More Data/Web Mining Links

- http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-1.html
  “Introduction to Data Mining” by Prof. Zdravko Markov
- http://www.nature.com/nature/webmatters/agents/agents.html
- http://www.w3.org/2001/sw/ W3C Semantic Web
- http://www.w3.org/RDF/ Resource Description Framework (RDF)
- http://www.w3.org/TR/webont-req/ OWL Web Ontology Language
- http://www.daml.org/ The DARPA Agent Markup Language
- http://research.microsoft.com/adapt/MSBNx/ Microsoft Bayesian Network Editor
- http://www.the-data-mine.com/ The Data Mine
- http://research.microsoft.com/~dmax/WinMine/Tooldoc.htm WinMine Toolkit
Data Mining

• Data mining finds valuable information hidden in large volumes of data.
• Data mining is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data.
• The computer is responsible for finding the patterns by identifying the underlying rules and features in the data.
• It is possible to "strike gold" in unexpected places as the data mining software extracts patterns not previously discernible or so obvious that no-one has noticed them before.
• Mining analogy:
  – large volumes of data are sifted in an attempt to find something worthwhile.
  – in a mining operation large amounts of low grade materials are sifted through in order to find something of value.
Data Mining Goals

• **Classification**
  – DM system learns from examples or the data how to partition or classify the data i.e. it formulates classification rules
  – Example - customer database in a bank
    • Question - Is a new customer applying for a loan a good investment or not?
    • Typical rule formulated:
      if STATUS = married and INCOME > 10000 and HOUSE_OWNER = yes
      then INVESTMENT_TYPE = good

• **Association**
  – Rules that associate one attribute of a relation to another
  – Set oriented approaches are the most efficient means of discovering such rules
  – Example - supermarket database
    • 72% of all the records that contain items A and B also contain item C
    • the specific percentage of occurrences, 72 is the confidence factor of the rule

• **Sequence/Temporal**
  – Sequential pattern functions analyze collections of related records and detect frequently occurring patterns over a period of time
  – Difference between sequence rules and other rules is the temporal factor
  – Example - retailers database can be used to discover the set of purchases that frequently precedes the purchase of a microwave oven
Stages of the Data Mining Process

- Data pre-processing
  - Heterogeneity resolution
  - Data cleansing
  - Data transformation
  - Data reduction
  - Discretization and generating concept hierarchies
- Creating a data model: applying Data Mining tools to extract knowledge from data
- Testing the model: the performance of the model (e.g. accuracy, completeness) is tested on independent data (not used to create the model)
- Interpretation and evaluation: the user bias can direct DM tools to areas of interest
  - Attributes of interest in databases
  - Goal of discovery
  - Domain knowledge
  - Prior knowledge or belief about the domain
Data Mining Applications

• Credit Assessment
• Stock Market Prediction
• Fault Diagnosis in Production Systems
• Medical Discovery
• Fraud Detection
• Hazard Forecasting
• Buying Trends Analysis
• Organizational Restructuring
• Target Mailing
• Knowledge Acquisition
• Scientific Discovery
• Semantics based Performance Enhancement of DBMS
Web Mining (recap)

- **Web content** mining - discovery of Web document content patterns (text mining).
- **Web structure** mining - discovery of hypertext/linking structure patterns
  - use hyperlinks to enhance text classification
  - page ranking
  - modeling and measuring the Web
- **Web usage** mining - discovery of web users activity patterns
  - mining web server logs
  - mining client machine access logs
Web Agents

• A good internet agent needs to be:
  – Communicative: Able to understand your goals, preferences and constraints.
  – Capable: Able to take actions rather than simply provide advice.
  – Autonomous: Able to act without the user being in control the whole time.
  – Adaptive: Able to learn from experience about both its tasks and about its users preferences.
Web Agent Research Sites

- [http://agents.umbc.edu/](http://agents.umbc.edu/) UMBC AgentWeb
- [http://www.isi.edu/integration/](http://www.isi.edu/integration/) Information Integration Research Group, University of South California
W3C Semantic Web and Resource Description Framework (RDF)

• The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF).

• The Resource Description Framework (RDF) integrates a variety of applications from library catalogs and world-wide directories to syndication and aggregation of news, software, and content to personal collections of music, photos, and events using XML as an interchange syntax. The RDF specifications provide a lightweight ontology system to support the exchange of knowledge on the Web.
OWL Web Ontology Language

• An *ontology* formally defines a common set of terms that are used to describe and represent a domain. Ontologies can be used by automated tools to power advanced services such as more accurate web search, intelligent software agents and knowledge management.