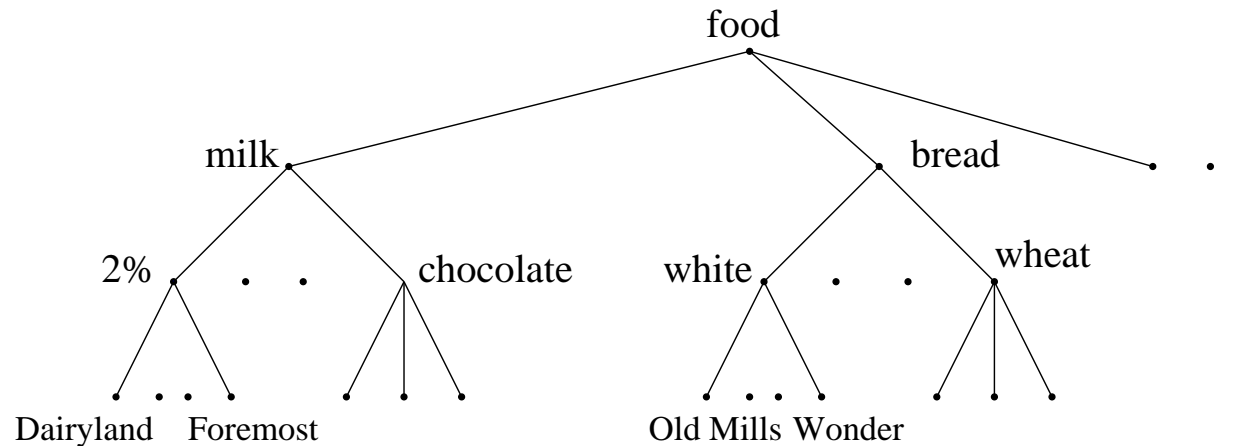
- An hierarchical information encoded transaction DB.

TID	Items
T_1	{111, 121, 211, 221}
T_2	{111, 211, 222, 323}
T_3	{112, 122, 221, 411}
T_4	{111, 121}
T_5	{111, 122, 211, 221, 413}

- Optimization at mining multiple-level association rules.
 - Progressive deepening vs. level-shared processing.
 - Candidate filtering vs. transaction table filtering.
- Variations at mining multiple-level association rules.
 - Level-crossed association rules:
 $2\% \text{ milk} \rightarrow \text{Wonder wheat bread.}$
 - Association rules with multiple, alternative hierarchies:
 $2\% \text{ milk} \rightarrow \text{Wonder bread.}$

Encoding Hierarchical Information in Transaction Database

- A taxonomy for the relevant data items.



- Conversion of bar_code into generalized_item_id.

GID	bar_code_set	category	content_spec	brand
112	{17325, 34823, 91265}	milk	2%	Foremost
141	{29394, 77454, 89157}	milk	skim	Dairy_land
171	{73295, 99184, 79520}	milk	chocolate	Dairy_land
212	{88452, 35672, 98427, 31205}	bread	wheat	Wonder
...	{..., ...}
711	{32514, 78152}	fruit_juice	orange	Minute_maid

Mining Multi-Level Associations by Progressive Deepening

- **A top-down, progressive deepening approach:**
 - **First find high-level strong rules:**
milk \rightarrow bread [20%, 60%].
 - **Then find their lower-level “weaker” rules:**
2% milk \rightarrow wheat bread [6%, 50%].
- **A shopping transaction database.**
 - **A *sales_transaction* table:** a set of $\langle T_i, \{i_p, \dots, i_q\} \rangle$.
 - **A *sales_item* (description) relation:** (bar_code, category, brand, producer, content_spec, size, storage_period, price).
- **Find the purchase patterns for fresh foods.**
 - find association rules
 - related to category, content_spec, brand
 - from sales_transactions T, sales_item I
 - where T.bar_code = I.bar_code and I.category = ‘food’
 - and I.storage_period < 21

Parallel and Distributed Mining of Association Rules

- **PDM (Park et al.'95):**
 - Use a hashing technique (DHP-like) to identify candidate k -itemsets from the local databases.
- **Count Distribution (Agrawal & Shafer'96):**
 - An extension of the Apriori algorithm.
 - May require a lot of messages in count exchange.
- **FDM (Cheung et al.'96).**
 - **Observation:** If an itemset X is globally large, there exists a partition D_i such that X and all its subsets are locally large at D_i .
 - Candidate sets are those which are also local candidates in some component database, plus some message passing optimizations.

Incremental Update of Discovered Association Rules

- Partitioned derivation and incremental updating.
- A fast updating algorithm, FUP (Cheung et al.'96).
 - View a database: original $DB \cup$ incremental db .
 - A k -itemset (for any k),
 - * large in $DB \cup db$ if large in both DB and db .
 - * small in $DB \cup db$ if small in both DB and db .
 - For those only large in DB , merge corresp. counts in db .
 - For those only large in db , search DB to update their itemset counts.
- Similar methods can be adopted for data removal and update.
- Principles are further developed for distributed/parallel mining of association rules.

Efficient Methods for Mining Association Rules

- The Apriori algorithm (Agrawal & Srikant'94).
 - At the first iteration, scan all the transactions and count the number of occurrences for each item. This derives the large (i.e., frequent) itemsets, L_1 .
 - At the k -th iteration, the candidate set C_k are those whose every $(k - 1)$ -item subset is in L_{k-1} . Scan DB and count the # of occurrences for each candidate itemset.
 - Rule extraction: confidence derived from support.
- Variations/extensions of Apriori.
 - AprioriTID (scan \bar{C}_k instead of DB) and AprioriHybrid.
 - DHP (Park, et al.'95): Use a hashing technique. A k -itemset is in C_k only if it is hashed into an entry whose value \geq min. support.
 - Partition (Savasere, et al.'95).

Mining Strong Association Rules in Transaction DBs

- **Transaction data analysis: Mining association rules.**
- **Applications: pattern association, market analysis, etc.**
- **Measurement of rule strength in a transaction DB.**

$$A \rightarrow B \text{ [support, confidence]}$$

$$\text{support} = \text{Prob}(A \cup B) = \frac{\#_of_trans_containing_all_the_items_in\ A \cup B}{total_#_of_trans}$$

$$\text{confidence} = \text{Prob}(B|A) = \frac{\#_of_trans_that_contain_both\ A\ and\ B}{\#_of_trans_containing\ A}$$

- **We are often interested in only strong associations, i.e.,**

$$\text{support} \geq \text{min_sup} \quad \text{and} \quad \text{confidence} \geq \text{min_conf.}$$

- **Examples.**

milk \rightarrow bread [5%, 60%].

tire \wedge auto_accessories \rightarrow auto_services [2%, 80%].

Integration of Data Mining & MDDB Techniques

- Mining knowledge using data cube structures.
 - Implementation of characterizer using data cube structures.
 - Storage of data mining results in data cubes.
 - Mining other kinds of knowledge (association rules, classifier, clustering, etc.) using data cubes.
- Mining multiple-level, multiple kinds of knowledge in data warehouses.
 - Extension of existing data mining techniques.
- Efficient construction of data cubes by data mining.
 - Determine data cube dimensions and generalization levels, based on data sparseness and other statistical information obtained by data mining.

Data Mining vs. Multi-Dimensional Databases

- **Characterizer:** works on any set of data and down to any granularity.
MDDB: usually on the whole DB, plus some slicing and dicing.
- **Characterizer:** usually not materialized, no index support.
MDDB: mostly are materialized or partially materialized, and/or use multidimensional indices.
- **Characterizer:** hierarchy can be automatically generated, adjusted, & balanced and thus generalization is flexible.
MDDB: mostly based on dimension “hierarchies”.
- **Data mining:** Many data mining functionalities, one of which is characterizer.
MDDB: not aimed for data mining. Data mining facilities should be further developed on top of it.

Characterizer vs. Multi-Dimensional Data Analyzer

- **Common features:**
 - Multi-dimensional analysis on large data sets.
 - Reduction of large relations to small, high-level ones.
 - Viewing data from different angles (“pivoting”).
 - Multiple-level views with “roll-up” and “drill down”.
 - Selection by “slicing” and “dicing”.
 - Grouping & aggregation on measured attributes, etc.
- **Strength of multi-dimensional analysis tools:**
 - Multiple-dimensional indexing structures.
 - Special storage structures + precomp. of aggregates.
 - Selective (and partial) instantiation of materialized views.
 - Data representation, visualization and interactive viewing facilities.

Mining Discriminant Rules: An Experiment

* The target class *					

disc_code	grant_order	amount	count%	d-weight	mark

Computer	Operating	0-20Ks	43.90%	62.07%	*
Computer	Operating	20Ks-40Ks	37.80%	56.36%	*
Computer	Operating	40Ks-60Ks	6.10%	83.33%	*
Computer	Operating	60Ks-	3.66%	100.00%	*
Computer	Equipment	0-20Ks	2.44%	66.67%	*
Computer	Equipment	20Ks-40Ks	1.22%	50.00%	*
Computer	Equipment	40Ks-60Ks	1.22%	100.00%	*
Computer	Infrastruct	0-20Ks	1.22%	100.00%	*
Computer	Infrastruct	40Ks-60Ks	1.22%	100.00%	*
Computer	Infrastruct	60Ks-	1.22%	50.00%	*

* The contrasting class *					

Computer	Operating	0-20Ks	40.74%	37.93%	*
Computer	Operating	20Ks-40Ks	44.44%	43.64%	*
Computer	Operating	40Ks-60Ks	1.85%	16.67%	*
Computer	Equipment	0-20Ks	1.85%	33.33%	*
Computer	Equipment	20Ks-40Ks	1.85%	50.00%	*
Computer	Equipment	60Ks-	3.70%	100.00%	*
Computer	Infrastruct	20Ks-40Ks	3.70%	100.00%	*
Computer	Infrastruct	60Ks-	1.85%	50.00%	*

DBMiner Query:

Distinguish Computer Science NSERC research grants of the province "B.C." from that of "Alberta" according to discipline, amount, and grant categories

```

use NSERC95
discover discriminant rule for "BC_Grants"
where O.province = "B.C."
in contrast to "Alberta_Grants"
where O.province = "Alberta"
from award A, organization O, grant_type G
where A.grant_code = G.grant_code and
      O.org_code = A.org_code
      and A.disc_code = "Computer"
in relevance to disc_code, grant_order, amount

```

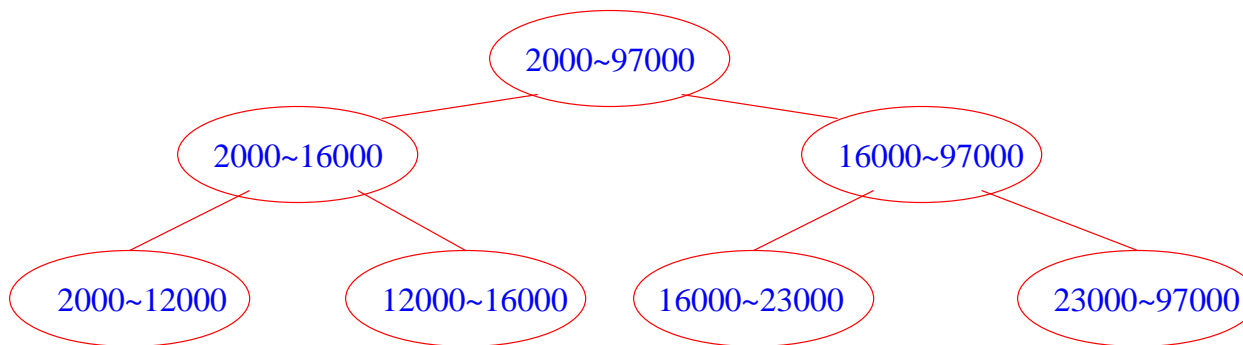
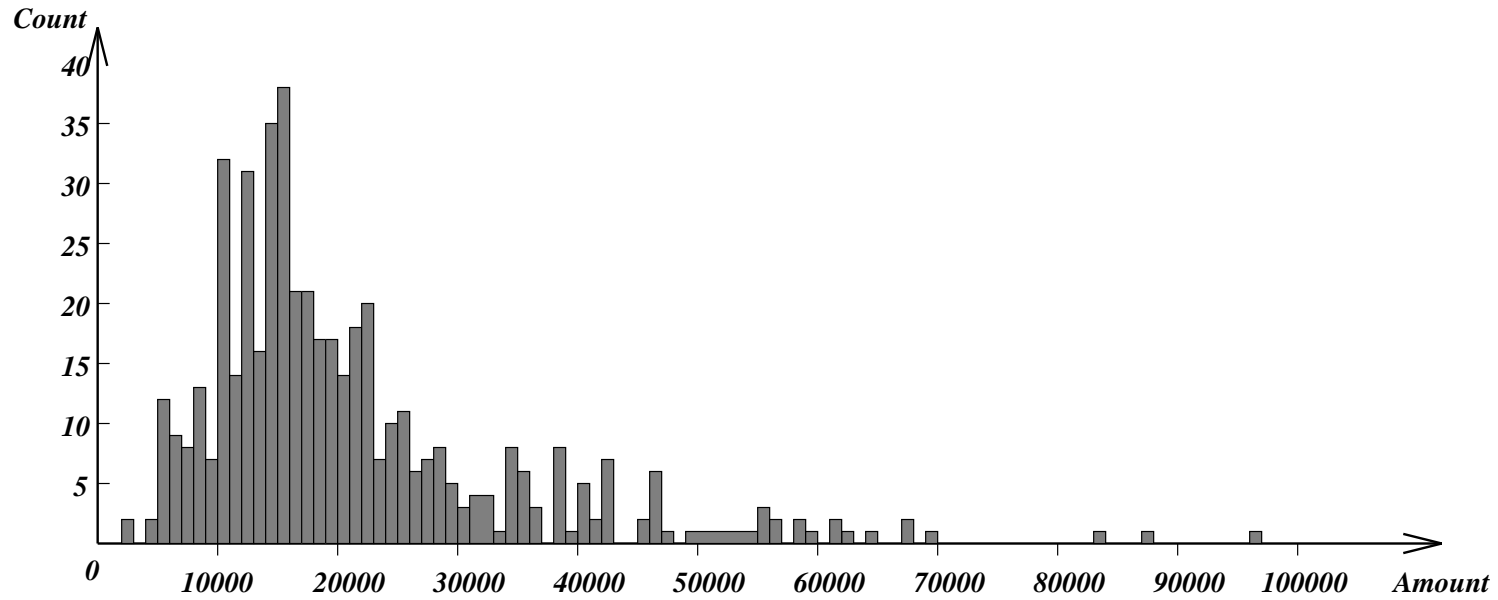
Discovery of Discriminant Rules

- Collect the relevant data respectively into the target class and the contrasting class.
- Extract the prime target relation/cube by A-O induction, and then generalize the concepts in the contrasting class to the same level as those in the prime target relation/cube, forming the prime contrasting relation/cube.
- To generate qualitative discriminant rules,
 - compare the tuples in two prime relations/cubes, mark those overlap in both, and
 - output only the unmarked tuples.
- To generate quantitative discriminant rules,
 - compute the discriminating weight for each property,
 - output those whose d-weight is close to 1 together with the d-weight.

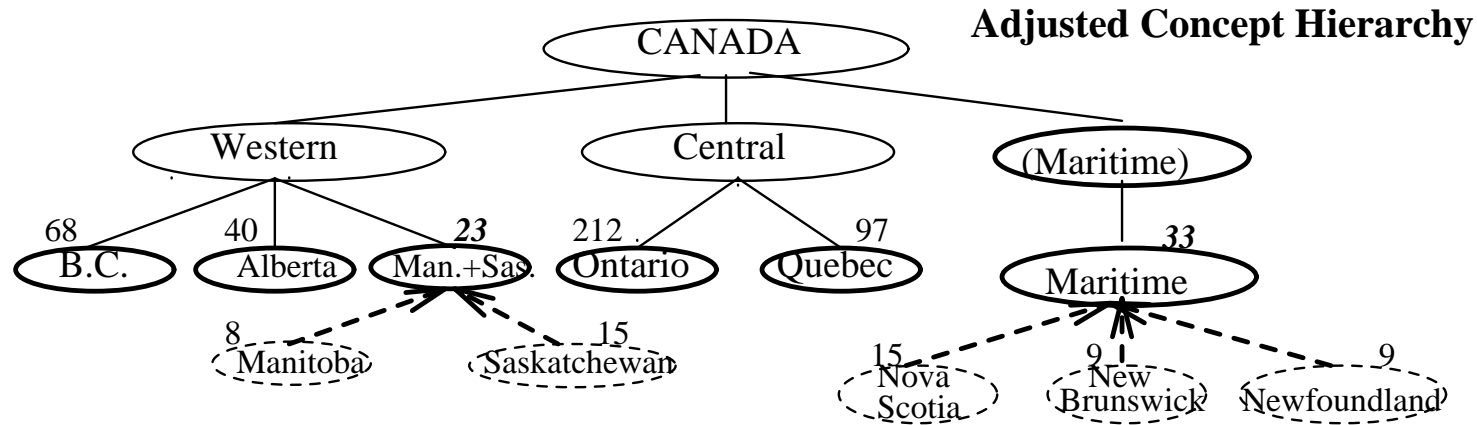
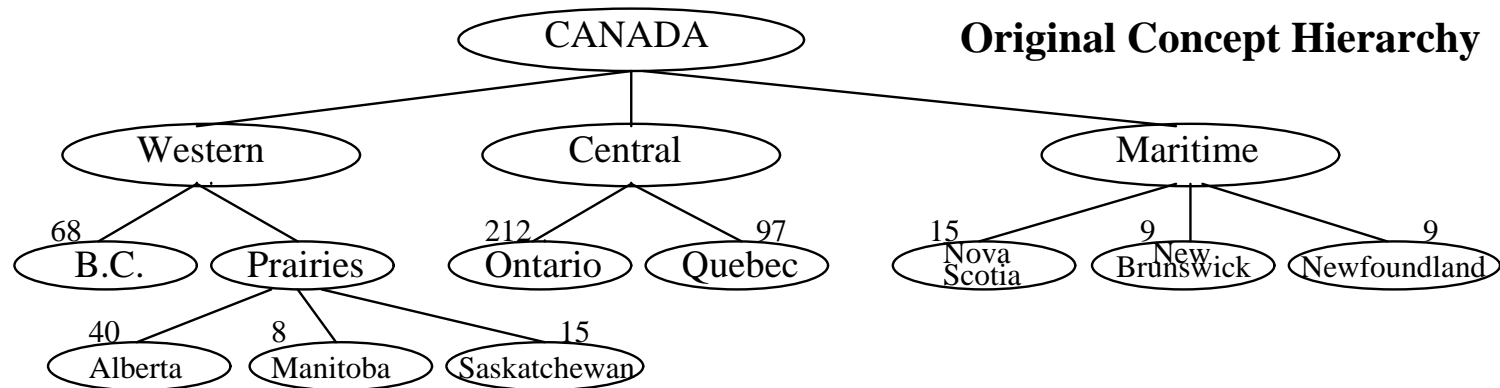
Methodologies of Multiple Level Data Mining

- **Progressive generalization (roll-up: easy to implement).**
- **Progressive deepening (drill-down: conceptually desirable).**
 - Start at a rather high level, find strong regularities at such a level
 - Selectively and progressively deepen the knowledge mining process down to deeper levels to find regularities at lower levels.
- **Interactive up and down:**
 - Roll-up and drill-down to different levels, including setting different thresholds and focuses.
- **Implementation: save a “minimally generalized relation”.**
 - Specialization of a generalized relation: Generalize the minimally generalized relation to appropriate levels.

Automatic Generation of Numeric Hierarchies



Dynamic Adjustment of Concept Hierarchies



Grant Distribution in Canadian CS Departments

org_name	count%	amount%
Toronto	7.92%	12.60%
Waterloo	8.87%	10.45%
British Columbia	5.85%	7.15%
Simon Fraser	4.34%	4.97%
Concordia	4.91%	4.81%
Alberta	4.15%	4.26%
Calgary	3.77%	4.21%
McGill	3.02%	4.12%
Victoria	3.96%	3.91%
Queen's	4.34%	3.90%
Carleton	3.40%	3.54%
Western Ontario	3.77%	3.25%
Ottawa	3.40%	2.87%
York	2.45%	2.41%
Saskatchewan	2.45%	2.36%
McMaster	2.26%	2.18%
Manitoba	2.64%	2.15%
Regina	2.26%	1.76%
New Brunswick	1.89%	1.24%
Guelph	1.51%	1.21%
Memorial Univ. of Nf	1.70%	1.18%
Dalhousie	1.32%	0.90%
Windsor	1.13%	0.78%
.....		

DBMiner Query:

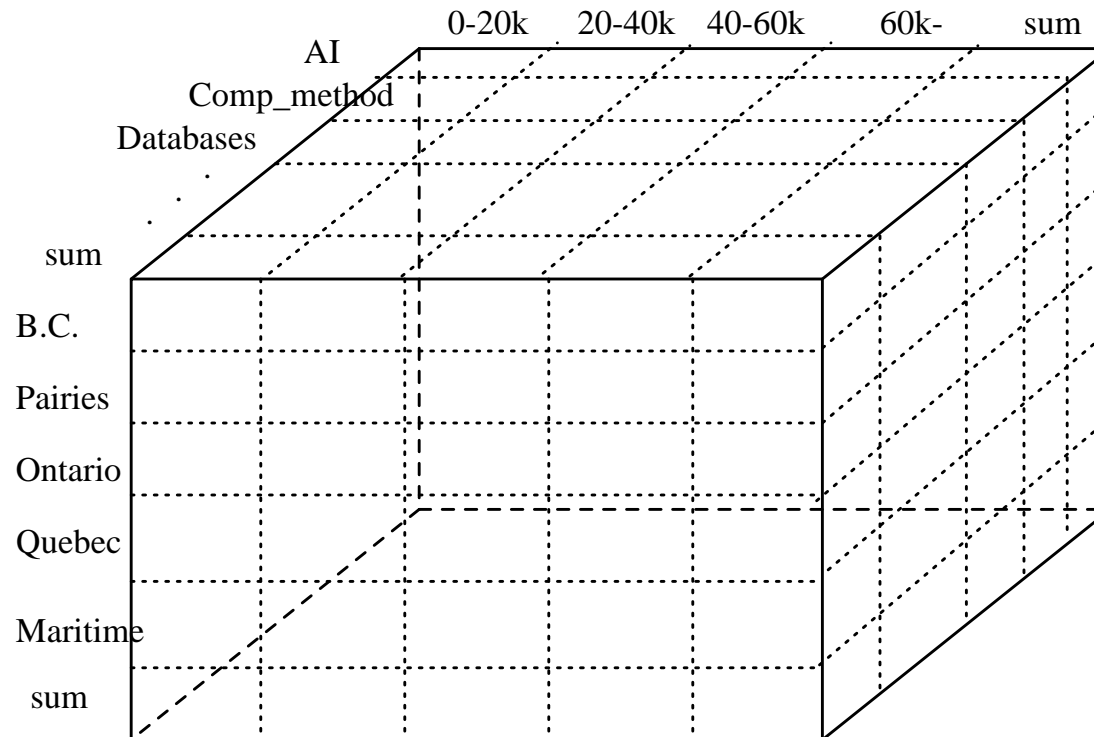
Find NSERC operating research grant distributions according to Canadian universities.

```

use NSERC96
find characteristic rule
for "CS_Organization_Grants"
from award A, organization O, grant_type G
where A.grant_code = G.grant_code and
      O.org_code = A.org_code
      and A.disc_code = "Computer"
      and G.grant_order = "Operating Grant"
related to amount, org_name, count(*)%, amount(*)%
set attribute threshold 1 for amount
unset attribute threshold for org_name

```

Data Cube Implementation of Characterization



- Each dim. of a cube represents generalized attr. values.
- A cube cell stores agr. values, e.g., count, amount.
- A “sum” cell stores dim. summation values.

Extraction of Feature Tables from Prime Relation

```

*      disc_code feature table by count%      *
*****
disc_code |amount                                     |Total      |
|         | 0-20Ks   20Ks-40Ks   40Ks-60Ks   60Ks-|         |
-----|-----|-----|-----|-----|-----|
AI        | 11.64%   6.34%   1.37%   0.68%| 20.03%|
COMP_METHODS| 7.36%   5.48%   1.03%   0.86%| 14.73%|
DATABASES  | 5.99%   2.23%   0.00%   0.17%| 8.39%|
HARDWARE   | 2.23%   1.20%   0.17%   0.68%| 4.28%|
SOFTWARE   | 11.30%   6.68%   1.03%   0.68%| 19.69%|
SYS_ORGANIZA| 4.28%   2.74%   0.68%   0.34%| 8.05%|
THEORY     | 11.82%   9.93%   2.40%   0.68%| 24.83%|
-----|-----|-----|-----|-----|
Total     | 54.62%  34.59%   6.68%   4.11%| 100.00%|

```

```

*      disc_code feature table by amount%      *
*****
disc_code |amount                                     |Total      |
|         | 0-20Ks   20Ks-40Ks   40Ks-60Ks   60Ks-|         |
-----|-----|-----|-----|-----|
AI        | 7.00%   6.99%   2.95%   2.41%| 19.35%|
COMP_METHODS| 4.31%   6.07%   2.11%   2.70%| 15.20%|
DATABASES  | 3.08%   2.61%   0.00%   0.48%| 6.17%|
HARDWARE   | 1.39%   1.37%   0.31%   3.94%| 7.01%|
SOFTWARE   | 6.73%   7.65%   2.21%   2.13%| 18.70%|
SYS_ORGANIZA| 2.83%   3.02%   1.39%   1.34%| 8.59%|
THEORY     | 6.73%  11.46%   4.86%   1.93%| 24.98%|
-----|-----|-----|-----|-----|
Total     | 32.06%  39.18%  13.84%  14.92%| 100.00%|

```

Discovery of Characteristic Rules: An Experiment

disc_code	amount	count%	amount%
AI	0-20Ks	11.64%	7.00%
AI	20Ks-40Ks	6.34%	6.99%
AI	40Ks-60Ks	1.37%	2.95%
AI	60Ks-	0.68%	2.41%
COMP_METHODS	0-20Ks	7.36%	4.31%
COMP_METHODS	20Ks-40Ks	5.48%	6.07%
COMP_METHODS	40Ks-60Ks	1.03%	2.11%
COMP_METHODS	60Ks-	0.86%	2.70%
DATABASES	0-20Ks	5.99%	3.08%
DATABASES	20Ks-40Ks	2.23%	2.61%
DATABASES	60Ks-	0.17%	0.48%
HARDWARE	0-20Ks	2.23%	1.39%
HARDWARE	20Ks-40Ks	1.20%	1.37%
HARDWARE	40Ks-60Ks	0.17%	0.31%
HARDWARE	60Ks-	0.68%	3.94%
SOFTWARE	0-20Ks	11.30%	6.73%
SOFTWARE	20Ks-40Ks	6.68%	7.65%
SOFTWARE	40Ks-60Ks	1.03%	2.21%
SOFTWARE	60Ks-	0.68%	2.13%
SYS_ORGANIZATION	0-20Ks	4.28%	2.83%
SYS_ORGANIZATION	20Ks-40Ks	2.74%	3.02%
SYS_ORGANIZATION	40Ks-60Ks	0.68%	1.39%
SYS_ORGANIZATION	60Ks-	0.34%	1.34%
THEORY	0-20Ks	11.82%	6.73%
THEORY	20Ks-40Ks	9.93%	11.46%
THEORY	40Ks-60Ks	2.40%	4.86%
THEORY	60Ks-	0.68%	1.93%
++++	13476738	100.00%	100.00%

DBMiner Query:

```

use NSERC96
discover characteristic rule
for "CS_Discipline_Grants"
from award A, grant_type G
where A.grant_code = G.grant_code
and A.disc_code = "Computer"
related to disc_code, amount,
count(*)%, amount(*)%

```

Two Implementations of Attribute-Oriented Induction

- Generalized relation vs. multi-dim. data cube approaches.
- Attribute-oriented induction: Four steps.
 1. Data collection: Collect the task-relevant data.
 2. PreGen: Prepare for generalization.
 - Derive the least generalized relation/cube.
 - Compute the desired level L_i for each attribute a_i based on its attribute threshold T_i (dynamic hierarchy adjustment if desired).
 3. PrimeGen: Derive the prime generalized relation/cube.
 4. Output transformation:
 - Roll-up or drill-down.
 - Map to generalized feature tables for a set of attributes.
 - Map to charts, curves, rules, etc.

Attribute-Oriented Induction: Basic Strategies

1. **Data focusing:** Focusing on the set of relevant data.
2. **Generalization vs. specialization:**
 - Generalization only (usually, no negative instances in DBs).
 - Avoid over generalization by **least commitment** — generalization on the smallest decomposable components.
3. **Attribute removal:** Remove
 - Nongeneralizable attrs with many distinct values (e.g., keys).
 - Attributes representing lower level concepts (e.g., address).
4. **Concept hierarchy ascension/generalization:**
 - Climbing concept trees vs. applying generalization operators.
5. **Count & aggr. value propagation & accumulation.**
6. **Attribute generalization control:**
 - Attribute threshold vs. desired level.

Mining Characteristic Rules: An Example Query

- **Characterization:** Data generalization/summarization at high abstraction levels.
- **An example query:** Find a characteristic rule for Computer Science research grants from the database 'NSERC96' in relevance to discipline, amount, and the distribution of count% and amount%.
- **Query in DMQL.**

```
use NSERC96
```

```
find characteristic rule for "CS_Discipline_Grants"
```

```
from award A, grant_type G
```

```
related to disc_code, amount, count(*)%, amount(*)%
```

```
where A.grant_code = G.grant_code and A.disc_code =  
"Computer"
```

Background Knowledge for Data Generalization

- Conceptual “hierarchies” and generalization operators.
 - Instance-based: $\{freshman, \dots, senior\} \subset undergraduate$.
 - Schema-based: $address(city, province, country)$.
 - Rule-based: $good(x) \leftarrow undergraduate(x) \wedge gpa(x) \geq 3.5$.
 - Aggregate or approx. operations, e.g., avg, sum , etc.
- Where to get such background knowledge?
 - Implicitly stored in databases, such as $address$.
 - Explicitly defined by experts, such as “ $physics \subset science$ ”.
 - Formed with different attribute combinations,
 $food(category, brand, content_spec, package_size, price)$.
 - Generated automatically by data distribution analysis.
- May need dynamic adjustment for a particular set of data.
- Choose from multiple hierarchies or try them in parallel.

High-level and Multiple-Level Data Characterization

- Generalizing, summarizing, and browsing data at high or multiple abstraction levels and from different angles.
 - Data in DBs are often expressed at primitive levels.
 - Knowledge is usually expressed at high levels.
 - Data may imply concepts at multiple levels:
Tom Jackson \in CS grad \subset student \subset person.
- Mining knowledge just at single abstraction level?
 - Too low level? — Raw data or weak rules.
 - Too high level? — Not novel, common sense?
- Mining knowledge at multiple levels:
 - Provides different views and different abstractions.
 - Progressively focuses on “interesting” spots and deepens the data mining process.

Knowledge Discovery Module of DBMiner

DBMiner: Discovery Modules

Characterizer

Discriminator

Classifier

**Association
Rule Finder**

**Meta-rule
Guided Miner**

Predictor

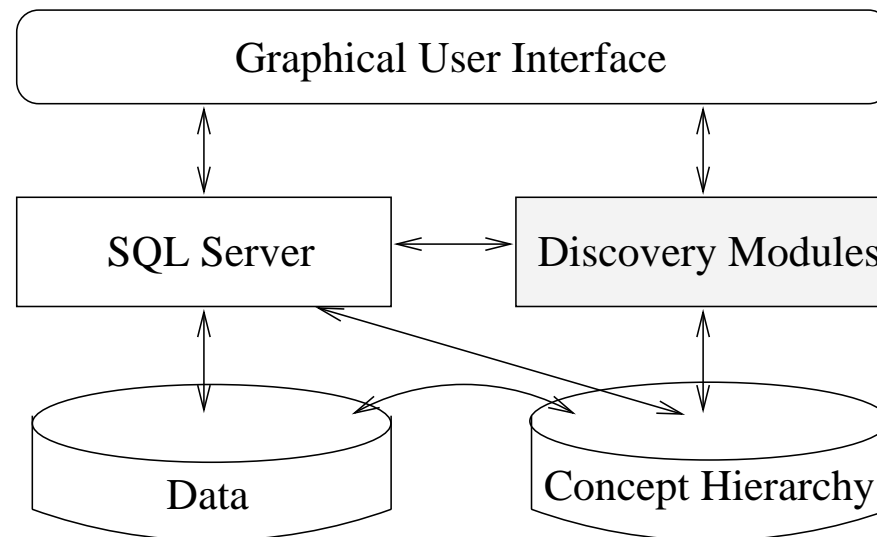
**Evolution
Evaluator**

**Deviation
Evaluator**

**Future
Modules**

Data Mining in Relational Databases: DBMiner Experience

- A generalization-based data mining tool for knowledge discovery in large relational databases.
- Multiple data mining function modules.
- Mining knowledge at multiple concept levels.
- DMQL: A data mining query language for DBMiner.



- **Architecture.**

A Data Mining Query Language: DMQL

- **Example 1.** Mining discriminant rules.

find discriminant rules

for `cs_grads` with `status = "graduate"`

in contrast to `cs_undergrads` with `status = "undergraduate"`

related to `gpa`, `birth_place`, `address`, `count(*)%`

from `student`

where `major = "cs"` and `birth_place = "Canada"`

- **Example 2.** Mining association rules.

find association rules

related to `gpa`, `birth_place`, `address`

from `student`

where `major = "cs"` and `birth_place = "Canada"`

with `support threshold = 0.05`

with `confidence threshold = 0.7`

Data Mining Interfaces

- Interactive mining (graphical input and output) vs. a data mining language.
- Specification of data mining tasks.
 - Data sets: any sets of data in DBs.
 - Mining task specification: kinds of knowledge or forms of rules to be mined.
 - Background knowledge (e.g., concept hierarchies): specification and manipulation.
 - Interestingness measurement: significance, confidence, thresholds, concept levels, etc.
- Transformation and manipulation of output results.
 - Roll-up vs. drill-down.
 - Multiple output forms: generalized relations, feature tables, charts, curves, and other visual outputs.

Mining Different Kinds of Knowledge in Large Databases

- Mining interesting knowledge at multiple abstraction levels.
- **Characterization:** Generalize, summarize, and possibly contrast data characteristics, e.g., grads/undergrads in CS.
- **Association:** Rules like “ $buys(x, milk) \rightarrow buys(x, bread)$ ”.
- **Classification:** Classify data based on the values in a classifying attribute, e.g., classify cars based on gas mileage.
- **Clustering:** Cluster data to form new classes, e.g., cluster houses to find distribution patterns.
- **Trend and deviation analysis:** Find and characterize evolution trend, sequential patterns, and deviation data, e.g., stock analysis.
- **Pattern analysis:** Find and characterize specified patterns in large DBs.

Classification of Data Mining Techniques

- **Different views, different classifications:**
 - the kinds of knowledge to be discovered,
 - the kinds of databases to be mined, and
 - the kinds of techniques adopted.
- **Knowledge to be mined: Summarization, characterization, association, classification, clustering, trend, deviation, and pattern analysis, etc.**
 - Mining knowledge at different abstraction levels: primitive level, high level, multiple-level, etc.
- **Databases to be mined: relational, transactional, object-oriented, active, spatial, temporal, textual, multi-media, heterogeneous, legacy, etc.**
- **Techniques adopted: database-oriented, machine learning, neural network, statistics, visualization, etc.**

Data Mining: Major Research Issues

- **Diversity of data mining tasks:** Summarization, characterization, classification, clustering, association, trend and deviation analysis, pattern analysis, etc.
- **Interactive mining of knowledge at multiple concept levels.**
- **Efficiency and scalability of data mining algorithms.**
- **Handling multiple-source, different kinds of data:** relational, transactional, object-oriented, active, spatial, temporal, textual, multi-media, heterogeneous, legacy, etc.
- **Expression and visualization of data mining results.**
- **Smooth integration with the existing database and data warehousing systems.**
- **Knowledge update, integration, and application.**
- **Data mining security:** guard against the invasion of privacy.

Motivation: “Necessity is the Mother of Invention”

- **Data mining (knowledge discovery in databases) —**
Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases.
[Other names: *knowledge mining from data, knowledge extraction, data dredging, data archaeology, data/pattern analysis,*]
- **Data explosion:** Automated data generation and gathering leads to tremendous amounts of data stored in databases.
- **We are drowning in data, but starving for knowledge!**
- **Data mining:** From data to “information,” “knowledge,” “regularity,” “overview,” and to “automatic construction of knowledge-bases.”

Data Mining Techniques: Tutorial Outline

- **Data Mining: Demands, Potential, and Major Issues.**
- **Classification of Data Mining Techniques.**
- **Generalization, Summarization, and Characterization.**
- **Mining Association Rules.**
- **Classification, Prediction, and Clustering Techniques.**
- **Discovery and Analysis of Patterns, Trends, and Deviations.**
- **Mining Knowledge in Advanced or Specialized Database Systems or Information Systems.**
- **Data Mining Systems and Prototypes.**
- **Data Mining Applications.**
- **Conclusions and Future Research.**

DATA MINING TECHNIQUES

(ACM-SIGMOD'96 CONFERENCE TUTORIAL)

Jiawei Han

Database Systems Research Laboratory

School of Computing Science

Simon Fraser University, Canada

June 1996