Data Mining & Association Rule Mining

Why Data Mining

- Credit ratings/targeted marketing:
  - Given a database of 100,000 names, which persons are the least likely to default on their credit cards?
  - Identify likely responders to sales promotions
- Fraud detection
  - Which types of transactions are likely to be fraudulent, given the demographics and transactional history of a particular customer?
- Customer relationship management:
  - Which of my customers are likely to be the most loyal, and which are most likely to leave for a competitor? :

Data Mining helps extract such information

Data mining

- Process of semi-automatically analyzing large databases to find patterns that are:
  - valid: hold on new data with some certainity
  - novel: non-obvious to the system
  - useful: should be possible to act on the item
  - understandable: humans should be able to interpret the pattern
- Also known as Knowledge Discovery in Databases (KDD)

Applications

- Banking: loan/credit card approval
  - predict good customers based on old customers
- Customer relationship management:
  - identify those who are likely to leave for a competitor.
- Targeted marketing:
  - identify likely responders to promotions
- Fraud detection: telecommunications, financial transactions
  - from an online stream of event identify fraudulent events
- Manufacturing and production:
  - automatically adjust knobs when process parameter changes

Applications (continued)

- Medicine: disease outcome, effectiveness of treatments
  - analyze patient disease history: find relationship between diseases
- Molecular/Pharmaceutical: identify new drugs
- Scientific data analysis:
  - identify new galaxies by searching for sub clusters
- Web site/store design and promotion:
  - find affinity of visitor to pages and modify layout

The KDD process

- Problem fomulation
- Data collection
  - subset data: sampling might hurt if highly skewed data
  - feature selection: principal component analysis, heuristic search
- Pre-processing: cleaning
  - name/address cleaning, different meanings (annual, yearly), duplicate removal, supplying missing values
- Transformation:
  - map complex objects e.g. time series data to features e.g. frequency
- Choosing mining task and mining method:
- Result evaluation and Visualization:

Knowledge discovery is an iterative process

Relationship with other fields

- Overlaps with machine learning, statistics, artificial intelligence, databases, visualization but more stress on
  - scalability of number of features and instances
  - stress on algorithms and architectures whereas foundations of methods and formulations provided by statistics and machine learning.
  - automation for handling large, heterogeneous data

Some basic operations

- Predictive:
  - Regression
  - Classification
  - Collaborative Filtering
- Descriptive:
  - Clustering / similarity matching
  - Association rules and variants
  - Deviation detection

Association Rules

Association rules

- Given set T of groups of items
- Example: set of item sets purchased
- Goal: find all rules on itemsets of the form a-->b such that
  - support of a and b > user threshold s
  - conditional probability (confidence) of b given a > user threshold c
- Example: Milk --> bread
- Purchase of product A --> service B

T

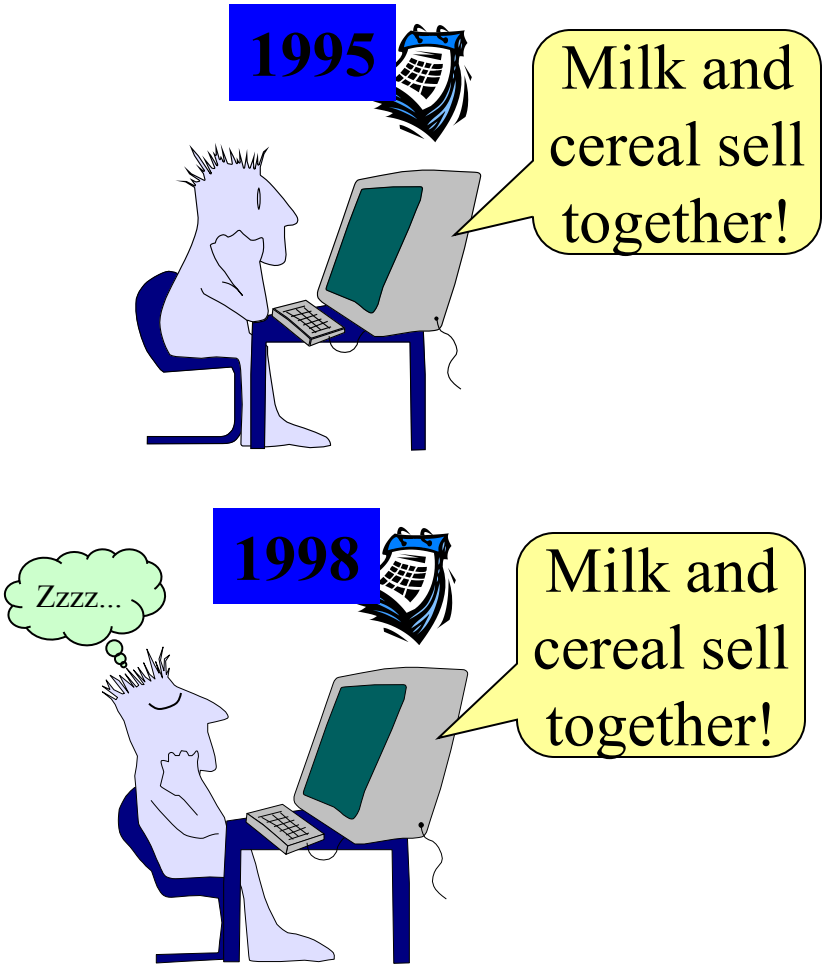
Milk, cereal
Tea, milk
Tea, rice, bread
cereal

Variants

- High confidence may not imply high correlation
- Use correlations. Find expected support and large departures from that interesting..
  - see statistical literature on contingency tables.
- Still too many rules, need to prune...

Prevalent ≠ Interesting

- Analysts already know about prevalent rules
- Interesting rules are those that *deviate* from prior expectation
- Mining's payoff is in finding *surprising* phenomena



What makes a rule surprising?

- Does not match prior expectation
  - Correlation between milk and cereal remains roughly constant over time
- Cannot be trivially derived from simpler rules
  - Milk 10%, cereal 10%
  - Milk and cereal 10% … surprising
  - Eggs 10%
  - Milk, cereal and eggs 0.1% … surprising!
  - Expected 1%

Applications of fast itemset counting

Find correlated events:

- Applications in medicine: find redundant tests
- Cross selling in retail, banking
- Improve predictive capability of classifiers that assume attribute independence
- New similarity measures of categorical attributes [Mannila et al, KDD 98]

Data Mining in Practice

Application Areas

Industry	Application
Finance	Credit Card Analysis
Insurance	Claims, Fraud Analysis
Telecommunication	Call record analysis
Transport	Logistics management
Consumer goods	promotion analysis
Data Service providers	Value added data
Utilities	Power usage analysis

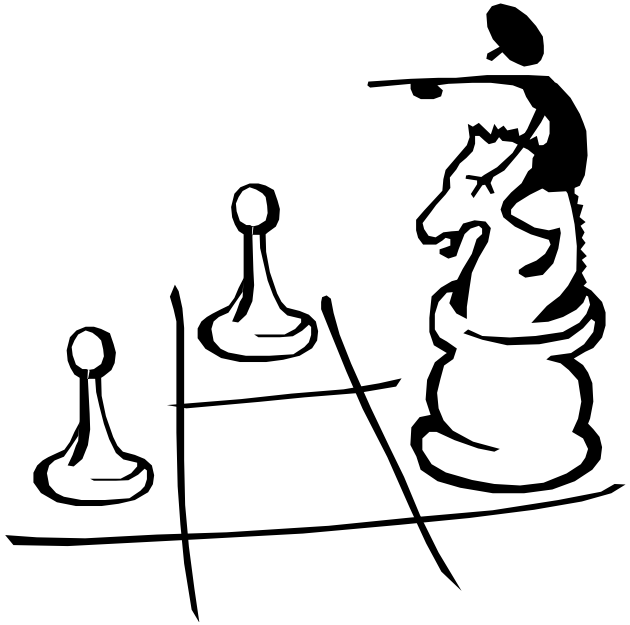
Why Now?

- Data is being produced
- Data is being warehoused
- The computing power is available
- The computing power is affordable
- The competitive pressures are strong
- Commercial products are available

Data Mining works with Warehouse Data

- Data Warehousing provides the Enterprise with a memory

Data Mining provides the Enterprise with intelligence



Usage scenarios

- Data warehouse mining:
  - assimilate data from operational sources
  - mine static data
- Mining log data
- Continuous mining: example in process control
- Stages in mining:
  - data selection → pre-processing: cleaning → transformation → mining → result evaluation → visualization

Mining market

- Around 20 to 30 mining tool vendors
- Major tool players:
  - Clementine,
  - IBM's Intelligent Miner,
  - SGI's MineSet,
  - SAS's Enterprise Miner.
- All pretty much the same set of tools
- Many embedded products:
  - fraud detection:
  - electronic commerce applications,
  - health care,
  - customer relationship management: Epiphany

Vertical integration:

Mining on the web

- Web log analysis for site design:
  - what are popular pages,
  - what links are hard to find.
- Electronic stores sales enhancements:
  - recommendations, advertisement:
  - Collaborative filtering: Net perception, Wisewire
  - Inventory control: what was a shopper looking for and could not find..

OLAP Mining integration

- OLAP (On Line Analytical Processing)
  - Fast interactive exploration of multidim. aggregates.
  - Heavy reliance on manual operations for analysis:
  - Tedious and error-prone on large multidimensional data
- Ideal platform for vertical integration of mining but needs to be interactive instead of batch.

State of art in mining OLAP integration

- Decision trees [Information discovery, Cognos]
  - find factors influencing high profits
- Clustering [Pilot software]
  - segment customers to define hierarchy on that dimension
- Time series analysis: [Seagate's Holos]
  - Query for various shapes along time: eg. spikes, outliers
- Multi-level Associations [Han et al.]
  - find association between members of dimensions
- Sarawagi [VLDB2000]

Data Mining in Use

- The US Government uses Data Mining to track fraud
- A Supermarket becomes an information broker
- Basketball teams use it to track game strategy
- Cross Selling
- Target Marketing
- Holding on to Good Customers
- Weeding out Bad Customers