THE 17TH EUROPEAN CONFERENCE
ON MACHINE LEARNING
AND THE 10TH EUROPEAN CONFERENCE
ON PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES

~ECML/PKDD 2006~

September 18th, 2006, Berlin, Germany

Title: Agent Intelligence through Data Mining

Presenters:

Andreas L. Symeonidis **Pericles A. Mitkas**

Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki

& Laboratory of Intelligent Systems and Software Engineering,

Informatics and Telematics Institute / CERTH

Thessaloniki, Greece

Contact Info:

Department of Electrical and Computer Engineering

Aristotle University of Thessaloniki -54124

Thessaloniki, Greece

E-mail: asymeon@iti.gr	E-mail: mitkas@auth.gr
Tel.: +30 2310 99 6399	Tel.: +30 2310 99 6390
Fax: +30 2310 99 6398	Fax: +30 2310 99 6398

References

A

1. Ackley D.H. & M.L. Littman, 1992. "*Learning from Natural Selection in an Artificial Environment"* in Artificial Life II Video Proceedings, C.G. Langton, Ed.. Redwood City, California, Addison-Wesley..
2. Adriaans P. & D. Zantige, 1996. *Data Mining*. Addison-Wesley.
3. Agent Working Group, 2000. *Agent Technology Green Paper*, Object Management Group.
4. Agrawal R., C. Aggarwal, & V. Prasad, 1999. "A tree projection algorithm for generation of frequent itemsets", in *Proceedings of High Performance Data Mining Workshop*, Puerto Rico.
5. Agrawal R. & R. Srikant, 1994. "Fast algorithms for mining association rules", in *Proceedings of the 20th VLDB Conference*, Santiago, Chile, pp. 487-499.
6. Agrawal R. & R. Srikant, 1995. "Mining Sequential Patterns", in *Proceedings of the International Conference on Data Engineering (ICDE)*, Taipei, Taiwan.
7. Amir A., R. Feldman, & R. Kashi, 1997. "A new and versatile method for association generation", *Information Systems*, vol. 22, no. 6-7, pp. 333-347.
8. Arthur B.W., 1994. "Inductive Reasoning and Bounded Rationality", *American Economic Review*, vol. 84, no. 2, pp. 406-411.
9. Athanasiadis I.N. & P.A. Mitkas, 2004. "An agent-based intelligent environmental monitoring system", *Management of Environmental Quality*, vol. 15, no. 3, pp. 229-237.

B

10. Bellifemine F., A. Poggi, & G. Rimassa, 2000. "Developing multi-agent systems with JADE", in Seventh International Workshop on Agent Theories, Architectures, and Languages, Boston MA.
11. Bigus J.P., 1996. Data Mining with Neural Networks Solving Business Problems from Application Development to Decision Support. Mc Graw-Hill.
12. Booker L., D.E. Goldberg, & J.H. Holland, 1989. "Classifier systems and genetic algorithms ", Artificial Intelligence, vol. 40, no. 1-3, pp. 235-282.
13. Bossel H., 1977. "Orientors of Nonroutine Behavior" in Concepts and Tools of Computer-Assisted Policy Analysis, H. Bossel Ed. pp. 227-265. Basel: Birkhauser, Verlag.
14. Bousquet F., C. Cambier, & P. Morand, 1994. "Distributed Artificial Intelligence and Object-Oriented Modelling of a Fishery", Mathematical Computation Modelling, vol. 20, no. 8, pp. 97-107.

C

15. Caglayan A., C. Harrison, & C.G. Harrison, 1997. Agent Sourcebook: A Complete Guide to Desktop, Internet, and Intranet Agents. John Wiley & Sons.
16. Carlsson C. & E.Turban, 2002. "DSS: directions for the next decade", Decision Support Systems, vol. 33, pp. 105-110.
17. Caswell H., 1989. Matrix population models: Construction, analysis, and interpretation. Sunderland, MA: Sinauer Associates.
18. Chen M.S., J. Han, & P.S. Yu, 1996. "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883.
19. Chen Z., 1999. Computational Intelligence for Decision Support. CRC Press, Boca Raton.
20. Choy K.L., W.B. Lee, & V. Lo, 2002. "Development of a case based intelligent customer-supplier relationship management system", Expert Systems with Applications, vol. 23, no. 3, pp. 281-297.
21. Choy K.L., W.B. Lee, & V. Lo, 2003. "Design of an intelligent supplier relationship management system: a hybrid case based neural network approach", Expert Systems with Applications, vol. 24, no. 2, pp. 225-237.
22. Crist T.O. & J.W. Haefner, 1994. "Spatial Model of Movement and Foraging in Harvester Ants (Pogonomyrmex) (II): The Roles of Environment and Seed Dispersion", Journal of Theoretical Biology, vol. 166, pp. 315-323.

D

23. Davenport T.H., 2000. "The future of enterprise system-enabled organizations", Information Systems Frontiers, vol. 2, no. 2, pp. 163-180.
24. Dean J., 1998. "Animats and what they can tell us", Trends in Cognitive Sciences, vol. 2, no. 2, pp. 60-67.
25. DeAngelis D. L.& L.J. Gross, 1992. Individual-based models and approaches in ecology: Populations, communities and ecosystems. Chapman and Hill, New York.
26. Durrett R. & S.A. Levin, 1994. "Stochastic spatial models: A user's guide to ecological applications", Philosophical Transactions of the Royal Society of London, vol. 343, (B), pp. 329-350.

E

27. Epstein J.M. & R.L. Axtell, 1996. Growing Artificial Societies: Social Science from the Bottom Up. The MIT Press, Washington.

F

28. Farquhar A., R. Fikes, & J. Rice, 1996. "The Ontolingua Server: A tool for Collaborative Ontology Construction", Knowledge Systems Laboratory, Stanford University, Technical Report KSL-96-26.

29. Fayyad U., 1996. "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the Technical Committee on Data Engineering, vol. 21, no. 1, pp. 39-48.
30. Fayyad U., G. Piatetsky-Shapiro, & P. Smyth, 1996. "Knowledge Discovery and Data Mining: Towards a unifying framework", in Proceedings of The Second International Conference on Knowledge Discovery and Data Mining, Portland, USA, pp. 82-88.
31. Ferber J., 1999. Multi-Agent Systems – An introduction to Distributed Artificial Intelligence. Addison-Wesley, London.
32. Fernandes A.A.A., 2000. "Combining Inductive and Deductive Inference in Knowledge Management Tasks", in Proceedings of the 11th International Workshop on Database and Expert Systems Applications - TAKMA 2000, IEEE Computer Society, pp. 1109-1114.
33. Freitas A.A., 1999. "On Rule Interestingness measures", Knowledge-Based Systems, vol. 12, no. 5-6, pp. 309-315.
34. Friedman-Hill E.J., 2003. Jess, The Expert System Shell for the Java Platform, version 6.1. Available: http://herzberg.ca.sandia.gov/jess.

G

35. Galitsky B. & R. Pampapathi, 2003. "Deductive and inductive reasoning for processing the claims of unsatisfied customers", in Proceedings of the 16th Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE 2003), Springer-Verlag, Heidelberg, pp. 21-30.
36. Ganti V., J. Gehrke, & R. Ramakrishman, 1999. "Mining Very Large Databases", Computer Magazine, vol. 32, no. 8, pp. 38-45.
37. Gasser L., 1991. "Social Conceptions of Knowledge and Action: DAI Foundations and Open Systems Semantics", Artificial Intelligence, vol. 47, pp. 107-138.
38. Genesereth M.R. & S. Ketchpel, 1994. "Software agents", Communications of the ACM, vol. 37, no. 7, pp. 48-53.
39. Goldberg D.E., 1989. Genetic Algorithms in Search, Optimization & Machine Learning. Addison-Wesley, Massachusetts.

H

40. Haeckel S.H. & R. Nolan, 1994. "Managing by wire", Harvard Business Review.
41. Haefner J.W. & T.O. Crist, 1994. "Spatial Model of Movement and Foraging in Harvester Ants (Pogonomyrmex) (I): The Roles of Memory and Communication", Journal of Theoretical Biology, vol. 166, pp. 299-313.
42. Han J. & M. Kamber, 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann, Burnaby.
43. Hillbrand E. & J. Stender, 1994. Many-Agent simulation and Artificial Life. IOS Press.
44. Holland J.H., 1975. Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor.
45. Holland J.H., 1987. "Genetic Algorithms and Classifier Systems: Foundations and Future Directions", in Proceedings of the second international conference on genetic algorithms and their applications, Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 82-89.

46. Holland J.H., 1995. Hidden order: How adaptation builds complexity. Addison-Wesley, Reading, MA.
47. Holsapple C.W.& M.P. Sena, 2004. "ERP plans and decision-support benefits", Decision Support Systems, to be published.
48. Hraber P.T., T. Jones, & S. Forrest, 1997. "The Ecology of Echo" in Artificial Life III, C.G. Langton Ed. Longman, Addison Wesley, pp. 165-190.

I

49. Information Discovery Inc 1999. Datamines for Data Warehousing.

J

50. Jennings N.R., 1993. "Commitements and Conventions: The Foundation of Coordination in Multi-Agent Systems", The Knowledge Engineering Review, vol. 2, no. 3, pp. 223-250.
51. Jennings N.R., J. Corera, I. Laresgoiti, E.H. Mamdani, F. Perriolat, P. Sharek, & L.Z. Varga 1996. "Using ARCHON to develop real-world DAI applications for electricity transportation management and particle accelarator control", IEEE Expert.
52. Jennings N.R., K. Sycara, & M.J. Wooldridge, 1998. "A roadmap of agent research and development", International Journal of Autonomopus Agents and Multi-Agent Systems, vol. 1, pp. 7-38.

K

53. Kaelbling L.P. & S.J.Rosenschein, 1990. Action and planning in embedded agents. The MIT Press, Cambridge.
54. Kargupta H., I. Hamzaoglou, & B. Stafford, 1996. "PADMA: PArallel Data Mining Agents for scalable text classification" in the Proceedings of High Performance Computing.
55. Kero B., L. Russell, S. Tsur, & W.M. Shen, 1995. "An Overview of Data Mining Technologies", in the KDD Workshop in the 4th International Conference on Deductive and Object-Oriented Databases, Singapore.
56. Knapik M. & J.Johnson, 1998. Developing Intelligent Agents for Distributed Systems. McGraw Hill.
57. Kodratoff Y., 1988. Introduction to Machine Learning. Pitman Publishing, London.
58. Koonce D.A., C-H. Fang, & S-C. Tsai, 1997. "A Data Mining tool for Manufacturing Systems", Computers ind.Engineering, vol. 33, no. 1-2, pp. 27-30.
59. Krebs F. & H. Bossel, 1996. "Emergent value orientation in self-organization of an animat", Ecological Modelling, vol. 96, pp. 143-164.
60. Kwon O.B. & J.J. Lee, 2001. "A multi agent intelligent system for efficient ERP maintenance", Expert Systems with Applications, vol. 21, pp. 191-202.

L

61. Langton C.G., 1994. Personal Communication.
62. Lee C., 1961. "An algorithm for path connections and its applications", IRE Trans Electron.Computers, vol. 10, pp. 346-365.
63. Levi S.D., P. Kaminsky, & S.E. Levi, 2000. Designing and managing the supply chain. McGraw-Hill, Illinois.
64. Looney C.G., 1997. Pattern Recognition Using Neural Networks: Theory and Algorithms for Engineers and Scientists. Oxford University Press.

M

65. MacQueen J., 1967. "Some methods for classification and analysis of multivariate observations", in Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, pp. 281-297.
66. Mahalingam K. & M.N. Huhns, 1997. "An Ontology Tool for Distributed Information Environments", IEEE Computer, vol. 30, no. 6, pp. 80-83.
67. Malone T.W., 1998. "Inventing the organizations of the twentieth first century: control, empowerment and information technology", in Sense and Respond: Capturing Value in the Network Era, S.P. Bradley & R. Nolan, Eds. Harvard Business School Press, Boston MA, pp. 263-284.
68. May R.M., 1973. Stability and Complexity in model ecosystems Princeton University Press, Princeton, N. J.
69. Mitkas P.A., A.L. Symeonidis, D. Kehagias, & I. Athanasiadis, 2002. "An agent framework for dynamic agent retraining: Agent academy", in Challenges and Achievements in e-business and e-work Prague, pp. 757-764.
70. Mitkas P.A., D. Kehagias, A.L. Symeonidis, & I. Athanasiadis, 2003. "A Framework for Constructing Multi-Agent Applications and Training Intelligent Agents", in Proceedings of the 4th International Workshop on Agent-Oriented Software Engineering (AOSE-2003), Springer-Verlag, Melbourne, Australia, pp. 1-16.
71. Mobasher B., 1999. "A Web personalization engine based on user transaction clustering", in Proceedings of the 9th Workshop on Information Technologies and Systems (WITS'99).
72. Mobasher B., R. Cooley, & J. Srivastava, 1999. "Creating adaptive web sites through usage-based clustering of URLs" in IEEE Knowledge and Data Engineering Workshop (KDEX'99).
73. Mobasher B., R. Cooley, & J. Srivastava, 2000. "Automatic personalization based on Web usage mining", Communications of the ACM, vol. 43, no. 8.
74. Mohammadian M., 2004. Intelligent Agents for Data Mining and Information Retrieval. Idea Group Inc..
75. Murrel D.J., J.M.J. Travis, & C. Dytham, 2002. "The evolution of dispersal distance in spatially-structured populations", Oikos, vol. 97, pp. 229-236.

N

76. Nwana H.S., 1995. "Software Agents: An Overview", The Knowledge Engineering Review, vol. 11, no. 3, pp. 205-244.

O

77. O' Conner M. & J. Herlocker, 1999. "Clustering items for collaborative filtering", in Proceedings of the ACM SIGIR Workshop on Recommender Systems, Berkeley, CA.

P

78. Papoulis A. 1991. Probability, Random Variables, and Stochastic Processes. McGraw-Hill.
79. Pecala S.W., 1986. "Neighborhood models of plant population dynamics. 2. Multispecies models of annuals", Theoretical Population Biology, vol. 29, pp. 262-292.
80. Peng Y., T. Finin, Y. Labrou, B. Chu, W. Tolone, & A. Boughannam, 1999. "A multi agent system for enterprise integration", Applied Artificial Intelligence, vol. 13, no. 1-2, pp. 39-63.
81. Perkowitz M. & O. Etzioni, 1998. "Adaptive Web sites: automatically synthesizing Web pages", in Proceedings of Fifteenth National Conference on Artificial Intelligence, Madison, WI.
82. Pilot Software 1999, White Paper: An introduction to Data Mining.

Q

83. Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. San Mateo, Morgan Kaufmann.

R

84. Ray T.S., 1992. "An approach to the synthesis of life" in Artificial Life II, C.G. Langton, C. Taylor, J.D. Farmer, and S. Rasmussen, Eds. Redwood City, CA, Addison-Wesley pp. 371-408.
85. Rosenschein J.S. & G. Zlotkin, 1994. "Designing Conventions for Automated Negotiation", AI Magazine, pp. 29-46.
86. Rousset F. & S. Gandon, 2002. "Evolution of the distribution of dispersal distance under distance-dependent cost of dispersal", Journal of Evolutionary Biology, vol. 15, pp. 515-523.
87. Rust R.T., V.A. Zeithaml, & K. Lemon, 2000. Driving customer Equity: How customer lifetime value is reshaping corporate strategy. The Free Press, New York.
88. Rygielsky C., J.C. Wang, & D.C. Yen, 2002. "Data mining techniques for customer relationship management", Technology in Society, vol. 24, no. 4, pp. 483-502.

S

89. Shahabi C., A. Zarkesh, J. Adibi, & V. Shah, 1997. "Knowledge discovery from users Web-page navigation", in Proceedings of Workshop on Research Issues in Data Engineering, Birmingham, England.
90. Shapiro J.,1999. "Bottom-up vs. top-down approaches to supply chain modeling", in Quantitative models for supply chain management, S. Tayur, R. Ganeshan, and M. Magazine Eds. Kluwer Publishing, pp. 737-759.
91. Simon H., 1996. The Sciences of the Artificial. MIT Press, MA, Cambridge.
92. Singh M.P., 1997. "Considerations on Agent Communication", in FIPA Workshop, FIPA97.
93. Spiliopoulou M. & L.C. Faulstich, 1999. "WUM: A Web Utilization Miner", in Proceedings of EDBT Workshop WebDB98, Valencia, Spain.
94. Spiliopoulou M., C. Pohle, & L.C. Faulstich, 1999. "Improving the effectiveness of a Web site with Web usage mining", in Workshop on Web Usage Analysis and User Profiling (WebKKD99), San Diego.
95. Stolfo S.J., A.L. Prodromidis, S. Tselepis, W. Lee, D.W. Fan, & P.K. Chan, 1997. "Jam: Java agents for meta-learning over distributed databases", in Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, AAAI Press Publisher, Newport Beach, CA , pp. 74-81.
96. Symeonidis A.L., D. Kehagias, & P.A. Mitkas, 2003. "Intelligent policy recommendations on enterprise resource planning by the use of agent technology and data mining techniques", Expert Systems with Applications, vol. 25, no. 4, pp. 589-602.
97. Symeonidis A.L., P.A. Mitkas, & D. Kehagias, 2002. "Mining patterns and rules for improving agent intelligence through an integrated multi-agent platform", in 6th IASTED International Conference, Artificial Intelligence and Soft Computing, Banff, Alberta, Canada.

T

98. Talavera L. & U. Cortes, 1997. "Inductive hypothesis validation and bias selection in unsupervised learning", in Proceedings of the 4th European Symposium on the Validation and Verification of Knowledge Based Systems, Leuven, Belgium, pp. 169-179.
99. The Data Mining Group, 2001. Predictive Model Markup Language Specifications (PMML), ver. 2.0. Available: http://www.dmg.org.
100. The FIPA Foundations, 2000. Foundation for Intelligent Physical Agents Specifications. Available: http://www.fipa.org.
101. The FIPA Foundations. FIPA-SL Specifications, 2000. FIPA SL Content Language Specification. Available: http://www.fipa.org/specs/fipa00008/SC00008I.html.
102. Turney P.D., 1993. "Robust Classification With Context-Sensitive Features", in 6th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, pp. 268-276.

U

103. UCI Group, 2004. UCI Machine Learning Repository. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html.

W

104. Webopedia, 2003. Online dictionary for computer and internet terms. Available: http://www.webopedia.com.
105. Weiss G., 2000. Multiagent Systems: A Modern Approach to Artificial Intelligence. The MIT Press, Massachuetts, USA.
106. Werner G.M. & M.G. Dyer, 1994. "Bioland: A Massively Parallel Simulation Environment for Evolving Distributed Foms of Intelligent Behavior", in Massively Parallel Artificial Intelligence, H. Kitano and J.A. Handler Eds. Menlo Park, California, AAAI Press/MIT Press.
107. Westerberg L. & U. Wennergren, 2003. "Predicting the spatial distribution of a population in a heterogeneous landscape", Ecological Modelling, vol. 166, pp. 53-65.
108. Wilson S.W., 1987. "Classifier Systems and the Animat Problem", Machine Learning, vol. 2, pp. 199-228.
109. Wilson S.W., 1991. "The Animat Path to AI", in From Animals to Animats: Proceedings of the First International Conference on the Simulation of Adaptive Behavior, J.A. Meyer and S.W. Wilson Eds. Cambridge, Massachussets, The MIT Press/Bradford Books.
110. Wilson S.W. & D.E. Goldberg, 1989. "A Critical Review of Classifier Systems", Proceedings of the Third International Conference on Genetic Algorithms, Morgan Kaufmann, Los Altos, California, pp. 244-255.
111. Witten I.H. & E. Frank, 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufman, New Zealand.
112. Wooldridge M. & N.R. Jennings, 1995. "Intelligent agents: Theory and practice.", The Knowledge Engineering Review, vol. 10, no. 2, pp. 115-152.
113. Wooldridge M., 1999. "Intelligent Agents". In Multiagent Systems, G. Weiss Ed. The MIT Press.
114. Worley J.H., G.R. Castillo, L. Geneste, & B. Grabot, 2002. "Adding decision support to workflow systems by reusable standard software components", Computers in Industry, vol. 49, pp. 123-140.

Y

115. Yeager L., 1994. "Computational Genetics, Physiology, Metabolism, Neural Systems, Learning, Vision, and Behavior, or Polyworld: Life in a New Context", in Artificial Life III, C.G.Langton Ed. Redwood City, California, Addison-Wesley.

Z

116. Zhang Z., C. Zhang, & S. Zhang, 2003. "An agent-based hybrid framework for database mining", Applied Artificial Intelligence, vol. 17, pp. 383-398.

CERTH

Agent Intelligence through Data Mining

AUTH

Agent Intelligence through Data Mining

Andreas L. Symeonidis

Research Associate
Informatics & Telematics Institute
Center for Engineering Research
and Technology – Hellas (CERTH)
Email:asymeon@iti.gr

Pericles A. Mitkas

Associate Professor
Electrical and Computer Engineering
Aristotle Univ. of Thessaloniki (AUTh)
and CERTH
Email: mitkas@auth.gr

Thessaloniki, GREECE

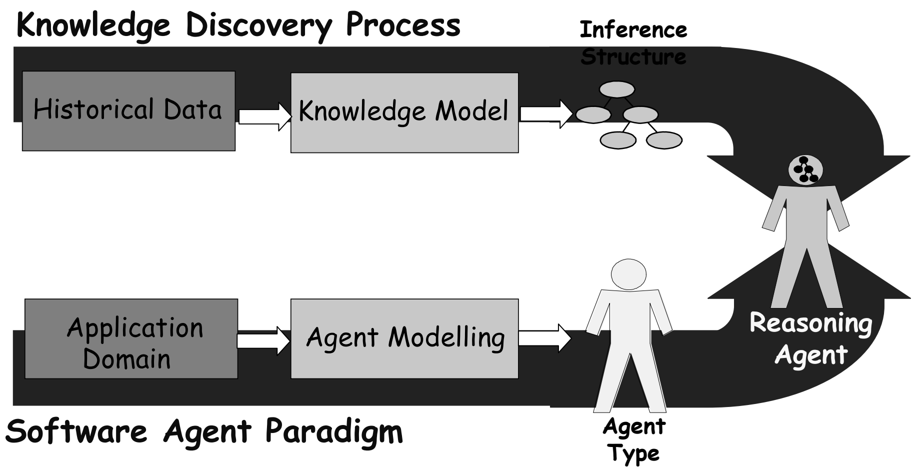
18/09/2006 ECML/PKDD 2006, Berlin, Germany 1

CERTH

Agent Intelligence through Data Mining

AUTH

The goal of the tutorial…



18/09/2006 ECML/PKDD 2006, Berlin, Germany 2

CERTH AUTH

Related Technologies

- **Data Mining (DM):** the extraction of interesting non-trivial, implicit, previously unknown and potentially useful information or patterns from data in large databases.
- **Software Agent (SA):** a software entity that acts autonomously (on behalf of another entity) in a goal-oriented manner. It is able to perceive its environment through sensors and act on it through effectors.
- **Multi-agent Systems (MAS):** a Software Engineering methodology for developing applications with the deployment of agents and agent primitives.

CERTH AUTH

Merging Technologies (2/2)

- Software agents have been repeatedly used for executing DM tasks, **but**
- DM results have not, yet, been **dynamically incorporated** to MAS

The reason: The inductive nature of DM and the lack of the appropriate tools hinders the unflustered incorporation of knowledge to MAS

Presentation Outline (1/3)

1 **Basic Primitives of DM technology**
- Data Preprocessing
- DM Techniques:
  - Cluster Analysis – Unsupervised Learning
  - Classification – Supervised Learning
  - Association Rules

2 **Data Mining & Semantics**

3 **Embedding Domain Knowledge into DM**

Presentation Outline (2/3)

4 **Data Mining Applications & Trends**

5 **Intelligent Agents**
- Definitions
- Attributes & Communication

6 **Agent Intelligence Infusion**
- The Levels of Intelligence Infusion
- Tools
- Methodologies

Presentation Outline (3/3)

7 MAS exploiting DM extracted intelligence

- An ERP add-on for intelligent CRM/SRM
- A near real-time EMS
- A decentralized maintenance management system
- A self-organizing MAS, "in danger"
- An agent-based, e-auction system

8 Open Issues

Part 1 - Introduction

Motivation

Data explosion problem

Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories

+

Knowledge Starvation

The need to see through and interpret all this "useless" data

=

Parts of the **KDD** process

Data warehousing and data mining

- ✓ **Data warehousing and on-line analytical processing**
- ✓ **Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases**

The KDD process

What is KDD?

Knowledge Discovery in Databases (KDD) is the extraction of interesting non-trivial, implicit, previously unknown and potentially useful information or patterns from data in large databases.

What is Data Mining?

Data Mining is the most important step in the KDD process, consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data.

Data Mining related technologies

- ✓ Machine Learning
- ✓ Knowledge Extraction – Extended retrieval
- ✓ Data/Pattern Analysis
- ✓ Statistical Analysis
- ✓ Business Intelligence

What is not data mining

- ✓ (Deductive) query processing.
- ✓ Expert systems or small ML/statistical programs

Data Mining vs. Machine Learning

- **The size of the dataset is different:**
  - ✓ For machine learning, datasets are loaded to main memory – thus small
  - ✓ For data mining, there is no such restriction (usually large datasets)
- **The objective is different:**
  - ✓ ML focuses on the inference mechanisms involved in the learning process
  - ✓ DM focuses on the business exploitation of extracted results

CERTH AUTH

Data Mining vs. Extended Retrieval

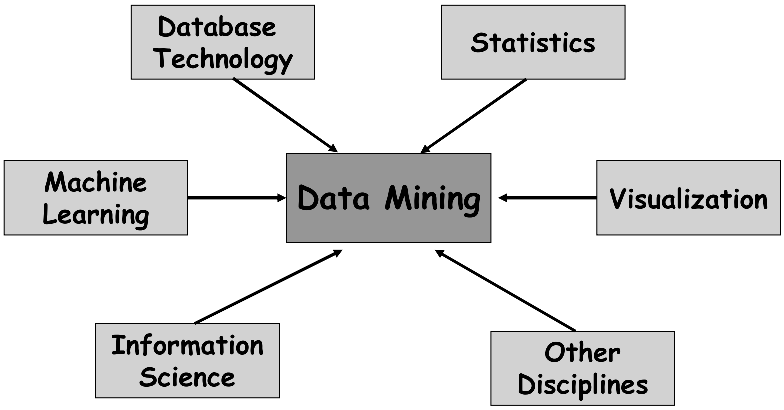
- **The approach is different:**
  - ✓ Extended Retrieval (ER) is based on individual examples (retrieved and stored as analogs), whereas DM is **ONLY** interested in a flood of data.
- **The extracted knowledge is different:**
  - ✓ DM generated knowledge contains the condensed information extracted from structured databases. In the case of ER knowledge comes through the mapping of structure information
- **ER and DM are complementary approaches**

CERTH AUTH

Data Mining vs. Statistical Analysis

- **The objective of statistical data analysis:**
  - ✓ To model the **underlying** structures in order to lead to **evaluation hypothesis**.
- **For SA Computational efficiency is not a concern**
- **Data for SA are static and clean and datasets are small**
- **The inference procedure is different:**
  - ✓ SA procedure involves repeated sampling under a given statistical model from an unknown distribution of the data
  - ✓ DM seeks to identify modeling procedures with a high probability of near-optimality over all possible distributions of data.

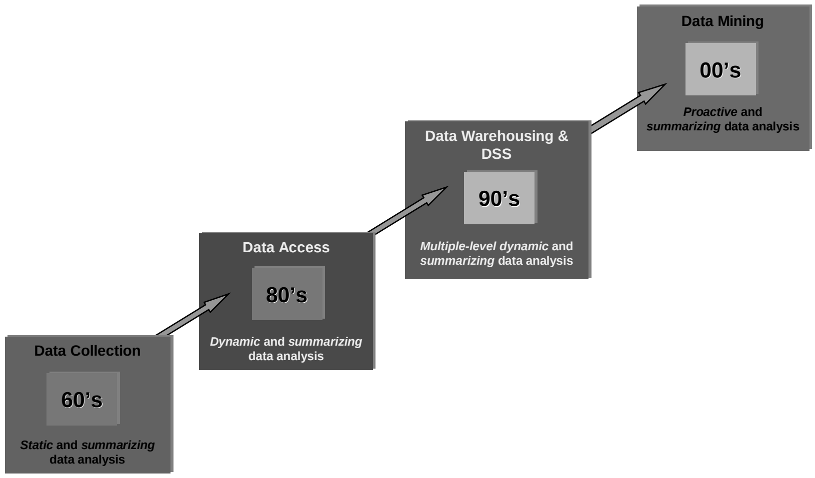
Data Mining: Confluence of Multiple Disciplines



Why data Mining?

- Data mining is a **computer-driven** application, as it is performed by computers and not by humans.
- Data mining solves the "**query formulation**" problem.
- Data mining confronts the visualisation and understanding of large data sets.

The evolution…



The KDD process in detail (1/3)

1. **Identify the goal of the KDD process:** Develop an understanding of the application domain and the relevant prior knowledge.
2. **Create a target data set:** Select a data set, or focus on a subset of variables or data samples, on which discovery will be performed.
3. **Clean and pre-process data:** Remove noise, handle missing data fields, account for time sequence information and known changes.

The KDD process in detail (2/3)

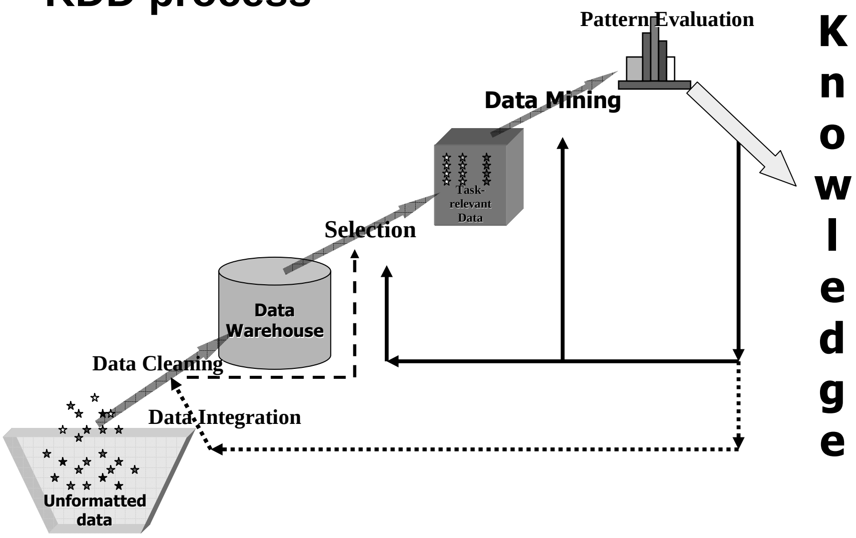
4. **Reduce and project data:**
   Find useful features to represent the data depending on the goal of the task.
5. **Identify data mining method:**
   Match the goals of the KDD process to a particular data mining method: e.g. summarization, classification, regression, clustering, etc.
6. **Choose a data mining algorithm:**
   Select method(s) to be used for searching for patterns in the data.
7. **Apply data mining**

The KDD process in detail (3/3)

8. **Evaluate data mining results:**
   Interpret mined patterns, possibly return to steps 1-7 for further iteration.
9. **Consolidate discovered knowledge:**
   Incorporate this knowledge into another system for further action, or simply document it and report it to interested parties.

Agent Intelligence through Data Mining

A graphical representation of the KDD process



18/09/2006 ECML/PKDD 2006, Berlin, Germany 22

Agent Intelligence through Data Mining

Data Mining: On What Kind of Data? (1/2)

- Relational databases
- Data warehouses
- Transactional databases
- Object-oriented and object-relational databases
- Spatial databases

18/09/2006 ECML/PKDD 2006, Berlin, Germany 23

Data Mining: On What Kind of Data? (2/2)

- Time-series data and temporal data
- Text databases and multimedia databases
- Heterogeneous and legacy databases
- Genomics databases
- World Wide Web

Data Mining Functionalities (1/6)

Concept description

- **Data characterization:** Summarize the features of the class under study (**target** class) in general terms.
  - ✓ E.g. Summarize the characteristics of customers who spent more than $1000 during 2003.
- **Data discrimination:** Compare the feature of the target class with one or a set of comparative classes (**contrasting** classes).
  - ✓ E.g. Create a comparative profile of customers that shop often to customers that shop rarely in our store
- Usual concept description outputs are:
  - ✓ **Characteristic rules**
  - ✓ **Discriminant rules**

Data Mining Functionalities (2/6)

Association (correlation and causality)

- The discovery of **association rules** showing attribute-value conditions that frequently occur
- Widely used for **market-basket** analysis and **transaction** data analysis

Example:

If the age of a customer is between 20 and 29 **and** his income is between $20.000 and $29.000, **then** he buys a PC with a certainty of 60%:

```
Age("20-29") ^ Income("$20..$29K") ⇒
              Buys("PC")
   [support = 2%, confidence = 60%]
```


Data Mining Functionalities (3/6)

Cluster analysis

- The identification of a finite set of **categories** or **clusters** to describe the data.
- Class label is **unknown**: The training data does not have any!!!
- **Clustering principle:** Objects that belong to the same cluster must be **similar** to each other, while objects that belong to different clusters must be **dissimilar** to each other.

Example

Identify homogeneous subpopulations of customers.

CERTH AUTH

Data Mining Functionalities (4/6)

Classification and Prediction

- The process of finding a set of **models** (or functions) that describe and distinguish data classes or concepts.
- The derived model is based on the analysis of **training** data, whose class label is **known.**
- The derived model can be presented as:
  - ✓ Classification rules (If-Then rules)
  - ✓ Decision tree (A flow-chart-like tree structure)
  - ✓ Mathematical formulae
  - ✓ Neural Networks

Examples

- ✓ Classify countries based on climate
- ✓ Classify cars based on gas mileage

CERTH AUTH

Data Mining Functionalities (5/6)

Outlier analysis

- Identify **outliers** in a set of data.
- **Outlier:** a data object that does not comply with the general behavior of the data
- It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis

Example

Discovery of fraudulent usage of credit cards.

Data Mining Functionalities (6/6)

Trend and evolution analysis

- Trend and deviation: regression analysis
- Sequential pattern mining, periodicity analysis
- Similarity-based analysis

Other pattern-directed or statistical analysis

Are all the “Discovered” Patterns Interesting?

A data mining system/query may generate thousands of patterns.
Not all of them are **interesting!!!**

Interestingness

A pattern is **interesting** if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm.

Are all the “Discovered” Patterns Interesting?

An interesting pattern represents **knowledge**

Different types of Interestingness

- **Objective**: based on statistics and the structure of discovered patterns
  - ✓ Support,
  - ✓ Confidence, etc.
- **Subjective**: based on user’s belief in the data
  - ✓ Unexpected,
  - ✓ Novel,
  - ✓ Actionable, etc.

Major Issues in Data Mining (1/3)

Mining methodology and user interaction

- Mining different kinds of knowledge in databases
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages and ad-hoc data mining
- Expression and visualization of data mining results
- Handling noise and incomplete data
- Pattern evaluation: the interestingness problem

CERTH AUTH

Major Issues in Data Mining (2/3)

Performance and scalability

- Efficiency and scalability of data mining algorithms
- Parallel, distributed and incremental mining methods

Issues relating to the diversity of data types

- Handling relational and complex types of data
- Mining information from heterogeneous databases and global information systems (WWW)

CERTH AUTH

Major Issues in Data Mining (3/3)

Issues related to applications and social impacts

- Application of discovered knowledge
  - ✓ Domain-specific data mining tools
  - ✓ Intelligent query answering
  - ✓ Process control and decision making
- Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
- Protection of data security, integrity, and privacy

Summary

- ✓ Mining can be performed on a variety of information repositories
- ✓ Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- ✓ Classification of data mining systems
- ✓ Major issues in data mining

Part 2-
Data Preprocessing

Major Tasks in Data Preprocessing

1. **Data cleaning**
2. **Data integration**
3. **Data transformation**
4. **Data reduction**
5. **Data discretization**

1. Data Cleaning

Real-world data tend to be "dirty" and incomplete…

Data cleaning tasks

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data

Missing Data

Data is not always available

- Many tuples have no recorded value for several attributes, i.e. customer income in sales data

Missing data may be due to:

- Equipment malfunction
- Inconsistency with other recorded data and thus deleted
- Data not entered due to misunderstanding
- Certain data may not be considered important at the time of entry
- History or changes of the data may not be registered

Missing data may need to be inferred.

Noisy Data

Noise: Random error or Variance in a measured variable.

Incorrect attribute values may be due to:

- Faulty data collection instruments
- Data entry problems
- Data transmission problems
- Technology limitation
- Inconsistency in naming convention

Other data problems which require data cleaning:

- Duplicate records
- Incomplete data
- Inconsistent data

2. Data Integration

Data integration: Integration of multiple databases, data cubes, or files. Data integration combines data from multiple sources into a coherent store.

Schema integration

✓ Integrate metadata from different sources

✓ Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-id#

Detecting and resolving data value conflicts

✓ For the same real world entity, attribute values from different sources are different

✓ Possible reasons: different representations, different scales, e.g., metric vs. British units

3. Data Transformation

- **Smoothing**: remove noise from data
- **Aggregation:** summarization, data cube construction
- **Generalization:** concept hierarchy climbing
- **Normalization:**
  scaled to fall within a small, specified range:
  - ✓min-max normalization
  - ✓z-score normalization
  - ✓normalization by decimal scaling
- **Attribute/feature construction:**
  New attributes constructed from the given ones

4. Data Reduction Strategies

- **Data reduction:** Obtains a reduced representation of the data set that is **much smaller** in volume, but yet produces the **same** (or almost the same) analytical results
- **Data reduction strategies**
  - ✓ Data cube aggregation
  - ✓ Dimensionality reduction
  - ✓ Numerosity reduction
  - ✓ Discretization and concept hierarchy generation

5. Data Discretization (1/2)

Part of data reduction but with particular importance

Discretization methods are applied both on <u>numeric</u> and <u>categorical</u> data

Discretization methods for categorical data

✓ Specification of a partial ordering of attributes explicitly at the schema level by users or experts

✓ Specification of a portion of a hierarchy by explicit data grouping

✓ Specification of a set of attributes, but not of their partial ordering

Mainstream DM techniques

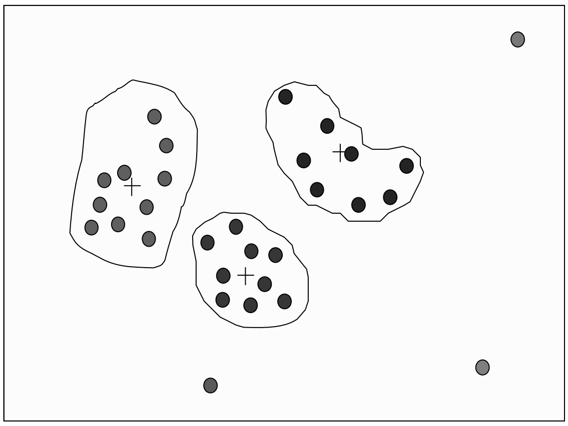
- **Clustering**
- **Classification**
- **Association Rules**

Part 3 – Clustering (Unsupervised Learning)

CERTH AUTH

Cluster Analysis

The identification of a finite set of <u>categories</u> or <u>clusters</u> to describe the data.



CERTH AUTH

Clustering: The Basics…

What is a Cluster?

Cluster is the collection of data objects that are:

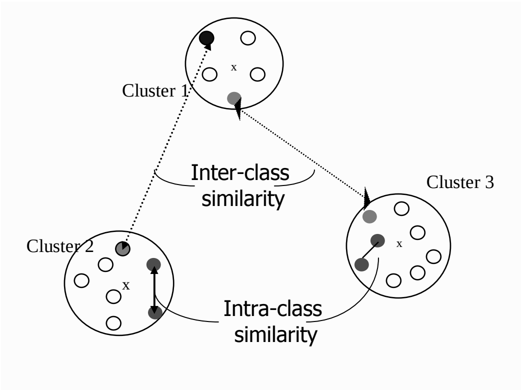
- ✓Similar to one another within the same cluster
- ✓Dissimilar to the objects in other clusters

Clustering as learning:

Clustering is unsupervised learning, which means that there are **<u>no predefined classes</u>** and **<u>no examples</u>** that would show what kind of desirable relations should be valid among the data.

Clustering criteria

- ✓ **High <u>intra-cluster</u> similarity**
- ✓ **Low <u>inter-cluster</u> similarity**



Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Interpretability and usability

CERTH AUTH

Steps to develop a clustering process

1 Feature selection

Select properly the features on which clustering is to be performed

2 Clustering algorithm

✓ *Proximity measure*.

✓ *Clustering criterion*

3 Validation of the results

The correctness of clustering algorithm results is verified using appropriate criteria and techniques

4 Interpretation of the results

CERTH AUTH

A Categorization of Clustering Methods

- **Technique used in order to define clusters**
  - ✓Partitional Methods
  - ✓Hierarchical Methods
  - ✓Density-Based Methods
  - ✓Grid-Based Methods
- **The type of variables**
  - ✓Statistical – Numerical Data
  - ✓Conceptual- Categorical Data
- **Theory used in order to extract clusters**
  - ✓Fuzzy Clustering
  - ✓Crisp Clustering
  - ✓Kohonen net clustering

Partitioning Algorithms: Basic Concept

- **Partitional method:** Decompose the data set into a set of k disjoint clusters.

Problem Definition

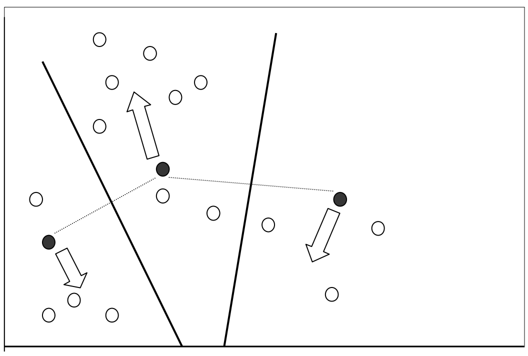
Given an integer k, find a partition of k clusters that optimizes the chosen partitioning criterion

$\Re$: Best partitioning algorithm representative:

The ***K-Means*** Method

The K-Means Clustering Method

$$E_K = \sum_k \left\| x_k - m_{c(x_k)} \right\|^2$$



CERTH AUTH

Comments on the K-Means Method

- **Advantages**
  - ✓*Relatively efficient*: O(tkn), where *n* is the number of objects, *k* is the number of clusters, and *t* is the number of iterations.
  - ✓Normally, *k, t* << n.
  - ✓Often terminates at a local optimum.
- **Problems**
  - ✓Applicable only to numerical data sets
  - ✓Need to specify the number of clusters in advance
  - ✓Unable to handle noisy data and outliers
  - ✓Not suitable to discover clusters with non-convex shapes

CERTH AUTH

The K-Medoids Clustering Method

A way of overcoming the K-Means Method problems is the K-Medoids method

Medoid: a representative object in clusters.

Most Known K-Medoids Algorithms:

- **PAM** (Kaufmann & Rousseeuw, 1987)
- **CLARA** (Kaufmann & Rousseeuw, 1990)
- **CLARANS** (Ng & Han, 1994): Randomized sampling

Hierarchical Clustering Algorithms

- **BIRCH (1996):** uses CF-tree and incrementally adjusts the quality of sub-clusters
- **CURE (1998):** is robust to outliers and identifies clusters of non-spherical shapes.
- **ROCK (1999):** is a robust clustering algorithm for Boolean and categorical data. It introduces two new concepts, that is a point's neighbours and links.

BIRCH - CF

- **What is the CF?** A triplet summarizing information about subclusters of objects. The CF of a subcluster is defined as:

$$CF = (N, LS, SS)$$

***N*:** Number of data points

LS: $\sum^{N}_{i=1}=X_i$ SS: $\sum^{N}_{i=1}=X_i^2$

- **Scales linearly:** Finds a good clustering with a single scan and improves the quality with a few additional scans.
- **Problem:** Handles only numeric data, and is sensitive to the order of the data record.

Agent Intelligence through Data Mining

CERTH AUTH

Density-Based Clustering Algorithms

- Clustering based on **density** (local cluster criterion), such as density-connected points.
- **Major features:**
  - ✓ Discover clusters of arbitrary shape
  - ✓ Handle noise
  - ✓ Need density parameters as termination condition
- **Representative algorithms:**
  - ✓ **DBSCAN:** Ester, et al. (KDD'96)
  - ✓ **DENCLUE:** Hinneburg & D. Keim (KDD'98)

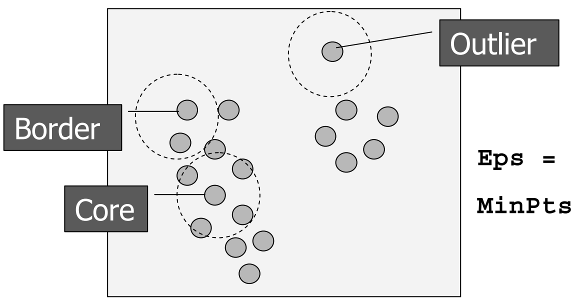
18/09/2006 ECML/PKDD 2006, Berlin, Germany 70

Agent Intelligence through Data Mining

CERTH AUTH

DBSCAN

- Relies on a **density-based** notion of cluster: A **cluster** is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise



`Eps = 1cm`

`MinPts = 5`

18/09/2006 ECML/PKDD 2006, Berlin, Germany 71

DBSCAN: The Algorithm

1. Arbitrary select a point *p*
2. Retrieve all points density-reachable from *p* w.r.t. **Eps** and **MinPts**.
3. If *p* is a **core** point, a cluster is formed.
4. If *p* is a **border** point, no points are density-reachable from *p* and DBSCAN visits the next point of the database.
5. Continue the process until all points are processed.

Part 4 - Classification (Supervised Learning)

CERTH AUTH

Classification

Classification can be described as a function that maps (classifies) a data item into ***one*** of the several predefined classes.

Requirements
✓ A well-defined set of classes
✓ A training set of pre-classified examples characterize the classification.

Goal

Induce a model that can be used to classify future data items whose classification is unknown.

CERTH AUTH

Classification Methods

- **Bayesian classification**
- **Decision Trees**
- **Neural Networks**

CERTH

AUTH

Bayesian classification

Aim: To classify a sample x to one of the given classes $c_1, c_2,.., c_N$ using a probability model defined according to Bayes theory

Requirements

✓ **A priori** probability for each class c_i.

✓ Conditional probability density function $p(x/c_i) \in [0,1]$

⇓ **Bayes Formula**

$$q(c_i / x) = \frac{p(x / c_i)p(c_i)}{\sum_{j=1}^{C} p(x / c_i)p(c_i)}$$ **Posterior probability**

A pattern is classified in the class with the highest posterior probability.

Problem: Complete knowledge of probability laws is necessary in order to perform the classification

CERTH

AUTH

Decision Trees

One of the most widely used techniques for classification and prediction.

Requirements

✓ Clusters (categories)

✓ A training set of pre-classified data.

Characteristics

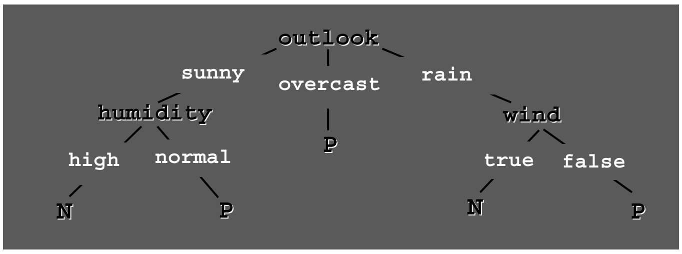
✓ **Internal node:** A test of an attribute

✓ **Branch descending of a node:** One of the possible values for this attribute

✓ **Leaf:** One of the defined classes

Decision Trees

A decision tree example:



The mechanism

- **Building phase:** The training data set is recursively partitioned until all the instances in a partition have the same class
- **Pruning phase:** Nodes are pruned to prevent over fitting and to obtain a tree with higher accuracy

Decision Tree Algorithms

Decision tree algorithms differ on the test criterion for partitioning a set of records

- **ID3, C4.5:** Information gain algorithms
- **CLS:** Examines the solution space of all possible decision trees to some fixed depth. It selects a test that minimizes the computational cost of classifying a record.
- **SLIQ, SPRINT:** select the attribute to test, based on the GINI index

ID3 : The Algorithm

The mechanism

Step 1:

- ✓ If all instances in C are positive, then create YES node and halt.
- ✓ If all instances in C are negative, create a NO node and halt.
- ✓ Otherwise select an attribute, A with values $v_1, \ldots, v_n$ and create a decision node.

Step 2:

Partition the training instances in C into subsets $C_1, C_2, \ldots, C_n$ according to the values of A.

Step 3:

Apply the algorithm recursively to each of the sets C_i.

ID3: Definitions (1/2)

$\Re$: How does ID3 decide which attribute is the best?

- A statistical property, called **information gain**, is used.
- The information needed to identify the class of an element of S, called **Entropy of S**, is:

$$Info(S) = \sum p(I) \log_2 p(I)$$

where $p(I)$ is the proportion of S belonging to class I.

ID3: Definitions (2/2)

- The information needed to identify the class of an element of S, **Info (S,A)**, after we partition S on basis of the value of an attribute A into sets Sv:

$$Info(S,A) = \sum[(|S_v|/|S|) \times Entropy(S_v)]$$

- **Gain(S, A)** is information gain of example set S on attribute A.

$$Gain(S, A) = Info(S) - Info(S,A)$$

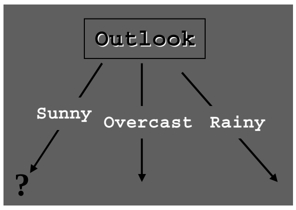
ID3: Example

Outlook	Temperature	Humidity	Wind	Play_ball
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

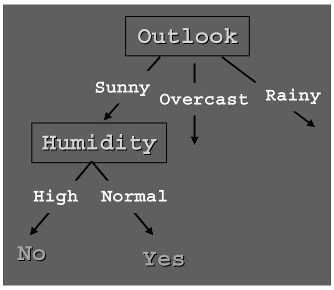
ID3 : Attribute Selection

Which attribute will be the root node ?

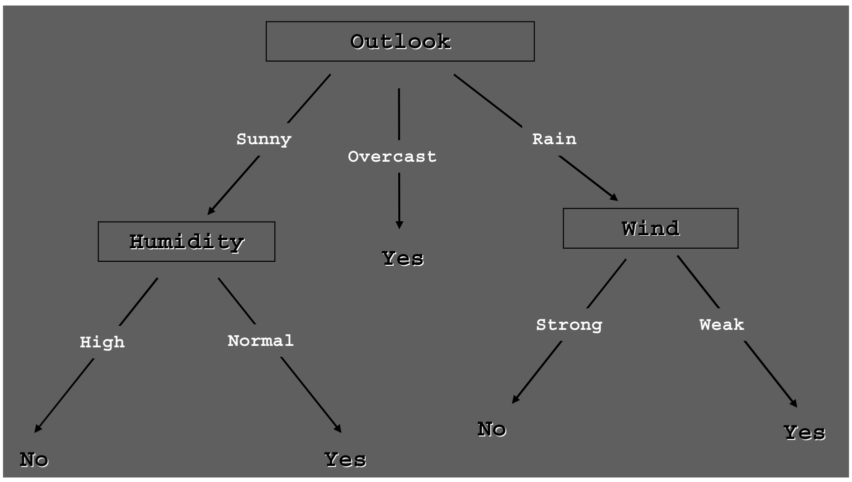
```
Gain(S, Outlook) = 0.246
Gain(S, Temperature) = 0.029
Gain(S, Humidity) = 0.151
Gain(S, Wind) = 0.048
```




```
Gain(Sunny, Humidity) = 0.970
Gain(Sunny, Temperature) = 0.570
Gain(Sunny, Wind) = 0.019
```



ID3 Example : Final Decision Tree



C4.5

C4.5 is an **extension of ID3** that accounts for:

- Unavailable values
- Continuous attribute values ranges
- Pruning of decision trees
- Rule derivation

PART 5 – Association Rules $S_1 \Rightarrow S_2$

Association mining: The idea…

The problem:

Given:

1) database of transactions,

2) each transaction is a list of items (purchased by a customer in a visit)

Find: All rules that correlate the presence of one set of items with that of another set of items

Association mining: The solution…

Association mining

Finding **frequent** patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.

The basic concepts on rule mining…

Support: There is enough support for the rule $\mathbf{S_1}$=>$\mathbf{S_2}$ if the number of records whose attributes include $\mathbf{S_1}$ **or** $\mathbf{S_2}$ is at least **minsupp**.

Confidence: We have enough confidence in an association rule if the ratio of records having attributes that include $\mathbf{S_1}$ **or** $\mathbf{S_2}$ over records having attributes that include S_1 is at least **minconf**.

Association Rule: a definition

An **association rule** is an expression of the form S_1=>S_2 where S_1, S_2 are sets of attributes with sufficient support and confidence.

When to pick an association rule?

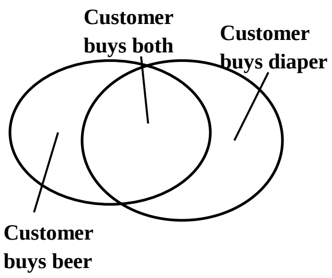
An association rule **A=>B** is thought to be **interesting** when:

$$\frac{P(A \cap B)}{P(A)} - P(B) > d$$

where **d** is an appropriate value

Counting Confidence and Support…

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F



What is the support and confidence for the rules:
1) $A \Rightarrow C$ **2)** $C \Rightarrow A$

1) A and C appear in 2 of the 4 transactions, so support is *50%*. In addition, A appears in 3 transactions, two of which contain C too, so confidence is *2/3=66,6%*.
2) A and C appear in 2 of the 4 transactions, so support is *50%*. In addition, A appears in all the transactions that C appears, so confidence is *2/2=100%*.

Mining Frequent Itemsets: the Key Step

Frequent itemsets: Sets of items that have minimum support

- A subset of a frequent itemset must also be a frequent itemset
  - ✓ i.e., if $\{A\ B\}$ is a frequent itemset, both $\{A\}$ and $\{B\}$ should be frequent itemsets
- Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)
- Use the frequent itemsets to generate association rules.

Association rule algorithms

- **Apriori**
- **DHP**
- **Trie Data Structure**
- **Iceberg Queries**

The Apriori Algorithm

Procedure Synopsis:

- Constructs a set of large items
- Counts the number of each set's appearances
- Determines large itemsets on a predetermined <u>minsupp</u>

The major Steps…

1) **Join Step:** C_k is generated by joining L_{k-1} with itself
2) **Prune Step:** Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

Candidate generation-Pseudo Code

In order to estimate C_k, Apriori uses $L_{k-1} \times L_{k-1}$. Then C_k consists of all (k-1)-itemsets.

C_k: Candidate itemset of size k

L_k : frequent itemset of size k

L_1 : {frequent items};

Pseudo-code:

```
for (k = 1; Lk !=∅; k++) do begin
 Ck+1 = candidates generated from Lk;
   for each transaction t in database do
     increment the count of all candidates in Ck+1 that are
contained in t
     Lk+1  = candidates in Ck+1 with min_support
   end
return ∪k Lk;
```

Agent Intelligence through Data Mining

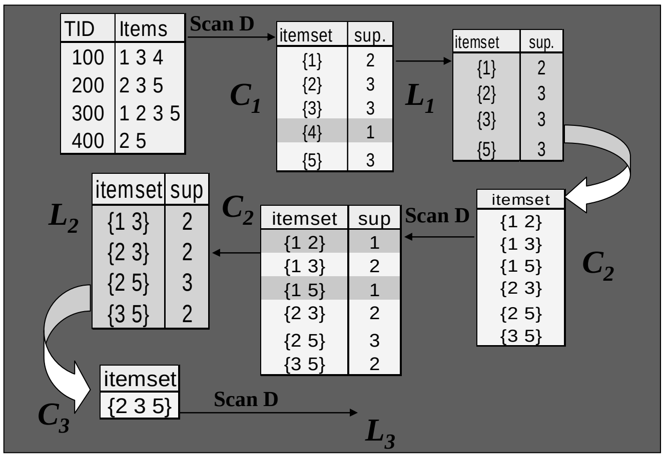
Candidate Generation Example…

1. L_3={abc, abd, acd, ace, bcd}
2. Self-joining: $L_3 \times L_3$
   ✓ ***abcd*** from ***abc*** and ***abd***
   ✓ ***acde*** from ***acd*** and ***ace***
3. Pruning:
   ✓ ***acde*** is removed because ***ade*** is not in L_3
4. C_4={abcd}

18/09/2006 ECML/PKDD 2006, Berlin, Germany 98

Agent Intelligence through Data Mining

Association Rule Extraction An Example…



18/09/2006 ECML/PKDD 2006, Berlin, Germany 99

Methods to Improve Apriori's Efficiency

- **Hash-based itemset counting:** A k-itemset whose corresponding hashing bucket count is below the threshold cannot be frequent
- **Transaction reduction:** A transaction that does not contain any frequent k-itemset is useless in subsequent scans
- **Partitioning:** Any itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB
- **Sampling:** Mine on a subset of given data, lower support threshold **+** a method to determine the completeness
- **Dynamic itemset counting:** Add new candidate itemsets only when all of their subsets are estimated to be frequent

Integration Algorithms (1/2)

DHP

Creates a hashing table that controls the legitimacy of k-itemsets and reduces dramatically the size of C_2 and the itemset database in result.

FTDA

It is used in order to extract association rules on quantitative values. Fuzzy methods are applied and fuzzy rules are extracted.

Integration Algorithms (2/2)

Partition

Divides the itemset database and applies only two scans in order to execute the algorithm.

DIC

Optimization through dynamic itemset counting.

TRIE Data Structure

Used to create covers: itemsets that have greater than or equal to a specified minimum support threshold.

Extracts exclusive association rules.

Exclusive Association Rules

Definition:

We say that, $S_1 \cup \neg S_2 \Rightarrow S_3$ is an excluding association, if S_1, S_2 and S_3 are mutually disjoint and the following conditions hold:

1. $\frac{supp(S_1 \cup S_3)}{supp(S_1)} < minconf$
2. $supp(S_1 \cup S_3) - supp(S_1 \cup S_2 \cup S_3) \geq minsupp$
3. $\frac{supp(S_1 \cup S_3) - supp(S_1 \cup S_2 \cup S_3)}{supp(S_1) - supp(S_1 \cup S_2)} \geq minconf$

Part 7 - Data Mining and Semantics

The different perspectives on Semantics

- The DM community refers to **semantics** as a way of representing knowledge by the use of “some” unified language
- The AI community deals with **semantics** as a way of creating meta-knowledge through the refinement of decision structures

DM Perspective – The Reasons…

- Each DM tool uses a different way of describing knowledge (same metrics, same concepts)
- Not all tools provide all dm techniques
- In order to incorporate results in your system, **reusability** is now mandatory

The Solution…

- **The development of a modeling language, that would allow users to:**
  - ✓Develop models within one vendor's application
  - ✓Use other vendors' applications to test, visualize, analyze, evaluate or otherwise use the models

The Language…

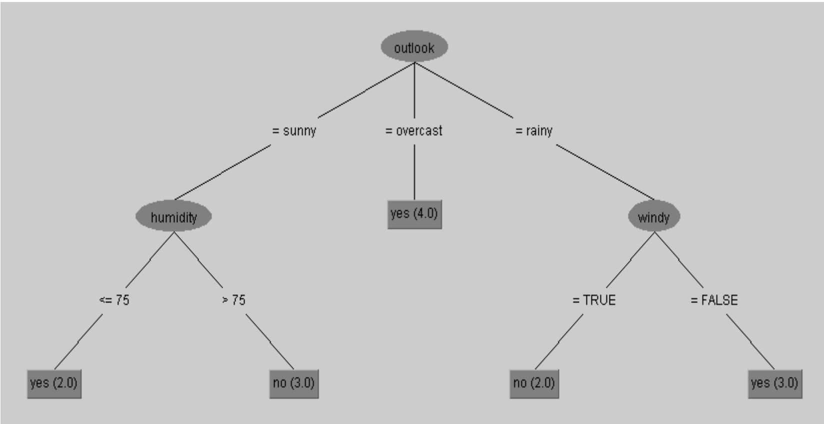
PMML

Predictive Markup Model Language

http://www.dmg.org

- ✓ Developed by the DMG (Oracle, SAS, SPSS, MineIt, etc.)
- ✓ **XML-based** language, with its own DTD, mistakes are difficult to occur
- ✓ It has a stable version (2.0)

A PMML document example



And the resulting PMML document is :

AI perspective –The Reasons…

- In order to incorporate meta-knowledge into data mining, new interestingness measures are applied, in order to indicate the **validity** of the knowledge extracted, according to domain understanding.
  - ✓ Fuzzy weights
  - ✓ Penalty matrices

Part 8 - New data mining algorithms: the AI way…

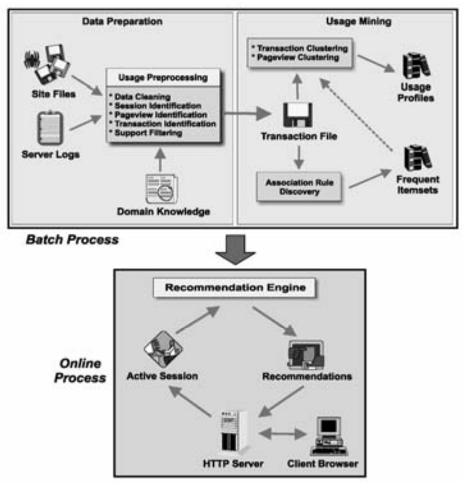
CERTH AUTH

The new agent-oriented algorithm…

- The need to develop an algorithm that deals with agent actions has led us to the development of **κ-Profile**.
- κ-Profile is a data mining mechanism that was first introduced for the **dynamic segregation of web roaming attitudes**.
- It is now being adapted in order to be able to **predict agent behaviors and actions**.

CERTH AUTH

The web roaming mechanism…



Offline

- ✓ Data extraction
- ✓ Data preprocessing
- ✓ Transaction identification
- ✓ Pageview identification
- ✓ Clustering
- ✓ Profiling

Online

- ✓ Active session
- ✓ Recommendation

Techniques and Algorithms used

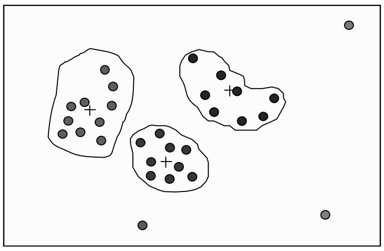
- **Weighted vectors**
  - ✓ FIS (**F**uzzy **I**nference **S**ystem)
- **Clustering**
  - ✓ Simple k-means
  - ✓ Maximin
- **Profiling**
  - ✓ The k-Profiler
- **Recommendation**
- **Evaluation**

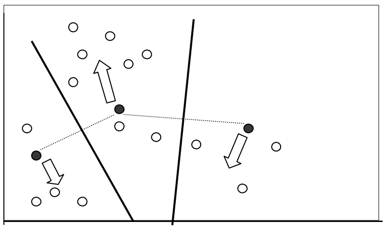
Data Preprocessing

- **Transaction vectors**
  - ✓ Outliers
  - ✓ Robots, spiders, etc.
- **Pageview vectors**
  - ✓ Homepage
  - ✓ Banners, etc.
  - ✓ Very frequent pages
  - ✓ Very rare pages
- **Fuzziness**

Clustering and Profiling

- **Maximin**
  - ✓ Decides on the number of clusters
  - ✓ Finds primary cluster centers
- **Simple k-means**
  - ✓ Creates transaction clusters
- **K-Profiler**
  - ✓ Creates the transaction profiles





κ-Profiler – From theory...

$T = \{t_1, t_2, \ldots, t_m\}$ – Transaction Set

$P = \{p_1, p_2, \ldots, p_n\}$ – Pageview Set

$t = \langle w(p_1,t), \ldots, w(p_n,t)\rangle$ – Transaction Vector

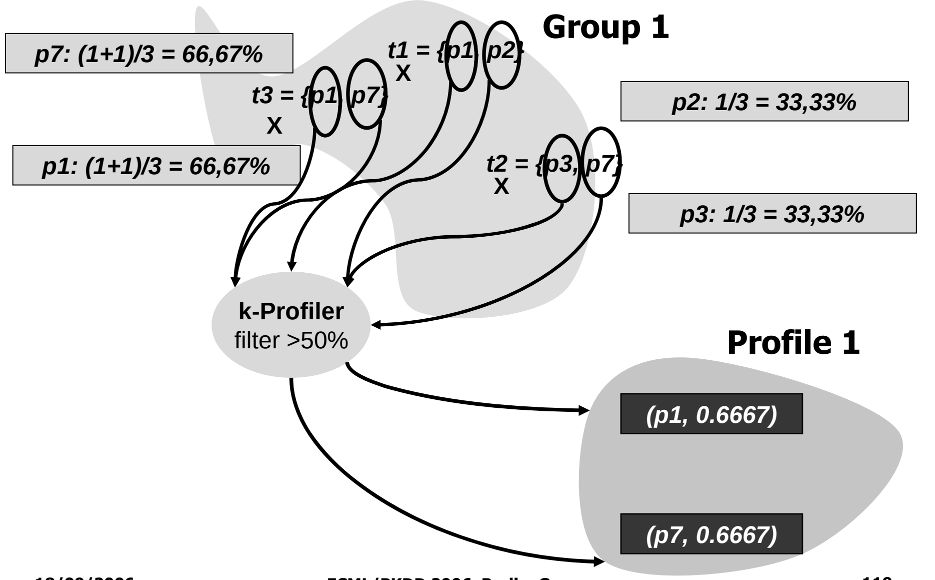
$TC = \{c_1, c_2, \ldots, c_k\}$ – Transaction Groups

$pr_c = \{\langle p, weight(p, pr_c)\rangle \mid p \in P, weight(p, pr_c) \geq \mu\}$ – User Profiles

- $weight(p, pr_c) =$

CERTH Agent Intelligence through Data Mining AUTH

... to practice



18/09/2006 ECML/PKDD 2006, Berlin, Germany 119

CERTH Agent Intelligence through Data Mining AUTH

The Recommendation engine

User profile $C = \{w\ \ , w\ \ , \ldots, w\ \ \}$

where $w =$

$match(S, C) =$

$Rec(S, p) =$

$UREC(S) = \{w \mid C \in UP, \text{ and } Rec(S, w) \geq \rho\}$

18/09/2006 ECML/PKDD 2006, Berlin, Germany 120

Work on agents

- Modeling agent actions according to the 'κ-Profile' paradigm
- Implementation of the agent behavior that shall incorporate the recommendation engine into the agents
- Testing and dissemination…

Part 9 - Data Mining Applications and Trends

CERTH AUTH

Basic Data Mining application domains

- ✓ Web mining
- ✓ Biomedical and DNA data analysis
- ✓ Financial data analysis
- ✓ Retail industry
- ✓ Telecommunication industry
- ✓ Market analysis and management
- ✓ Risk analysis and management
- ✓ Fraud detection and management

CERTH AUTH

Web mining - Data on the Web

- **Primary data (Web content):**
  - ✓ Mainly **text**, with some multimedia content and mark-up commands.
  - ✓ Underlying databases (not directly accessible).
- **Secondary data (Web usage):**
  - ✓ Access **logs** collected by a Web server.
  - ✓ A variety of navigational information collected by Web clients.

Approaches to Web mining

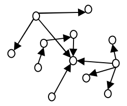
- **Web content mining:**
  - ✓ Pattern discovery in Web content data.
  - ✓ Mainly mining **unstructured** textual data.
- **Web structure mining**
  - ✓ Pattern discovery in the **Web graph**.
  - ✓ The graph is defined by the hyperlinks.
- **Web usage mining**
  - ✓ Discovery of interesting **usage patterns**.
  - ✓ Mainly in server logs.

Web content mining

- **Information Access**
  - ✓ Document category modeling
  - ✓ Construction of thematic hierarchies
- **Fact Extraction**
  - ✓ Extraction of product information, presented in different formats
- **Information Extraction**
  - ✓ Structured "event" summaries from large textual corpora

AUTH

Web structure mining (1/2)



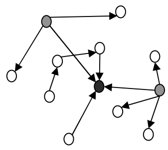
- Information retrieval can be improved by:
  - ✓ Identifying **authoritative** pages.
  - ✓ Identifying **resource index** pages.
  - ✓ Summarizing common references.
- Linked pages often contain complementary information (e.g. product offers).
- Structural analysis of a Web site facilitates its improvement.

Agent Intelligence through Data Mining

AUTH

Web structure mining (2/2)

- **Social network analysis**:
  - ✓ Nodes with large fan-in (authorities) provide high quality information.
  - ✓ Nodes with large fan-out (hubs) are good starting points.



- Disconnected subgraphs correspond to different social (e.g. research) communities.


Web Usage Mining (1/2)


Personalization:

- Better service for the user:
  - ✓ Reduction of the information overload.
  - ✓ More accurate information retrieval and extraction.
- Customer relationship management:
  - ✓ Customer segmentation and targeted advertisement.
  - ✓ Customer attraction and retention strategy.
  - ✓ Service improvement (site structure and content).

Web Usage Mining (2/2)

Data collection — Collection of usage data by the server and the client.

↓

Data pre-processing — Data cleaning, user identification, session identification

↓

Pattern discovery — Construction of user models

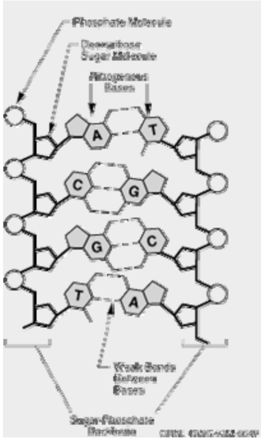
↓

Knowledge post-processing — Report generation, visualization, personalization module.

↓

Biomedical Data Mining and DNA Analysis (1/2)

- **DNA sequences:** 4 basic building blocks (nucleotides):
  - ✓ adenine (A)
  - ✓ cytosine (C)
  - ✓ guanine (G)
  - ✓ thymine (T)
- **Gene:** a sequence of hundreds of individual nucleotides arranged in a particular order



Biomedical Data Mining and DNA Analysis (2/2)

- Tremendous number of ways that the nucleotides can be ordered and sequenced to form distinct genes
- Semantic integration of heterogeneous, distributed genome databases
  - ✓ **Current:** highly distributed, uncontrolled generation and use of a wide variety of DNA data
  - ✓ Data cleaning and data integration methods developed in data mining will help

Data Mining for Financial Data Analysis (1/2)

- Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality
- Design and construction of **data warehouses** for multidimensional data analysis and **data mining**
  - ✓ View the debt and revenue changes by month, by region, by sector, and by other factors
  - ✓ Access **statistical** information such as max, min, total, average, trend, etc.
- **Loan payment prediction/consumer credit policy analysis**
  - ✓ feature selection and attribute relevance ranking
  - ✓ Loan payment performance
  - ✓ Consumer credit rating

Data Mining for Financial Data Analysis (2/2)

- **Classification and clustering of customers for targeted marketing**
  - ✓ Multi-dimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group
- **Detection of money laundering and other financial crimes**
  - ✓ Integration of data from multiple DBs (e.g., bank transactions, federal/state crime history DBs)
  - ✓ **Tools:** Data visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)

Data Mining for Retail Industry

- **Retail industry:** Huge amounts of data on sales, customer shopping history, etc.
- **Applications of retail data mining:**
  - ✓ Identify customer buying behaviors
  - ✓ Discover customer shopping patterns and trends
  - ✓ Improve the quality of customer service
  - ✓ Achieve better customer retention and satisfaction
  - ✓ Enhance goods consumption ratios
  - ✓ Design more effective goods transportation and distribution policies

Data Mining for Telecommunications

- A rapidly expanding and highly competitive industry and a great demand for data mining
  - ✓ Understand the business involved
  - ✓ Identify telecommunication patterns
  - ✓ Catch fraudulent activities
  - ✓ Make better use of resources
  - ✓ Improve the quality of service
- Multidimensional analysis of telecommunication data
  - ✓ Intrinsically multidimensional: calling-time, duration, location of caller, location of callee, type of call, etc.

Data Mining Tools - Criteria (1/3)

- **Data types:**
  - ✓ Relational
  - ✓ Transactional
  - ✓ Text
  - ✓ Time sequence
  - ✓ Spatial, etc.
- **Data sources:**
  - ✓ ASCII text files, multiple relational data sources
  - ✓ support ODBC connections (OLE DB, JDBC)?
- **Need for multiple dimensional views in selection**

Data Mining Tools - Criteria (2/3)

- **Data mining functions and methodologies:**
  - ✓ One vs. multiple data mining functions
  - ✓ One vs. variety of methods per function
- **Coupling with DB and/or data warehouse systems:**
  - ✓ Four forms of coupling: no coupling, loose coupling, semi-tight coupling, and tight coupling.
- **Scalability**
  - ✓ Row (or database size) scalability
  - ✓ Column (or dimension) scalability
  - ✓ Course of dimensionality: it is much more challenging to make a system column scalable that row scalable

Data Mining Tools - Criteria (3/3)

- **Visualization tools**
  - ✓ "A picture is worth a thousand words"
  - ✓ Visualization categories: data visualization, mining result visualization, mining process visualization, and visual data mining
- **System issues**
  - ✓ Running on only one or on several operating systems?
  - ✓ Client/server architecture?
  - ✓ Provide Web-based interfaces and allow XML data as input and/or output?

The Generations of DM systems

- **First generation**: Systems that performed classification or clustering and were based on a certain algorithm
- **Second generation**: Systems that provide better support throughout the whole KDD process
- **Third generation**: Systems that deal with the business end user rather than advanced data analysis

CERTH AUTH

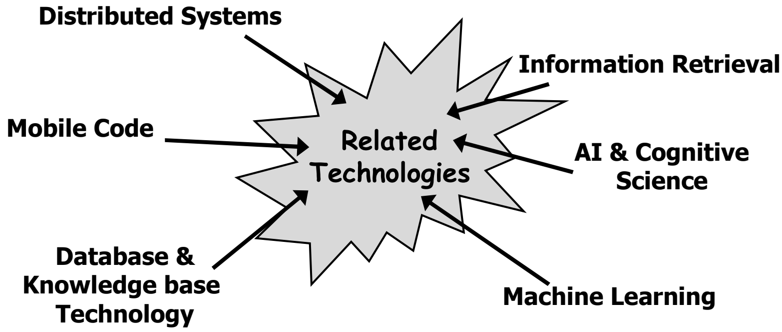
Part 10 – Intelligent Agent Technology

CERTH AUTH

Positioning agents in the software development process

- Agent technology is the **next step in object-oriented programming.**
- It satisfies all the requirements, while it supports major key properties, since agents are:
  - ✓ **autonomous,**
  - ✓ **goal-oriented,**
  - ✓ **cooperative,**
  - ✓ **communicative,**
  - ✓ **adaptive...**

Agents: A system-building paradigm



$$\frac{\text{agents}}{2004} = \frac{\text{objects}}{1982} = \frac{\text{structured programming}}{1974}$$

Intelligent Agents: An Intro…

- **Intelligent Agents (IAs)** are considered to be very important because they promise to change the way that people interact with computers.
- The underpinning concepts of IAs can be traced back to the early years of Artificial Intelligence - 40 years ago
- Research on IAs is considered to be in its first stages

CERTH AUTH

Definitions of Agents

- There is not a consensus definition of IAs!!!
- There are several operational definitions:
  - ✓ Dictionary Definition
  - ✓ Common Definition Patterns
  - ✓ Definition of IAs based on their characteristics

CERTH AUTH

Dictionary Definition

"An agent is one who acts for or in the place of another by authority from him… "

CERTH AUTH

Common Definition Patterns

An agent is a software entity that…

- Acts **autonomously** in a goal-oriented manner
- Functions **continuously**
- Acts on behalf of another entity (human, IA)
- Is able to **perceive its environment** through sensors and act on it through effectors
- Employs some degree of **knowledge**

CERTH AUTH

Attributes of an Agent (1/2)

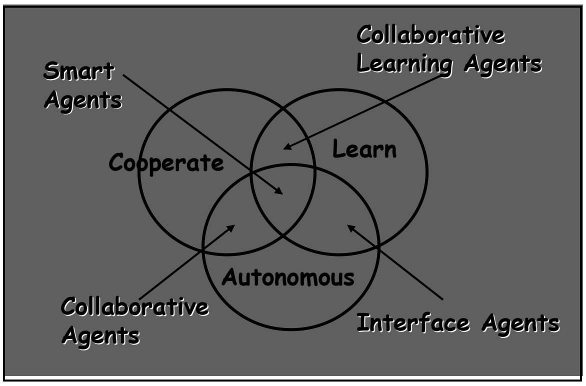
- **Autonomy**
- **Interactivity**
  - ✓ **Reactivity**
  - ✓ **Pro-activity**
- **Adaptivity**
- **Sociability**
- **Collaborative Behavior**

Attributes of an Agent (2/2)

- Competitiveness
- Temporal Continuity
- Personality
- Mobility
- Learning

Taxonomies of Agents (1/2)

Nwana's 3 dimensional classification:



Taxonomies of Agents (2/2)

We can also classify agents by:

- **What they do** (Web Search, Information Filtering, Notification, Financial Services, Entertainment, etc.)
- **Where they act** (Desktop, Internet, Intranet)
- **The degree of the characteristics they enjoy** (Mobile, Collaborative, Reactive, etc.)
- **A Combination of the above** (Hybrid Agents)

Underlying Technologies

- Technologies for Developing Agents
  - ✓ Internal Infrastructure
  - ✓ Programming Languages
  - ✓ Standards for distributed computing
- Agents Communication

Internal Infrastructures

- Knowledge-based expert systems
- Agents and Objects
- AI technologies (Neural Networks, Genetic Algorithms, Fuzzy Systems)

Programming Languages

✓ Telescript
✓ Smalltalk
✓ JAVA
✓ C++
✓ Other languages and development frameworks

Standards for Distributed Systems

- **CORBA** (Common Object Request Broker Architecture)
- **DCOM** (Distributed Component Object Model)
- **JAVA / RMI** (Remote Method Invocation)

Agent Communication

- **KQML** (Knowledge and Querying Manipulation Language)
- **KIF** (Knowledge Interchange Format)
- **FIPA** (Agent Communication Language)

Some Important Issues of Agent Design

- **Multi-agent** architecture vs. **Single-agent** architecture
- **Mobility**
- **Security**

Why MAS instead of a Single Agent? (1/2)

- A single Agent that handles a great amount of tasks lacks **performance**, **reliability**, **maintainability**, etc.
- MAS can provide **modularity**, **flexibility**, **modifiability**, **extensibility**, due to Distributed Environments

Why MAS instead of a Single Agent? (2/2)

- A single Agent cannot obtain extensive and specialized knowledge
- A MAS can access more knowledge resources
- Applications requiring Distributed Computing are better supported by MAS
- Intelligence, in neuroscience terms, can be approached by a multi-processing system such as a wide distributed environment of MAS

Mobility

- **Mobility:** The transportation of an Agent to remote services resources
- **Rationale for mobility:** Improved performance
- **New requirements are introduced:**
  - ✓ Presence of an agent server
  - ✓ Security
  - ✓ Common standards for IAs Communication
  - ✓ Complexities about system maintenance and IAs identification in a distributed environment

Security

- ✓ Unauthorized disclosure
- ✓ Unauthorized alteration
- ✓ Unauthorized damage
- ✓ Unauthorized copy and replay
- ✓ Denial of Service
- ✓ Repudiation
- ✓ Spoofing and Masquerading

Current Status of Agent Applications (2/2)

- ✓ Electronic Commerce
- ✓ Information Retrieval and Knowledge Management
- ✓ Mobile Computing
- ✓ Planning and Scheduling
- ✓ Scientific Applications

Current Status of Agent Applications

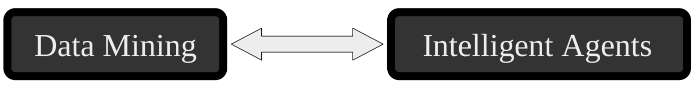
- ✓ **Distributed Project Management**
- ✓ **Manufacturing**
- ✓ **Networking**
- ✓ **Other applications** (economics, business, military, etc.)
- ✓ **Tools for Agent Development**

Standards and Organizations

- ✓ **AgentCities**
- ✓ **AgentLink**
- ✓ **FIPA**
- ✓ **Knowledge Sharing Effort**
- ✓ **OMG**
- ✓ **Other Organizations** (Agentx Working Group, International Foundation for MAS, Ontology.Org)

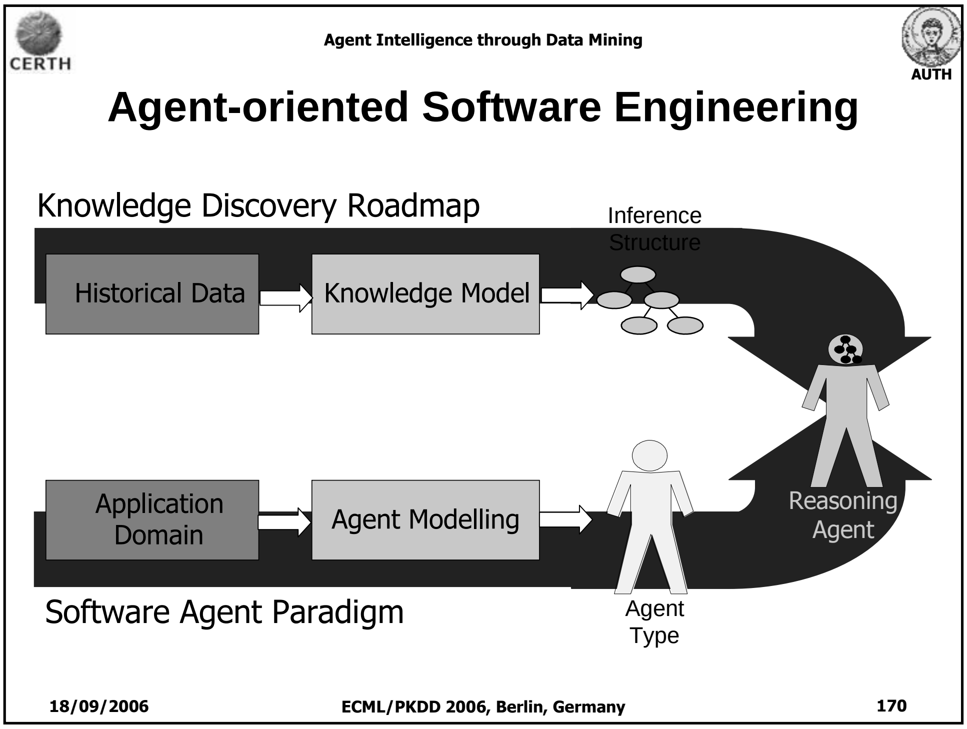
Part 11 – Agent Intelligence & Data Mining

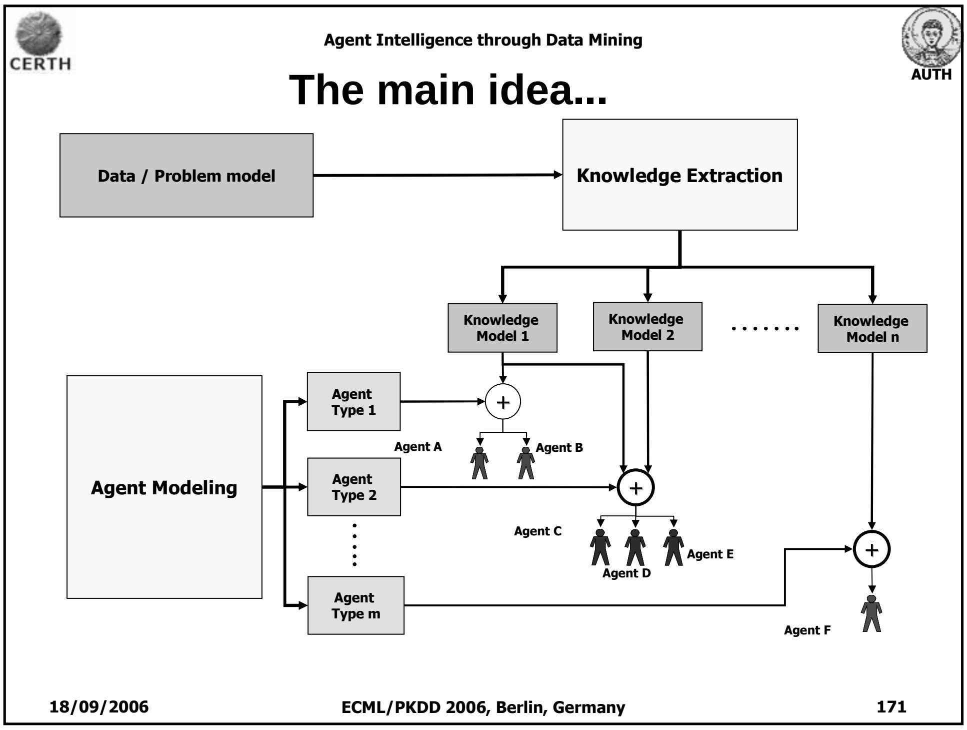
The main problem…



Do they need each other?

Symbiosis?


Agent Intelligence through Data Mining
CERTH
AUTH
Agent-oriented Software Engineering
Knowledge Discovery Roadmap
Inference Structure
Historical Data
Knowledge Model
Application Domain
Agent Modelling
Reasoning Agent
Software Agent Paradigm
Agent Type
18/09/2006
ECML/PKDD 2006, Berlin, Germany
170


Agent Intelligence through Data Mining
CERTH
AUTH
The main idea...
Data / Problem model
Knowledge Extraction
Knowledge Model 1
Knowledge Model 2
....
Knowledge Model n
Agent Modeling
Agent Type 1
Agent Type 2
Agent Type m
Agent A
Agent B
Agent C
Agent D
Agent E
Agent F
18/09/2006
ECML/PKDD 2006, Berlin, Germany
171

The main goal

- The development of a **unified methodology** that:
  - ✓ Takes logic limitations into account
  - ✓ Is supported by the appropriate tools
  - ✓ Has been applied to a satisfactory number of applications, in order to:

Provide the capability of **dynamically incorporating knowledge** to SAs and MAS. This knowledge has been extracted with the use of DM techniques

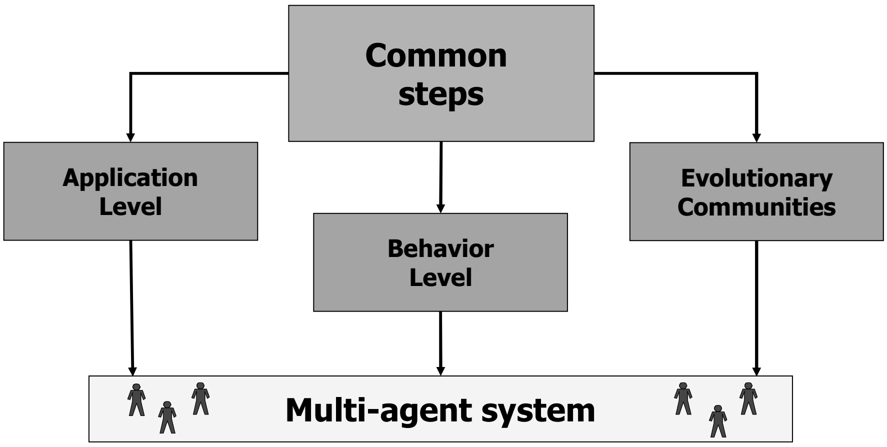
Knowledge diffusion levels

- Extracting knowledge on a MAS **application** level:
  - ✓ Perform data mining techniques on application data, in order to discover useful rules – associations – patterns.
- Extracting knowledge on a MAS **behavioral** level:
  - ✓ Perform data mining techniques on agent behavior data, in order to predict their behaviors and, thus, reduce system work load. The extracted knowledge is related to agent actions.
- Extracting knowledge on **evolutionary agent communities:**
  - ✓ Deploy evolutionary DM techniques, in order to study societal issues. It has to do with the satisfaction of the goal of a community, which evolves and learns through interaction.

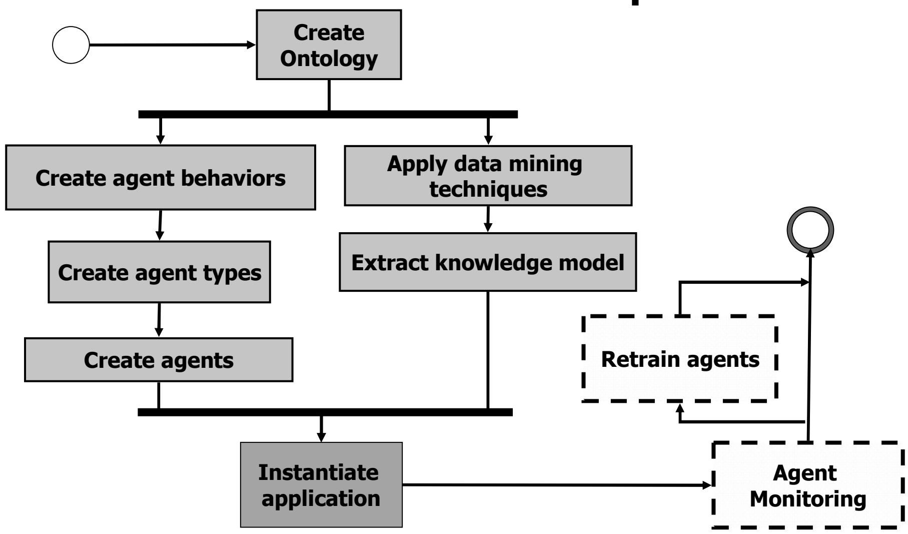
Defining training

Training: The process of **dynamically incorporating** DM-extracted knowledge models to SAs and MAS.

The Methodology



The Common steps



Application-oriented DM…

Application-oriented DM…

- Data mining (clustering, classification…) is performed on application-specific data
- For example, develop a MAS that decides whether a golf game will be played, depending on humidity, outlook, wind, etc.
- **The decision is extracted through data mining**

Prerequisites…

- A significant data volume must be available (the bigger the better)
- The DM expert has to decide on the best way to exploit the resulting knowledge
- The multi-agent system architecture has to seriously take into consideration specific limitations, such as the **safety** and **soundness** of the application.

Advantages…

- The size of knowledge bases of intelligent agents can dramatically affect their **performance**.
- Agent **retraining** on new data is feasible
- The retraining process can be **automated** and **dynamic**, providing versatility to MAS implementation.

Tools and techniques…

An integrated tool for embedding data mining extracted intelligence into agents is Agent Academy (AA):

http://sourceforge.net/projects/agentacademy

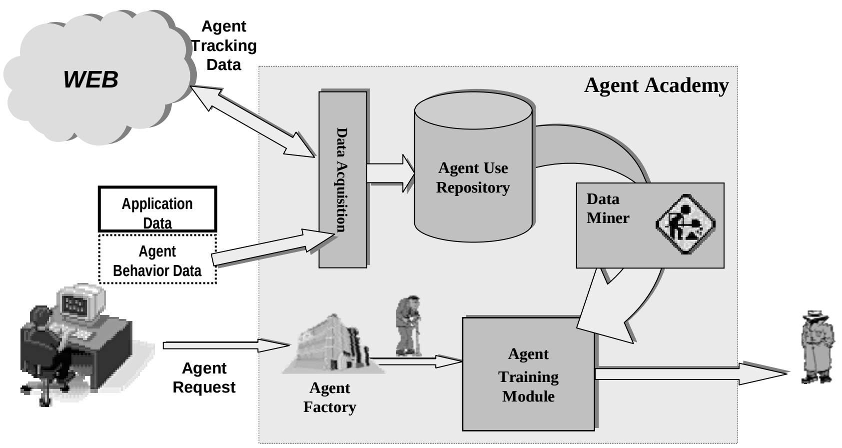
The AA hypothesis…

Using Agent Academy and its data-mining capabilities, the MAS developer can reduce the effort required to:

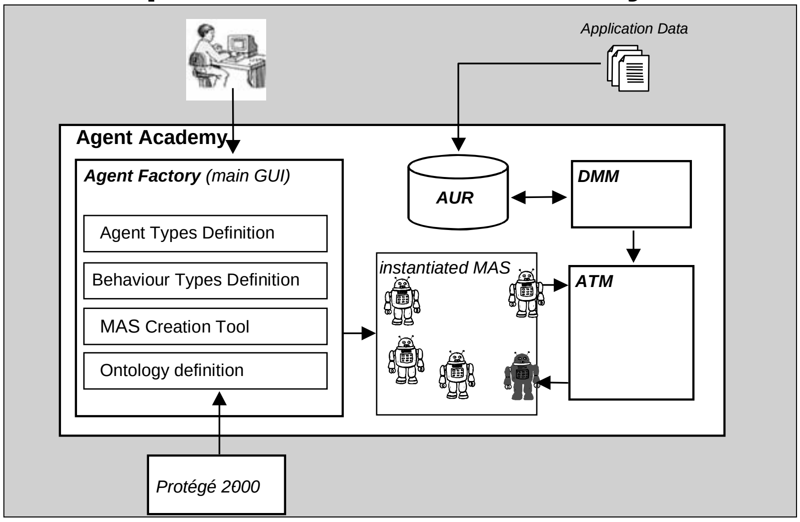
a. Develop **new** or expand **existing** applications with intelligent agents, and

b. Upgrade them, as needed, by **retraining** the agents deployed

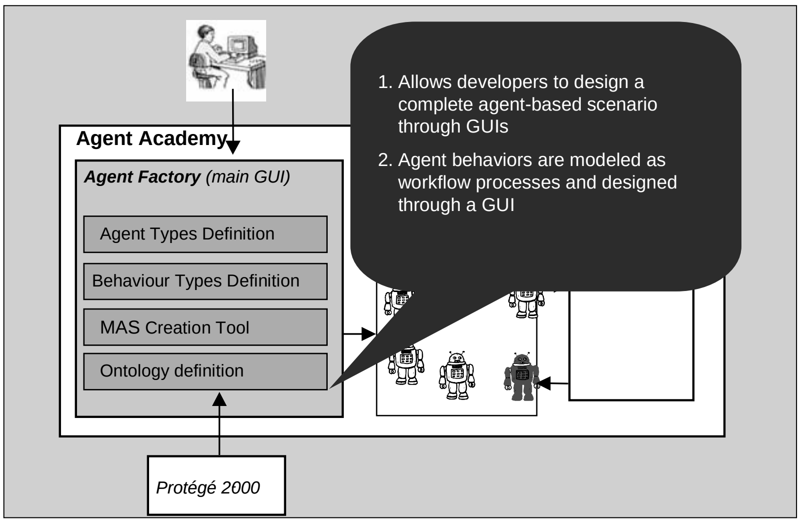
The AA-1.0 architecture…



Components of the AA-1.0 System

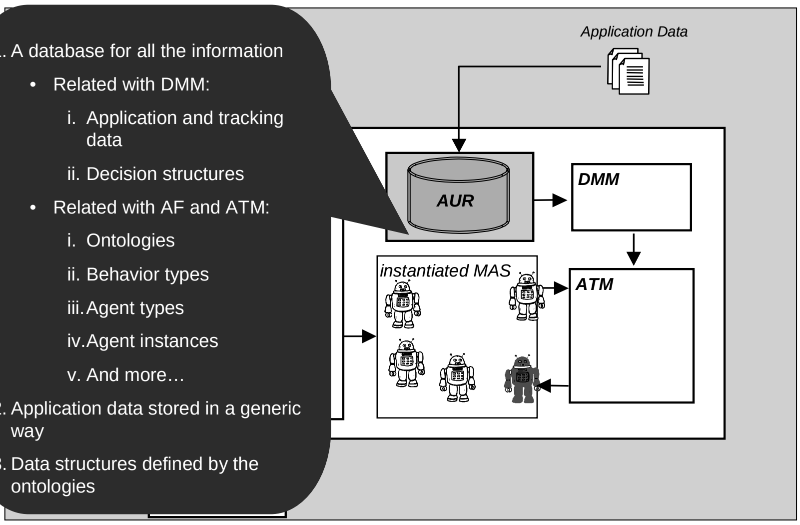


Agent Factory



Agent Use Repository

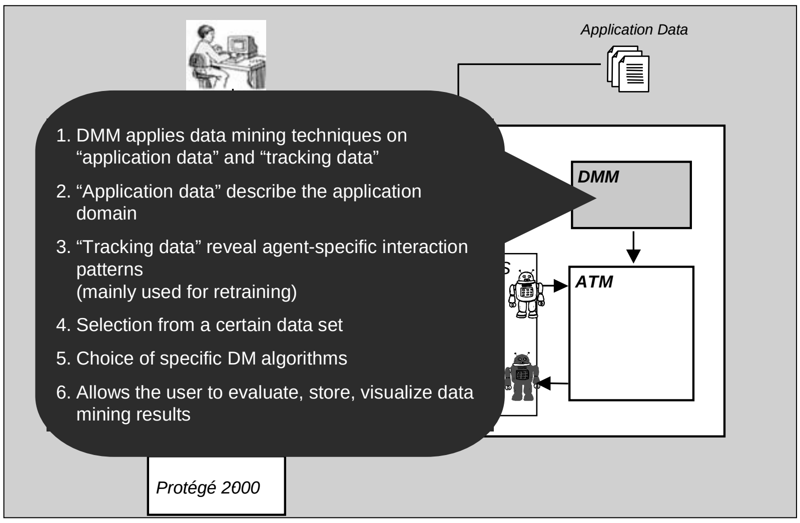
1. A database for all the information
   - Related with DMM:
     i. Application and tracking data
     ii. Decision structures
   - Related with AF and ATM:
     i. Ontologies
     ii. Behavior types
     iii. Agent types
     iv. Agent instances
     v. And more…
2. Application data stored in a generic way
3. Data structures defined by the ontologies



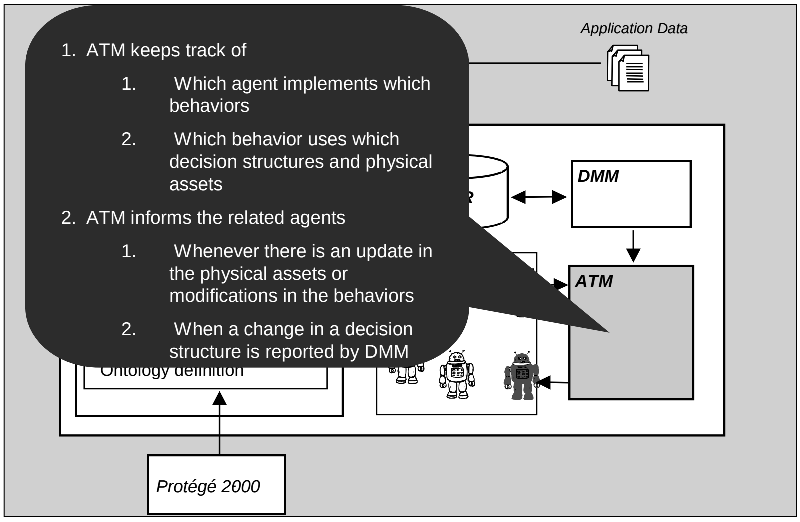
CERTH Agent Intelligence through Data Mining AUTH

Data Mining Module (DMM)

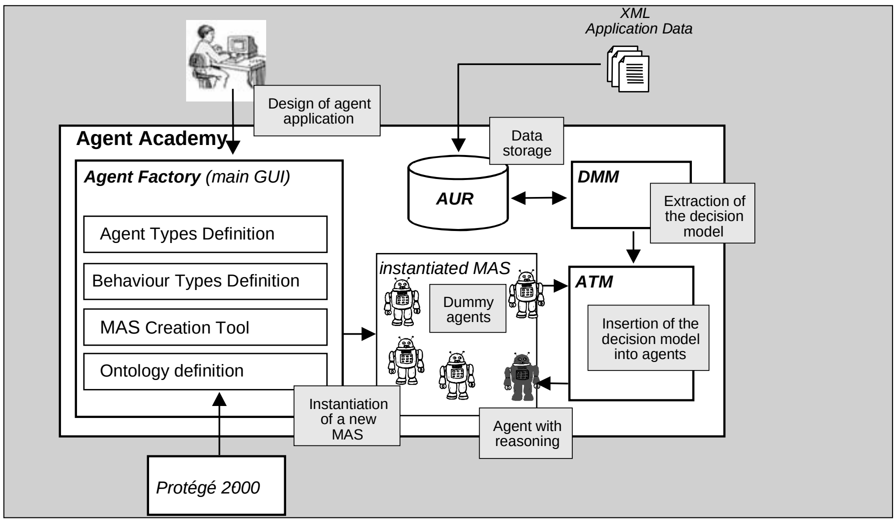
1. DMM applies data mining techniques on "application data" and "tracking data"
2. "Application data" describe the application domain
3. "Tracking data" reveal agent-specific interaction patterns
   (mainly used for retraining)
4. Selection from a certain data set
5. Choice of specific DM algorithms
6. Allows the user to evaluate, store, visualize data mining results



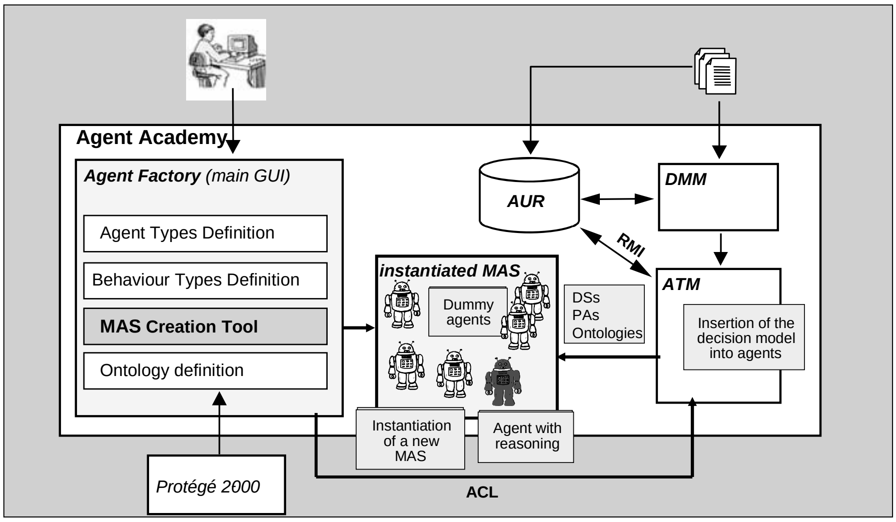
Agent Training Module (ATM)



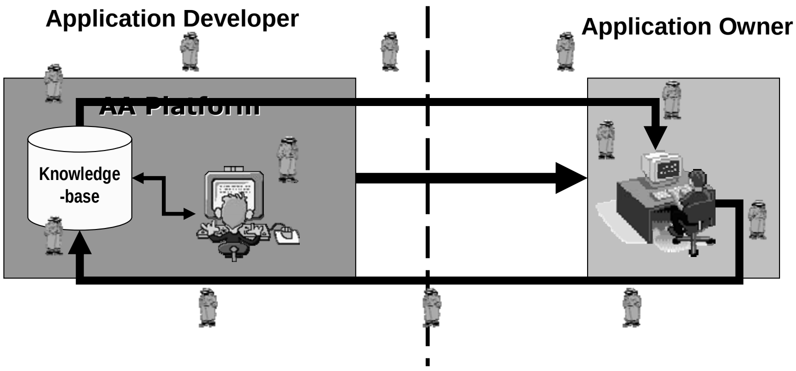
The Integrated Platform



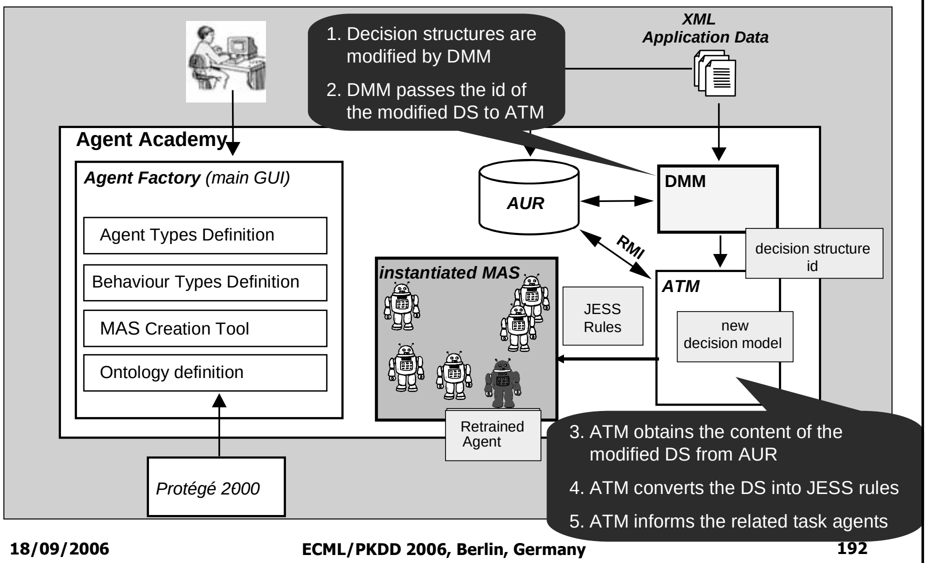
Multi-Agent Scenario Creation



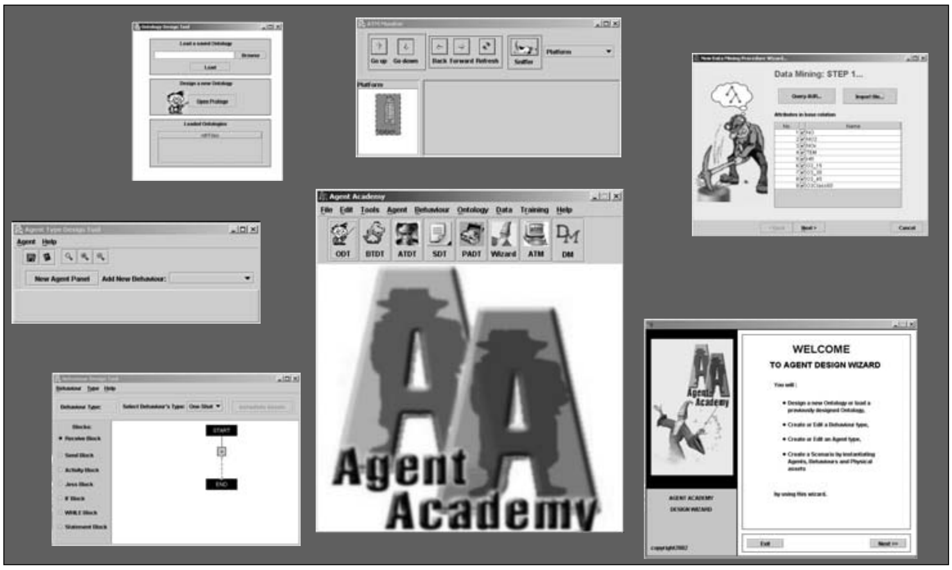
Agent Retraining



Retraining through DMM



AA-1.0…



Agent Intelligence through Data Mining

CERTH AUTH

Nevertheless…

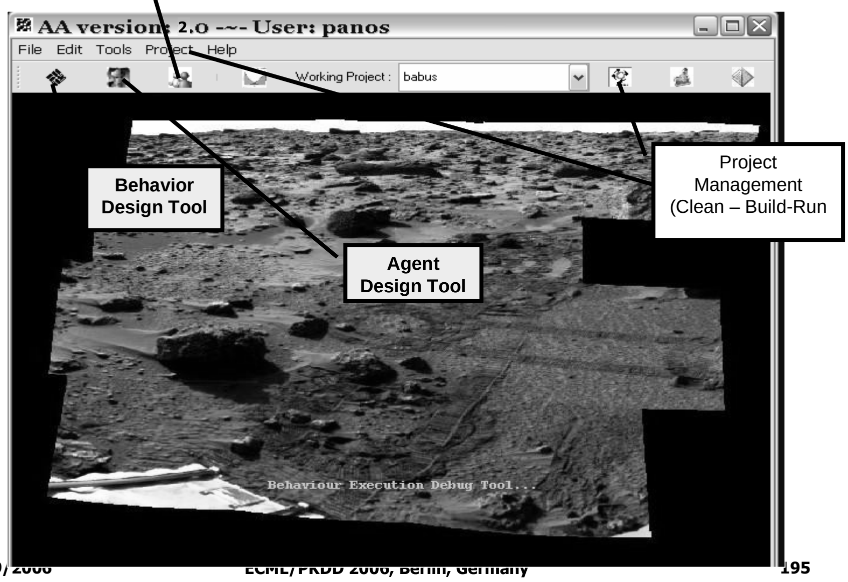
- Although ambitious, AA-1.0 was not very easy to use.
- No software code is produced (agent code is not transparent)
- Training and retraining is very complicated
- AA-1.0 is a MAS itself, inserting limitations on module communication
- This is why AA-2.0 has been released

18/09/2006 ECML/PKDD 2006, Berlin, Germany 194

Agent Intelligence through Data Mining

CERTH AUTH

AA-2.0 Main Frame



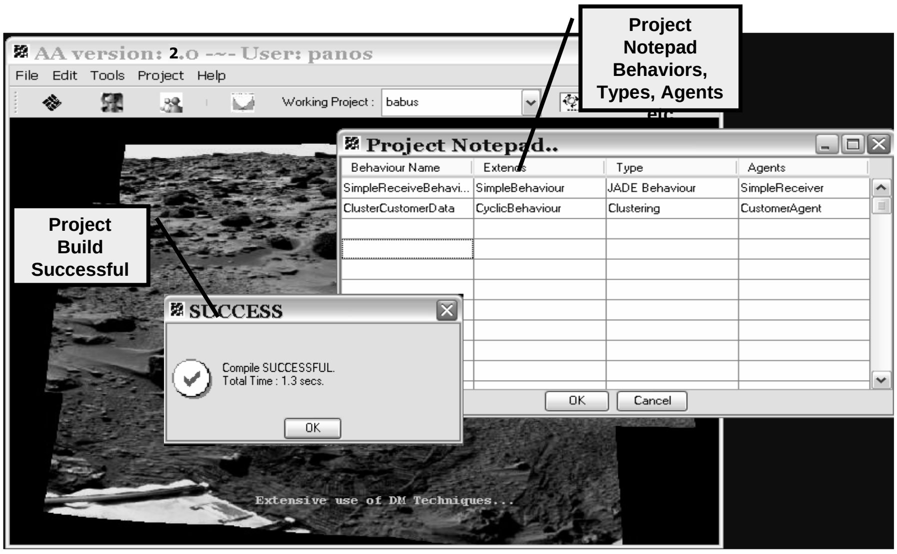
18/09/2006 ECML/PKDD 2006, Berlin, Germany 195

CERTH

Agent Intelligence through Data Mining

AUTH

AA-2.0 Project Actions



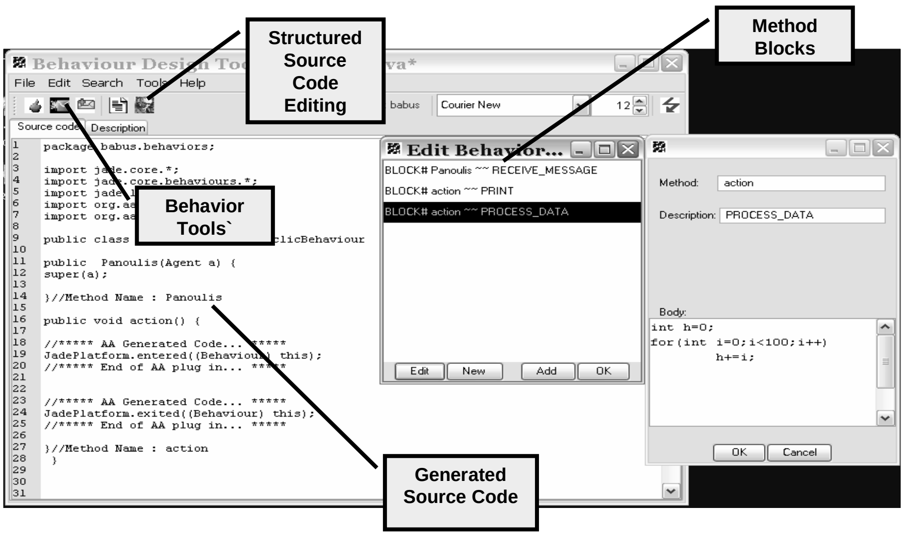
18/09/2006 ECML/PKDD 2006, Berlin, Germany 196

CERTH

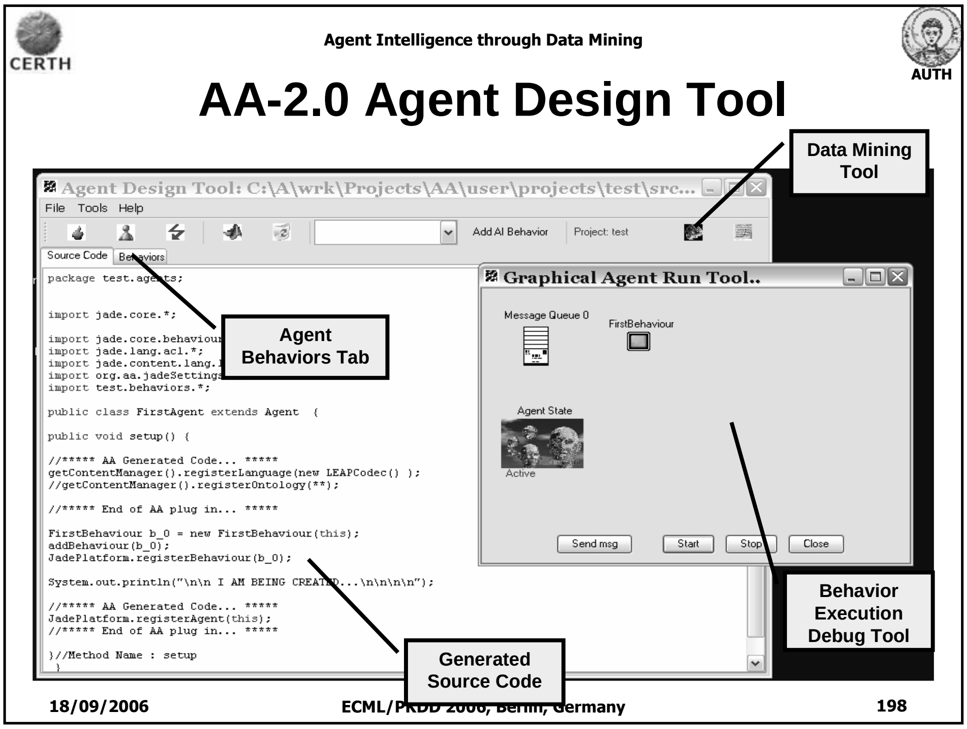
Agent Intelligence through Data Mining

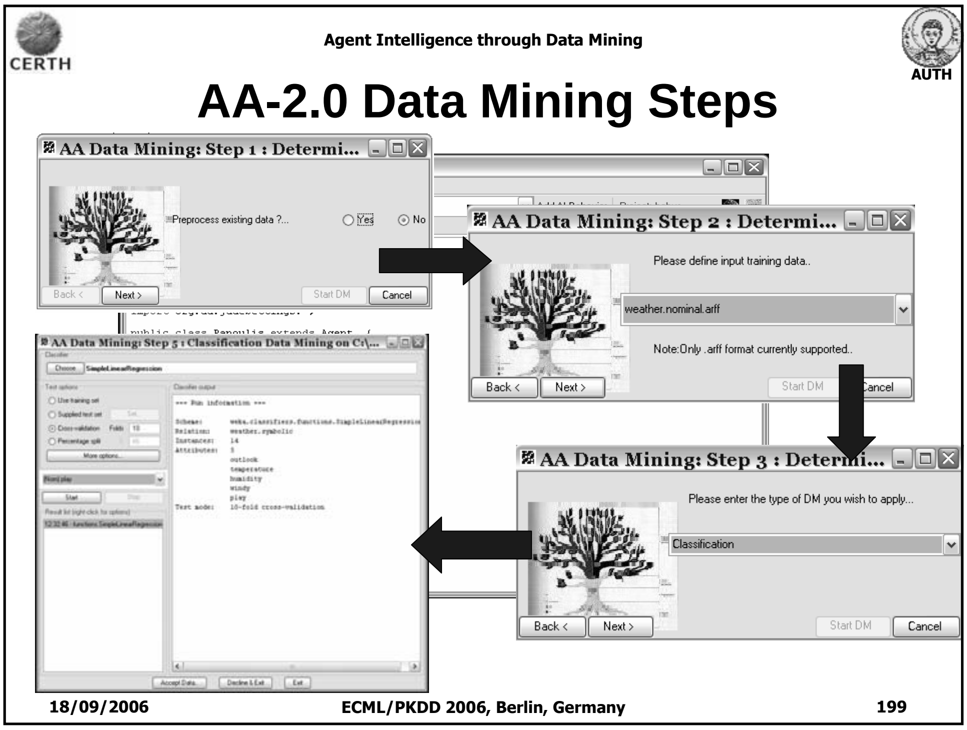
AUTH

AA-2.0 Behavior Design Tool

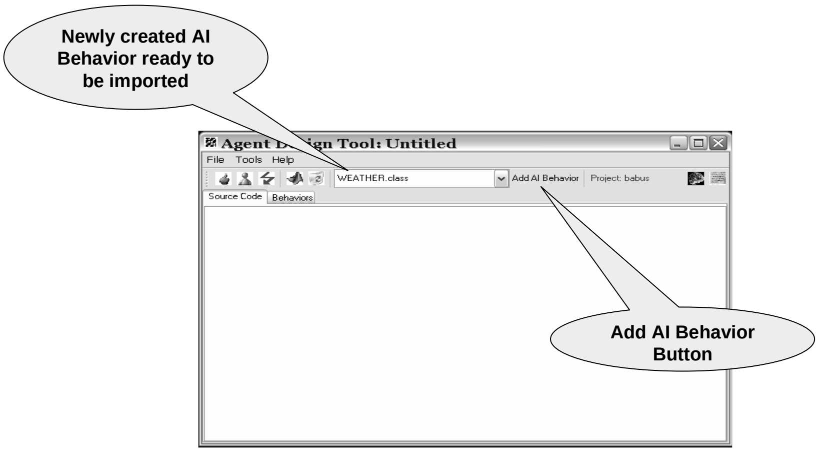


18/09/2006 ECML/PKDD 2006, Berlin, Germany 197

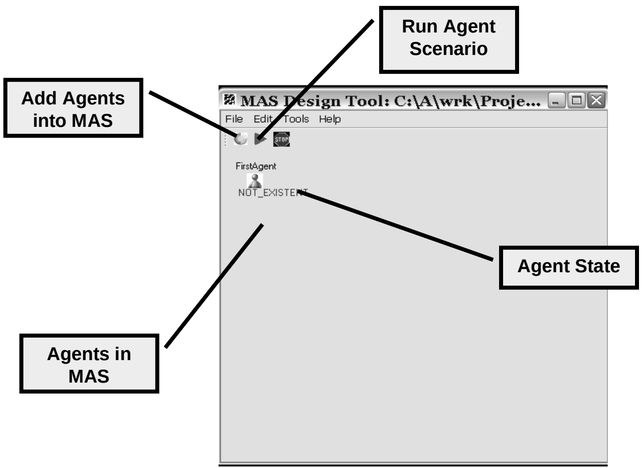

CERTH
AUTH
Agent Intelligence through Data Mining
AA-2.0 Agent Design Tool
Data Mining Tool
Agent Behaviors Tab
Graphical Agent Run Tool..
Behavior Execution Debug Tool
Generated Source Code
18/09/2006
198


CERTH
AUTH
Agent Intelligence through Data Mining
AA-2.0 Data Mining Steps
AA Data Mining: Step 1 : Determi...
Preprocess existing data ?...
AA Data Mining: Step 2 : Determi...
Please define input training data..
weather.nominal.arff
Note:Only .arff format currently supported.
AA Data Mining: Step 3 : Determi...
Please enter the type of DM you wish to apply...
Classification
18/09/2006
ECML/PKDD 2006, Berlin, Germany
199

AA-2.0 DM Integration



AA-2.0 MAS Design Tool

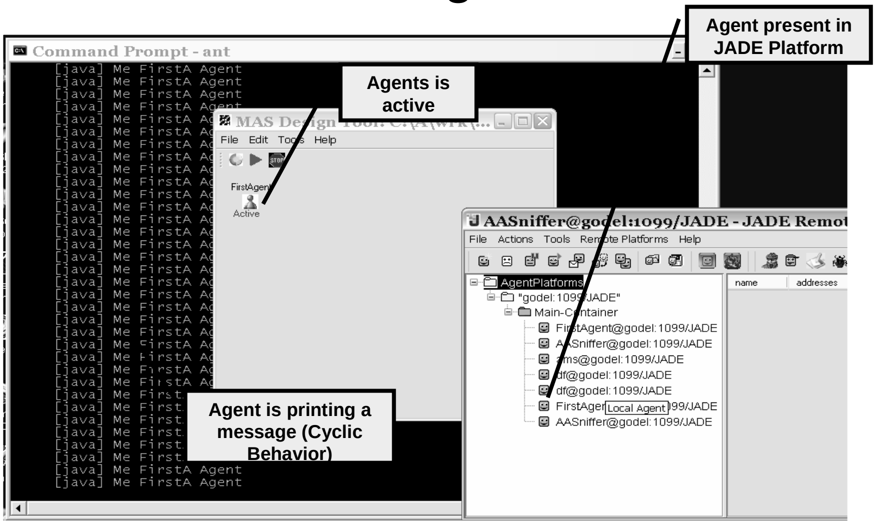


CERTH

Agent Intelligence through Data Mining

AUTH

AA-2.0 Running Scenarios



18/09/2006 ECML/PKDD 2006, Berlin, Germany 202

CERTH

Agent Intelligence through Data Mining

AUTH

Agent behavior-oriented DM…

18/09/2006 ECML/PKDD 2006, Berlin, Germany 203

Why Behavior-oriented DM (1/2)

- **Interaction** among agents is crucial for the efficiency of MAS
  - ✓ New agents enter into the system without the necessary knowledge and skills
  - ✓ New agents are not able to learn from the others' behavior
  - ✓ It is not possible to define and represent a priori the relevant knowledge the agents need for the interaction

Why Behavior-oriented DM (2/2)

- In order to improve its behavior, a new agent should act **consistently** with the knowledge and the behaviors (***culture***) of the other agents.
- A way for supporting multi-agent interaction based on the idea of

Implicit Culture (IC)

http://www.science.unitn.it/~pgiorgio/ic/

The main idea of Implicit Culture…

- A situation of the agent environment is represented as a set:

<a, σ, t>

where:

✓ a: set of agents

✓ σ: set of scenes (environment + actions)

✓ t: time slot

- The goal is to predict the next **executed action**
- **IC is realized through SICS…**

Systems for Implicit Culture Support (SICS)

- SICS main goals are to:

✓ **Establish** the implicit culture phenomenon and

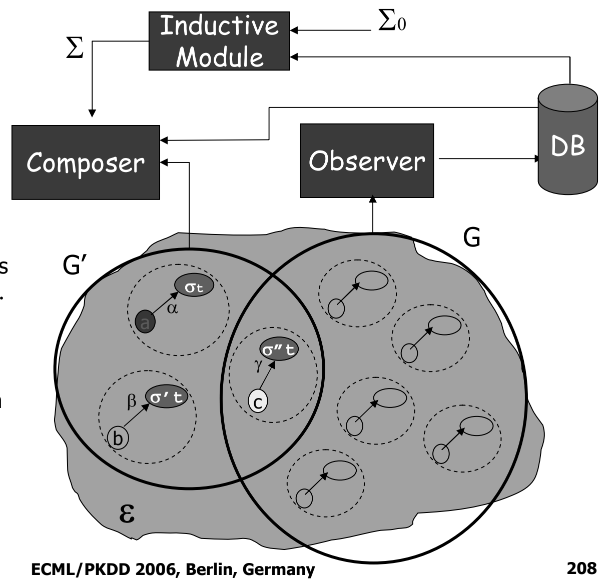
✓ **Propose** the next expected situated actions based on the already "played" scenes.

The SICS architecture...

Observer stores in a data base the situated executed actions of the agents of G.

Inductive Module that, using the data of the DB and the a priori theory Σ_0, induces a cultural constraint theory Σ.

Composer proposes to a group G' a set of scenes such that the expected situated actions satisfy Σ.

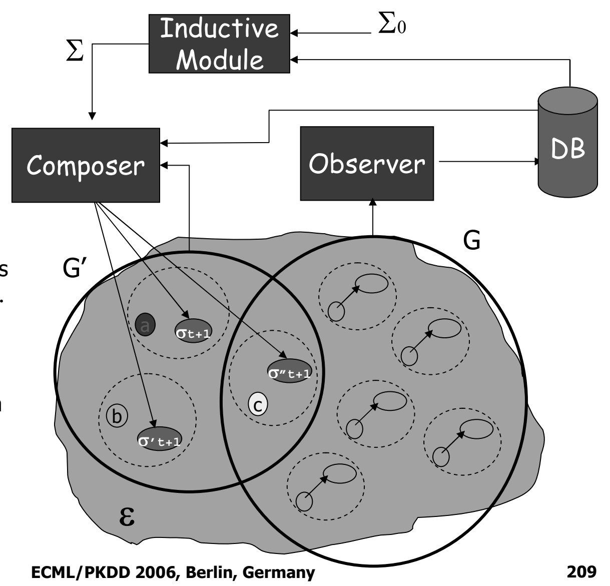


The SICS architecture...

Observer stores in a data base the situated executed actions of the agents of G.

Inductive Module that, using the data of the DB and the a priori theory Σ_0, induces a cultural constraint theory Σ.

Composer proposes to a group G' a set of scenes such that the expected situated actions satisfy Σ.

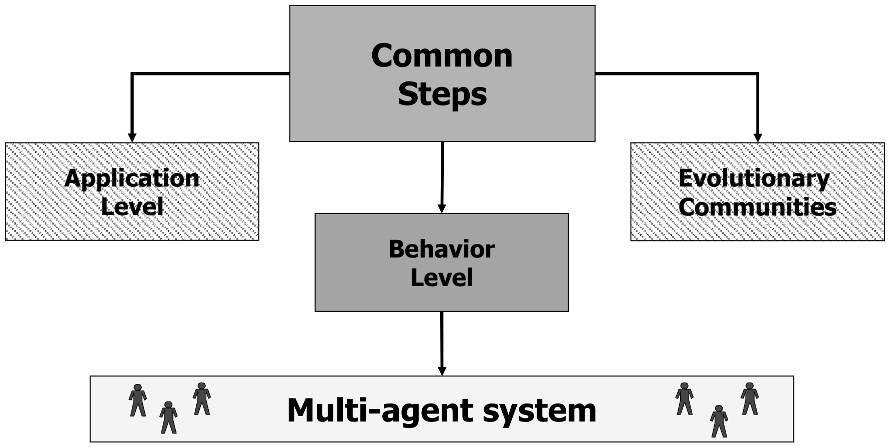


CERTH

Agent Intelligence through Data Mining

AUTH

The methodology



18/09/2006 ECML/PKDD 2006, Berlin, Germany 210

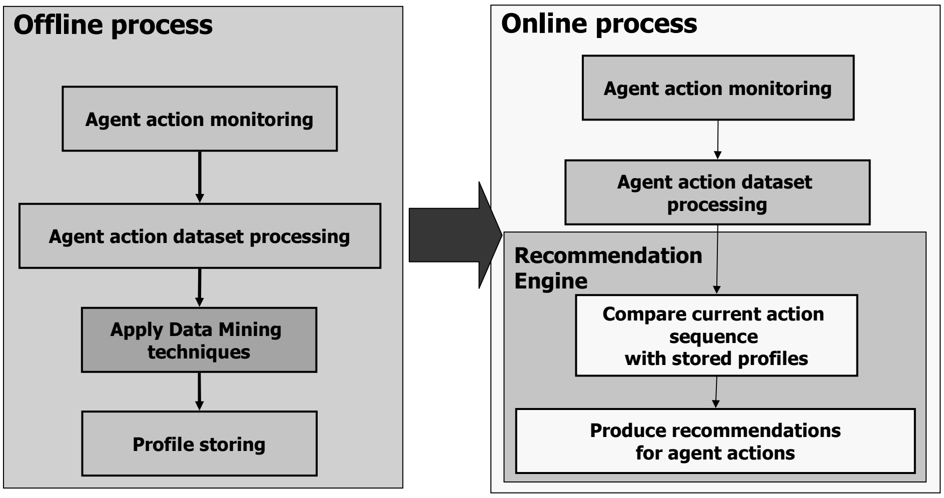
CERTH

Agent Intelligence through Data Mining

AUTH

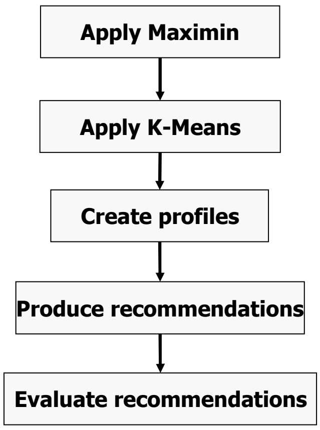
18/09/2006 ECML/PKDD 2006, Berlin, Germany 211

Behavior level: The training framework



The developed DM algorithm

- **κ-Profile** is a mechanism, which:
  - ✓ Represents agent actions,
  - ✓ Creates agent action profiles,
  - ✓ Produces recommendations, and
  - ✓ Evaluates them.



The parameters to determine...

- In order to develop a MAS with recommendation abilities (that follows the presented methodology), the parameters that have to be specified are:
  1. All possible agent actions.
  2. The specific MAS goal.
  3. The fuzzy inference engine parameters.

Knowledge extraction in agent communities

The orientation…

- Self-organization is 100% pure AI (Artificial Intelligence) oriented
- A lot of research work on this topic…

The goal of self-organizing MAS…

- To achieve a certain MAS goal, through "**self-learning**"
- Self-learning has to do with the **penalties** or **awards** the MAS receives whenever it performs an action.
- The most difficult part in developing a self-organizing MAS is to find the best award/penalty function…

The main technologies…

- Neural Networks
  - ✓ Mainly used for **classification** (the output is categorical)
  - ✓ Really efficient when confronted with a specific problem
- Genetic Algorithms
  - ✓ Mainly used for **optimization** problems, where dynamic programming cannot be used
  - ✓ The correct representation of the problem is crucial

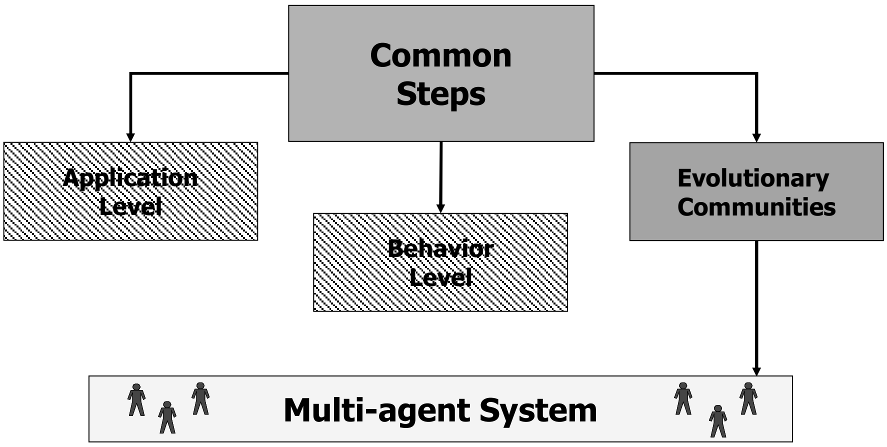
Agent Communities (1/2)

- Agent communities simulate problems that are complicated, heterogeneous and non-linear.
- The internal structure of such systems is inherently complicated, due to continually varying interactions.
- The goal of agent communities is collective, and its evaluation is done through indicators.

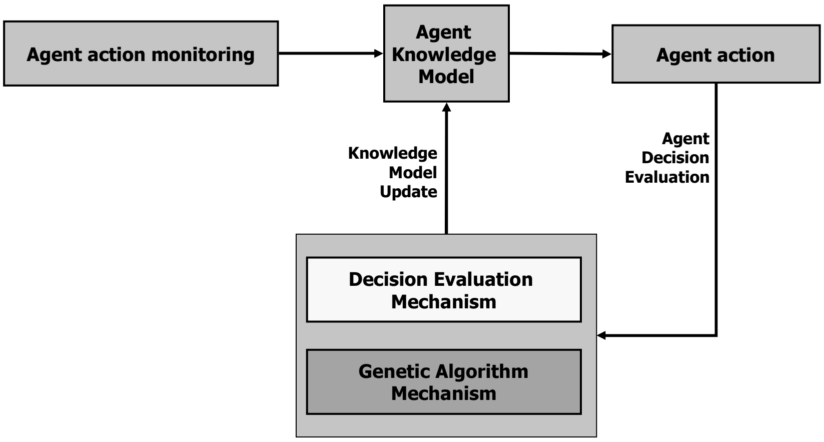
Agent Communities (2/2)

- Agent communities do not offer historical data, for DM techniques to be applied on and knowledge models to be extracted.
- The presented approach is based on way evolutionary techniques can be exploited in order to augment agent community intelligence.
- The knowledge extraction mechanism employed evaluates agent decisions (rewards or punishes) and updates the agent decision model.

The methodology



Evolutionary communities – training framework



Part 12 – MAS exploiting DM-extracted intelligence

CERTH AUTH

The presented systems…

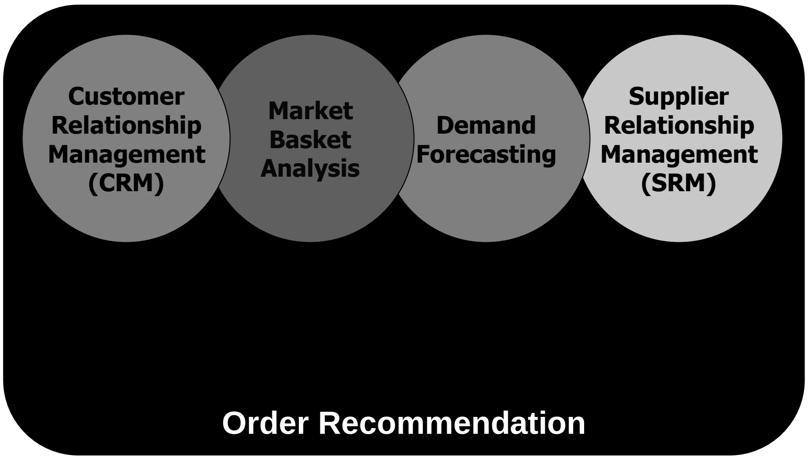
- An ERP add-on, providing intelligent policy recommendations on customer and supplier management.
- An intelligent environmental monitoring system.
- A decentralized maintenance management system.
- A self-organizing agent community “in danger”.
- An agent-based, e-auction system.

CERTH AUTH

The ERP add-on

- Provides an **Intelligent Shell on-top of an ERP**, working in close cooperation with it.
- Incorporates intelligence inside a business process: the handling of a **customer’s order**.
- **Makes a specific recommendation for a specific customer and a specific order**.
- Manipulates important ERP data with efficient algorithms for **reducing information overload**.

Main Goals (2/2)



Why do it with agents?

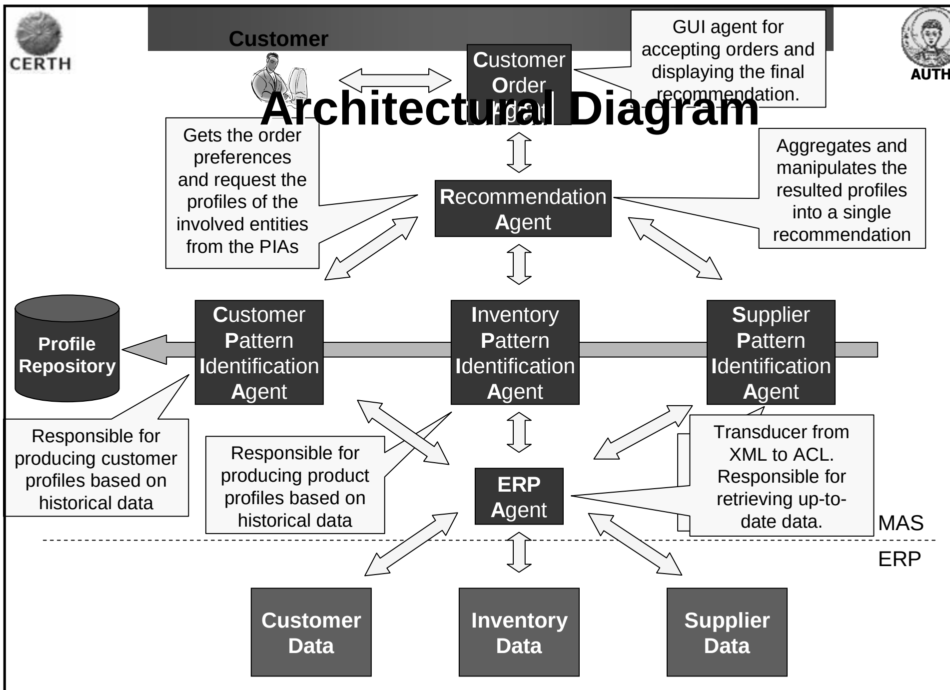
- Multiple Loci of Control:
  - ✓Changes in the goals and behaviors of one agent adequate to adjust the whole system.
- Software & Business Engineering Perspective:
  - ✓Coordination is done easily both in the design and the development phase.
- Adoption of Agents - no questions to us

CERTH AUTH

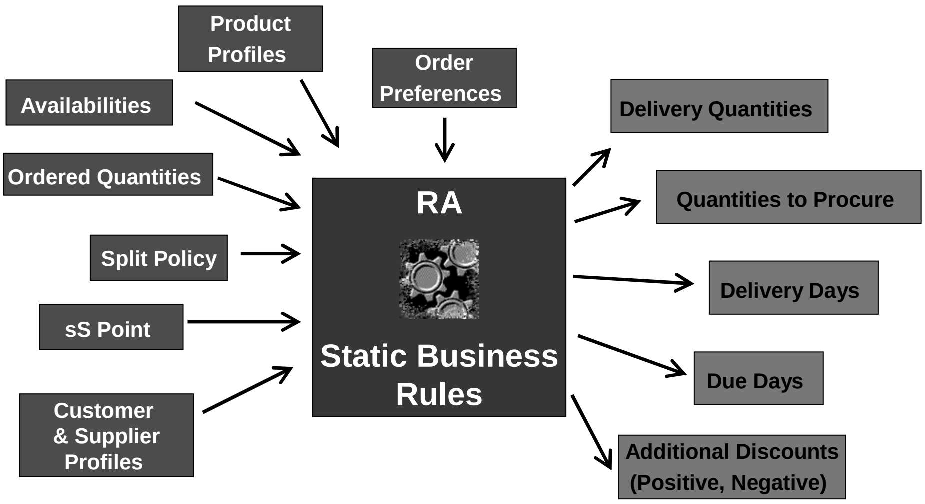
Why do it with data mining?

- Has been widely used for CRM
- Extended to SRM
- With Clustering
  - ✓ Deal with noisy data (missing, unknown or erroneous)
  - ✓ Find non-trivial, unknown, implicit, potentially useful patterns
  - ✓ Learn and provide quick responses
- Market Basket Analysis and Association Rules
  - ✓ Discover customer oriented buying habits
  - ✓ Discover market oriented buying habits

Architectural Diagram



Final Recommendation



Conclusions (1/2)

- Added value without rebuilding or reengineering the system
- Reduction of work effort through adaptability of business rules and automatic generation of the recommendation
- Reusable and customizable system (match different ERP platforms with minimum effort)

CERTH AUTH

Conclusions (2/2)

Improvement	Classic ERP	ERP + DKE
Market Basket Analysis	No	Yes
Recommendations	Indirectly, through reports	Automatically
Autonomy	No	Yes
Adaptability	Low	High
Customer Management & Pricing Policy	No	Yes
Supplier Management	No	Yes

CERTH AUTH

The system…

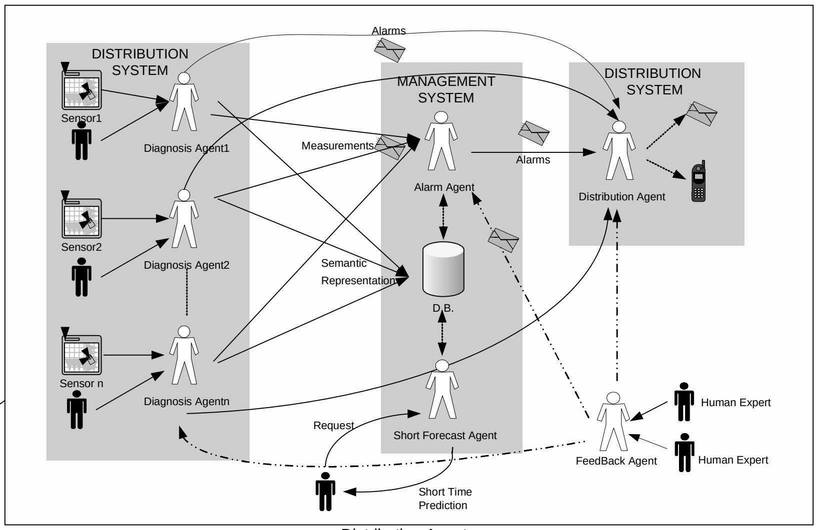
- Real-time Environmental Monitoring
- Agents detect ozone alarms and notify interested parties
- The application domain
  - ✓ Real–time systems management
  - ✓ Environmental assessment
  - ✓ Emergency real-time systems
  - ✓ Real time applications
  - ✓ Decision support systems

CERTH

Agent Intelligence through Data Mining

AUTH

The architectural diagram



Distribution Agent

18/09/2006 ECML/PKDD 2006, Berlin, Germany 236

CERTH

Agent Intelligence through Data Mining

AUTH

Diagnosis Agents

- Diagnosis Agents:
  - ✓Get measurements at real-time from sensors
  - ✓Verify that the sensors operate properly
- Inductive reasoning engines
  - ✓**Validate** incoming measurements
  - ✓**Predict** a missing ozone measurement level

18/09/2006 ECML/PKDD 2006, Berlin, Germany 237

CERTH AUTH

Knowledge Extraction

- From raw data to data-driven reasoning
  - ✓ Quinlan's C4.5 for decision tree induction
  - ✓ Two reasoning engines in Diagnosis Agent:
    1. Incoming Data Validation
    2. Missing Measurement Estimation

CERTH AUTH

Validation Reasoning Engine

Attributes

O3	The current ozone value
O3_30	The ozone value 30 min ago
O3_90	The ozone value 90 min ago
MinMax60	The difference between the maximum and the minimum ozone value in the last 60 min
MinMax150	The difference between the maximum and the minimum ozone value in the last 150 min
O3val	The corresponding validation tag (valid/erroneous)

Confusion Matrix

Validation Decision Model

Records classified as :	valid	erroneous
No. records in class 'valid':	34,454	21
No. records in class 'erroneous':	63	420

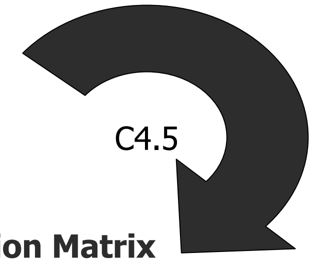
Decision Tree Parameters

Size of Decision Tree:	29 (15 Leaves)
Correctly classified records:	99.71%

Estimation Reasoning Engine

Attributes

NO	concurrent value of NO concentration
NO2	concurrent value of NO2 concentration
NOX	concurrent value of NOx concentration
TEM	concurrent value of Temperature
HR	concurrent value of Relative Humidity
O3_15	ozone value 15 min ago
O3_30	ozone value 30 min ago
O3Class	missing ozone value level (low/med)



Confusion Matrix

Estimation Decision Model

Records classified as :	low	med
No. Records in class 'low':	9905	2,351
No. Records in class 'med':	752	4,384

Decision Tree Parameters

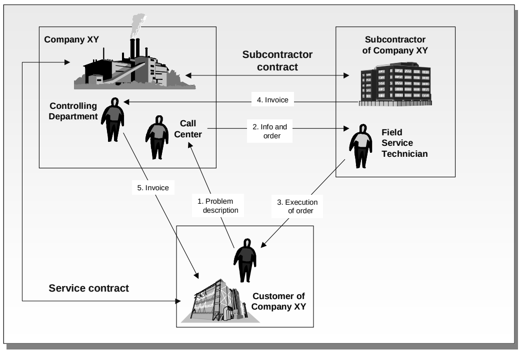
Size of Decision Tree:	29 (15 Leaves)
Correctly classified records:	93.80%

Conclusions on EMS

- In this scenario, agents act as **mediators**, delivering validated information in a distributed environment.
- Inductive Reasoning Agents seem suitable for building Intelligent Environmental Software Applications

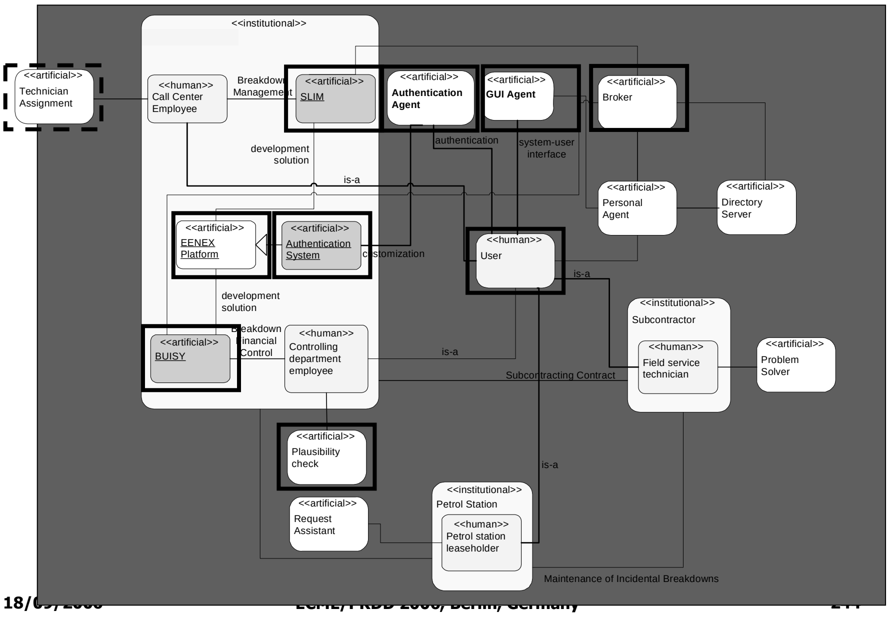
Distributed Service Management

Management of physical (maintenance) service to be carried out at the customer's location.



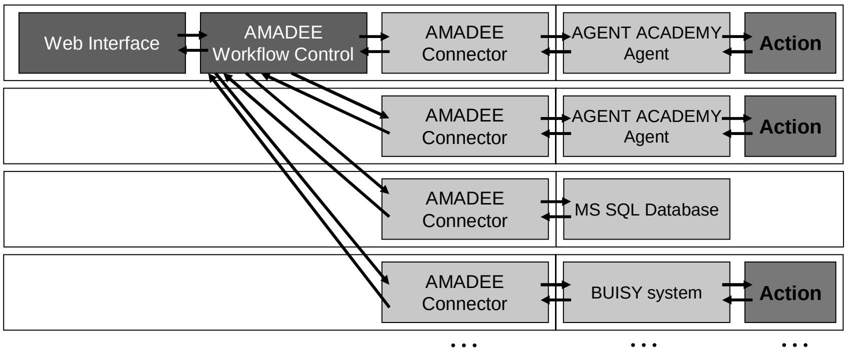
- Web based application
- Process monitoring
- Automated helpdesk
- Decision support
- Proactive event predictor
- Adaptability
- Customizability

Architectural Diagram



CERTH AUTH

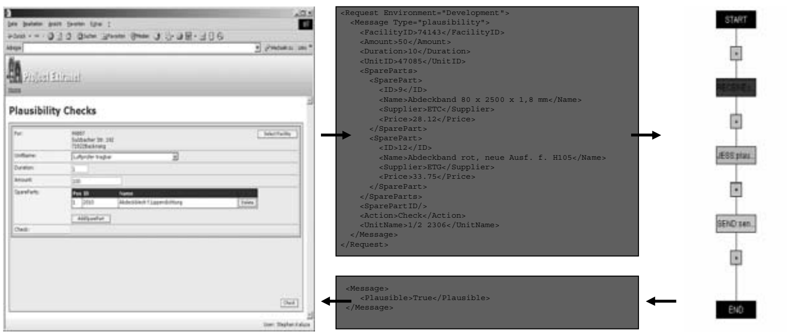
Embedding Agents into a Legacy Architecture

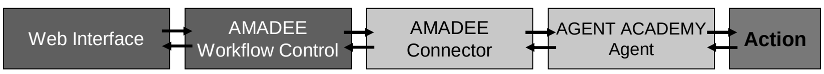


-Full process Control through AMADEE Workflow Control
-Seamless embedding in client‘s and AA architecture

CERTH AUTH

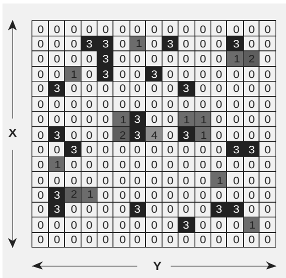
Agent Communication Flow (for Plausibility checks)



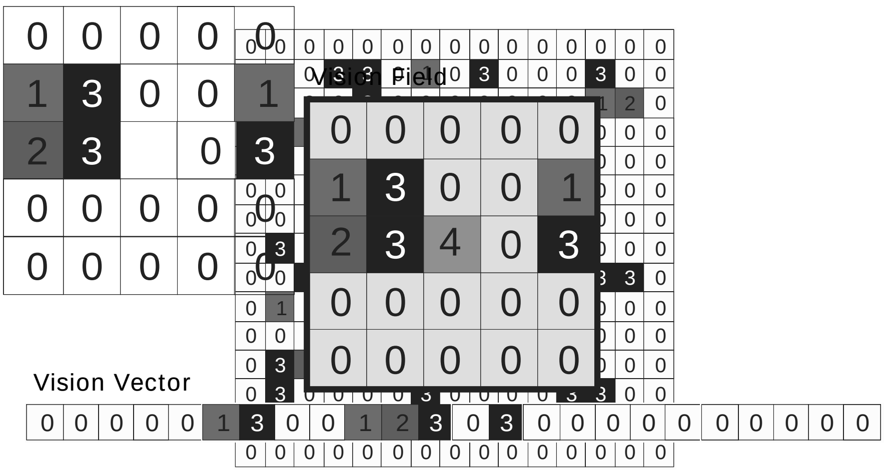


Biotope: A self-organizing agent community

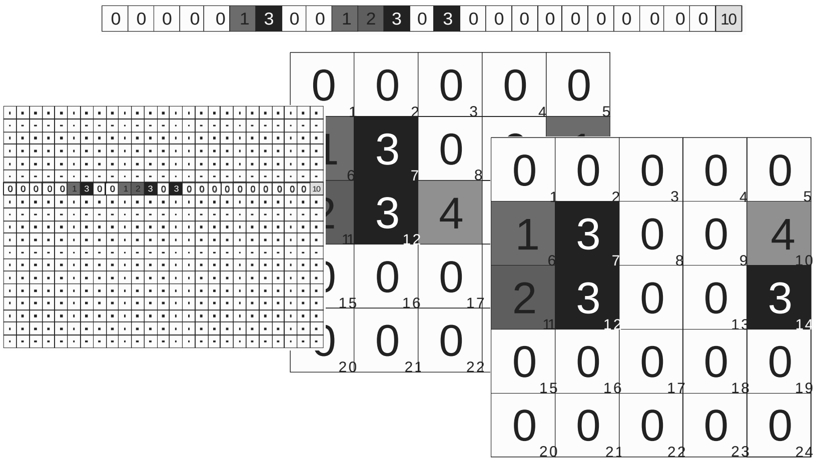
- Simulates a **parametrical** ecosystem, with food, traps and obstacles.
- The living organisms of this ecosystem are **intelligent agents**.
- The agents live in an unknown, probably “**hostile**” environment.



Agent Sight



Agent Movement



Agent Intelligence

- They perceive part of the environment (according to their **vision capabilities**), and learn what to eat and what to avoid.
- They **communicate** with each other (when "in sight") and exchange knowledge

Vision Vector

0	0	0	0	0	1	3	0	0	1	2	3	0	3	0	0	0	0	0	0	0	0	0	0

The scope of Biotope

- To use the simulating environment for evolution experiments.
- To develop the suitable tools for monitoring the evolution of the ecosystem.
- To monitor and model agent behaviors.
- To **enhance** agent intelligence by the use of genetic algorithms

Agent Reproduction

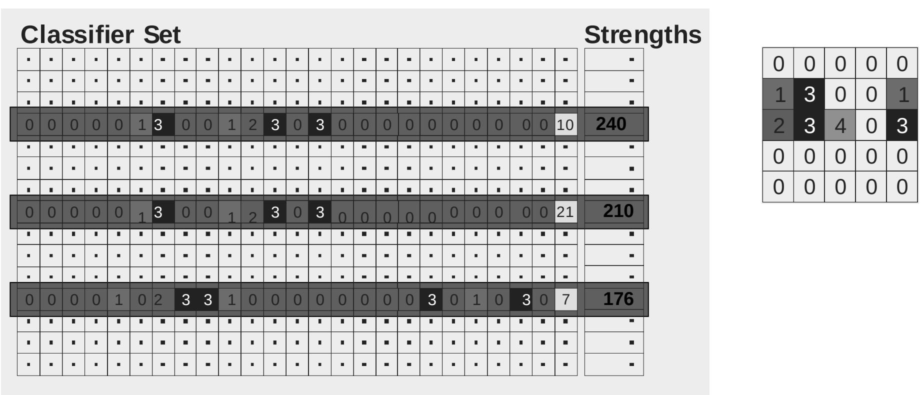
- Reproduction condition: Energy > 800 units
- Reproduction Outcome
  - 1 offspring
  - Inherits part of the parent's Classifier set
- Dispersal distance d_s
  - Distance: $Exp(d_s) \sim (1/m) \cdot exp(-d_s/m)$
  - Orientation: $\theta \sim (0,2\pi)$
- Energy variation
  - Parent: exponential decrease 400 --> 0
  - Offspring: exponential increase 200 --> 600

CERTH Agent Intelligence through Data Mining AUTH

18/09/2006 ECML/PKDD 2006, Berlin, Germany 253

CERTH Agent Intelligence through Data Mining AUTH

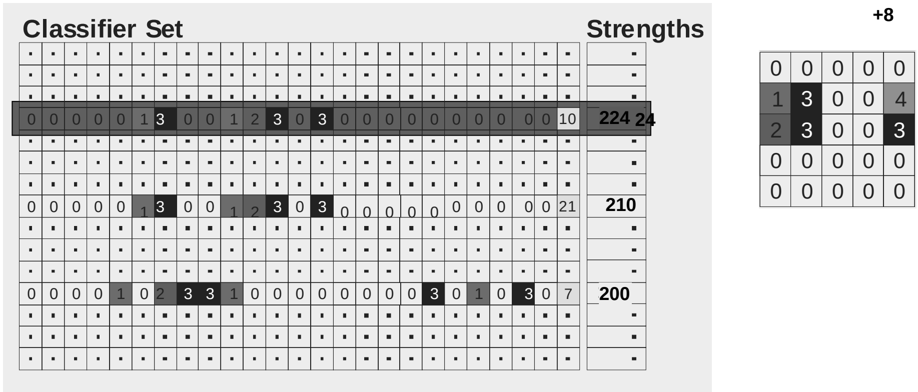
Rule evaluation



18/09/2006 ECML/PKDD 2006, Berlin, Germany 254

CERTH Agent Intelligence through Data Mining AUTH

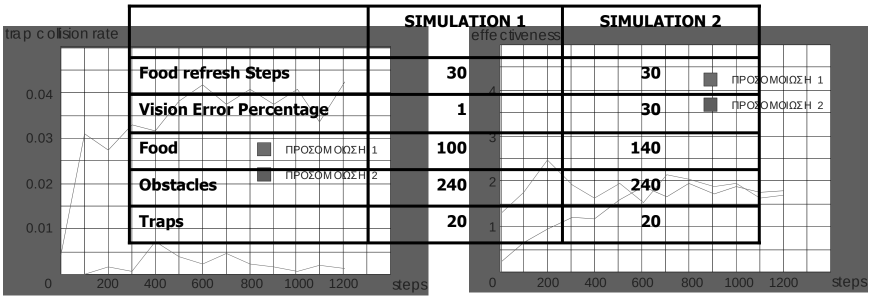
Rule evaluation



18/09/2006 ECML/PKDD 2006, Berlin, Germany 255

CERTH Agent Intelligence through Data Mining AUTH

Different Environments - Conclusions (1/2)…

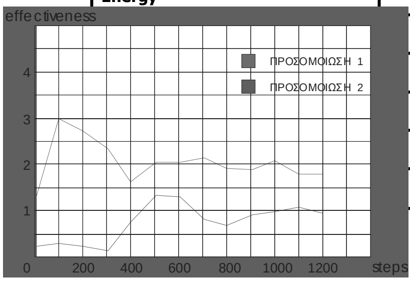


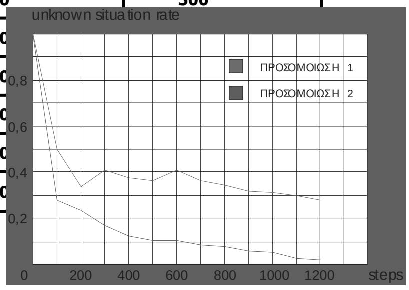
	SIMULATION 1	SIMULATION 2
Food refresh Steps	30	30
Vision Error Percentage	1	30
Food	100	140
Obstacles	240	240
Traps	20	20

18/09/2006 ECML/PKDD 2006, Berlin, Germany 256

Different Agents - Conclusions (2/2)…

	SIMULATION 1	SIMULATION 2
Energy	500	500
	3000	
	3000	
	200	
	200	
	20	





Environmental Indicators

- Resource availability
- Environmental variety
- Environmental reliability

CERTH AUTH

Agent Performance Indicators

- Energy
- Effectiveness
- Aging
- Food Consumption Rate
- Trap Collision Rate
- Reproduction Rate
- Unknown Situation Rate

CERTH AUTH

Conclusions & Future Directions

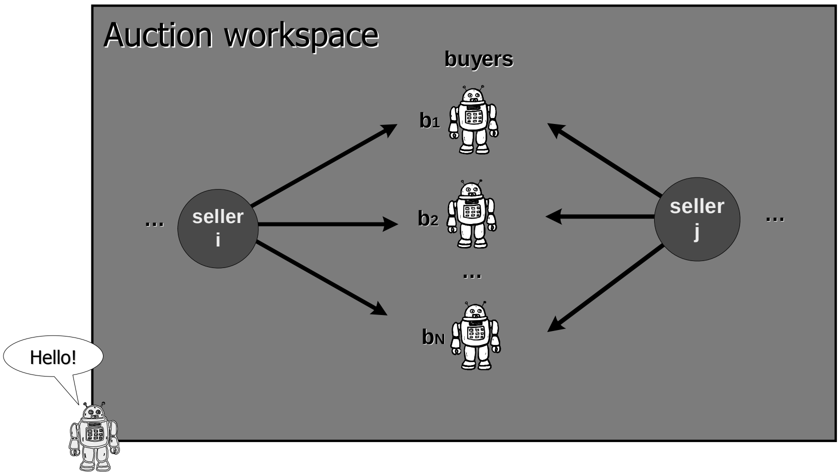
- In agent communities, the emphasis is on how to model the **common** problem.
- The combination of Genetic Algorithms, a classifier system, and an advanced agent communication framework, proves capable of handling dynamic and complex problems.
- The Biotope infrastructure can be used to model and simulate ***distributed computational systems***, where the agents are the computational entities, food represents the resources of the system, traps resource losses, and obstacles represent system incompatibilities.
- A test case could be ***a community of agents roaming the Web***, either collaborating or competing over its digital resources, while fragmentarily perceiving their environments.

An agent-based e-auction environment

- Protocol resembles to English double auctions
- Many different strategies deployed by agents
- Agents: multiple **Buyers** and multiple **Sellers** that trade over many goods

Agent Intelligence through Data Mining

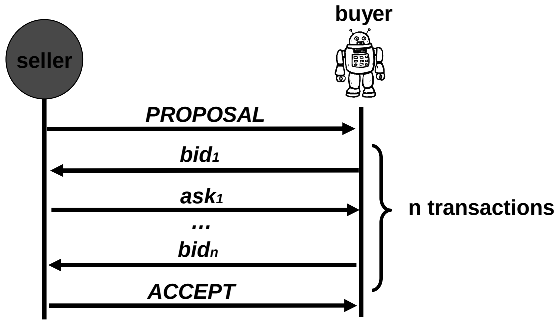
The auction environment



CERTH Agent Intelligence through Data Mining AUTH

Negotiating agents

- Negotiations take place between a buyer and a seller in order to reach an agreement

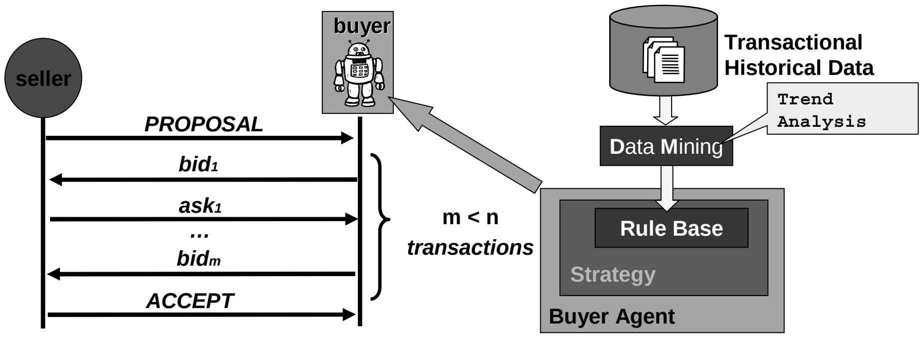


18/09/2006 ECML/PKDD 2006, Berlin, Germany 263

CERTH Agent Intelligence through Data Mining AUTH

Improving the behavior of a buyer

- Reducing communication by training



18/09/2006 ECML/PKDD 2006, Berlin, Germany 264

CERTH AUTH

How can DM improve decisions in auctions?

- The number of messages in a negotiation can be decreased
- DM can provide a tool for identifying hidden patterns (trends) in the history of previous transactions
- Evaluation criteria can be used to update-improve an agent’s rule base (e.g. fuzzy criteria)

CERTH AUTH

Part 13 – Open Issues

CERTH AUTH

Data mining and intelligent agents (1/2)

- How to determine **safety** and **soundness** in multi-agent systems
- How to specify a methodology for developing intelligent (through data mining) multi-agent applications
- When and how to perform agent **retraining**

CERTH AUTH

Data mining and intelligent agents (2/2)

- How to develop self-improving agents through dm
- How to develop tools and techniques for data mining agent behavior
- How to specify dm metrics that take **semantics** seriously into account
- How to **evaluate** agent "intelligence"

Agent Intelligence through Data Mining

CERTH AUTH

Thank you

Andreas L. Symeonidis

Research Associate
Informatics & Telematics Institute
Center for Engineering Research
and Technology – Hellas (CERTH)
Email:asymeon@iti.gr

Pericles A. Mitkas

Associate Professor
Electrical and Computer Engineering
Aristotle Univ. of Thessaloniki
Email: mitkas@eng.auth.gr

18/09/2006 ECML/PKDD 2006, Berlin, Germany 269