Data Mining

Introduction to Modern Information Retrieval from Databases and the Web

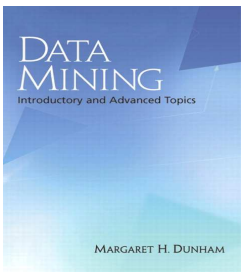
Administrivia

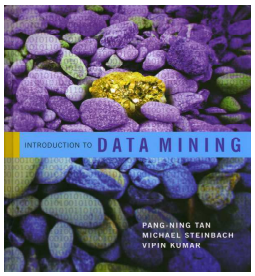
- Instructor:
  - Kostis Sagonas (MIC, Hus 1, 352)
- Course home page:
  http://user.it.uu.se/~kostis/Teaching/DM-05/
- If you want to be enrolled in the course, you have to also send, by November 11, a mail with your name, your UU e-mail address, your personnummer, and the course you are registering to:
  kostis@it.uu.se
- Assistants:
  - Per Gustafsson pergu@it.uu.se
  - Tobias Lindahl tobiasl@it.uu.se

Course Structure

- Course has theoretical and practical aspects
  - Need both in modern data mining!
- Written examination = tests understanding of main concepts/ideas covered in the course
- Assignments = practice and hands-on experience
  - Hand-in to & corrected by the assistants
  - Assignment deadlines are strict but you get one "free" extension (if you really need it)
  - Absolute final deadline: Easter 2006

Recommended Books





Papers from the recent DM literature

- In addition to lecture slides, various papers from the recent research on Data Mining are available at the course's homepage.
- Some of these handouts are **required** reading.

Course Syllabus

• Introduction to data mining	(today)
• Basic data mining techniques	(1 lecture)
• Classification	(2 lectures)
• Clustering	(2 lectures)
• Association rule mining	(3 lectures)
• Sequence mining	(1 lecture)
• Web mining	(1 lecture)
• Search engines	(1 lecture)
• Privacy-preserving data mining	(2 lectures)

Introduction to Data Mining

Lecture's Overview

- Motivation: Why data mining?
- What is data mining?
- Data mining: On what kind of data?
- Data mining functionality
- Are all the patterns interesting?
- Classification of data mining systems
- Major issues in data mining

Historical Introduction

Originally, "data mining" was a statistician's term for overusing data to draw invalid inferences

Bonferroni's theorem warns us that if there are too many possible conclusions to draw, some will be true for purely statistical reasons, with no physical validity.

(In)Famous example: David Rhine, a "parapsychologist" at Duke University in the 1950's tested students for "extrasensory perception" by asking them to guess 10 cards - red or black. About 1/1000 of them guessed all 10 cards correctly, and he declared them to have ESP. When he retested them, he found that they did no better than average. His conclusion:

"Telling people that they have ESP, causes them to lose it!"

"Necessity is the Mother of Invention"

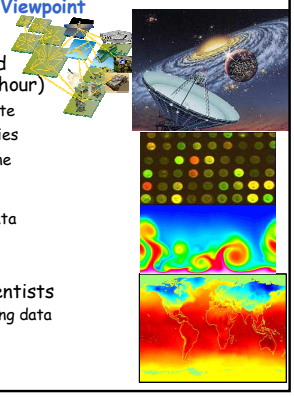
- Data explosion problem
  - Automated data collection tools and mature database (DB) technology lead to tremendous amounts of data stored in information repositories such as DBs and data warehouses
- We are drowning in data, but starving for knowledge!

Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at supermarkets or big department stores
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

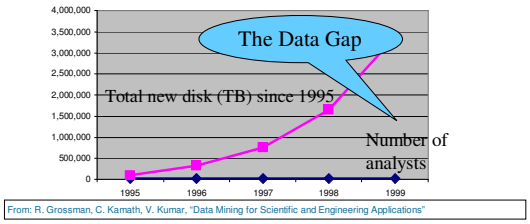
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in hypothesis formation



Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

What is Data Mining?

- Data mining (knowledge discovery in databases):
  - Efficient extraction of interesting (valid, non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- Alternative names and their "inside stories":
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- What is not data mining?
  - (Deductive) query processing
  - Expert systems or small ML/statistical programs.

Database vs. Data Mining Processing

- Query
  - Well defined
  - SQL
- Data
  - Operational
- Output
  - Precise
  - Subset of database

- Query
  - Poorly defined
  - No "standard" query language
- Data
  - Not operational
- Output
  - Fuzzy
  - Not a subset of database

Database Query Processing vs. Data Mining

- Database (OLTP/OLAP) Query
  - How many sofas did IKEA stores sell in the first 3 months of 2003 in Sweden vs. in Denmark ?
  - Find all IKEA customers who have bought a sofa worth more than 10.000 kronor.
  - Find all IKEA customers who have bought a sofa.
- Data Mining Queries
  - Which IKEA countries had anomalous sofa sales in the first 3 months of 2003 ? (classification)
  - Identify all buyers of IKEA sofas who have similar buying patterns. (clustering)
  - What other items do customers of IKEA frequently buy on the same visit that they purchase a sofa ? (association rules)
  - What are the most frequently bought sequences of items after the purchase of a sofa ? (sequence mining)

Why Data Mining? — Potential Applications

1. Database analysis and decision support
   - Market analysis and management
     - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
   - Risk analysis and management
     - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
   - Fraud & terrorist detection
- Other Applications
  - Text mining / spam filtering / document classification
  - Stream data mining
  - Web mining
  - Biology & Bio-Informatics (DNA data analysis)
  - Astronomy

Market Analysis and Management (1)

- Where are the data sources for analysis?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find *clusters* of "model" customers who share the same characteristics: interests, income level, spending habits, etc.
- Determine customer purchasing patterns over time
  - Conversion of single to a joint bank account: marriage, etc.
- Cross-market analysis
  - Associations/co-relations between product sales
  - Prediction based on the association information

Market Analysis and Management (2)

- Customer profiling
  - data mining can identify what types of customers buy what products (clustering or classification)
- Identify customer requirements
  - identify the "best" products for different customers
  - use prediction techniques to find what factors will attract new customers
- Provide summary information
  - various multi-dimensional summary reports
  - statistical summary information (data central tendency and variation)

Corporate Analysis and Risk Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - contingent claim analysis to evaluate assets
  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)
- Resource planning:
  - summarize and compare the resources and spending
- Competition:
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

Fraud Detection and Management (1)

- Applications
  - widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.
- Approach
  - use historical data to build models of fraudulent behavior and employ data mining to help identify similar instances
- Examples
  - auto insurance: detect groups of people who stage accidents to collect on insurance
  - money laundering: detect suspicious money transactions
  - medical insurance: detect professional patients or rings of doctors or rings of references

Fraud Detection and Management (2)

Detecting inappropriate medical treatment

- Australian Health Insurance Commission identifies that in many cases blanket screening tests were requested (saved Australian $1m/yr).

Detecting telephone fraud

- Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.
- British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.

Retail

- Analysts estimate that 38% of retail shrink is due to dishonest employees.

Other Successful Applications

Sports

- IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat

Astronomy

- JPL and the Palomar Observatory discovered 22 quasars with the help of data mining

Internet Web Surf-Aid

- IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Examples of "success stories"

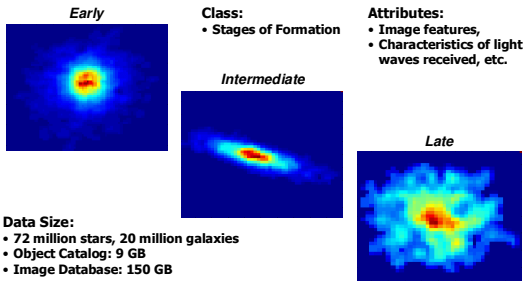
- Decision trees constructed from bank-loan histories, allow banks to decide whether to grant a loan and at which rate.
- Patterns of traveler behavior manage the sale of discounted tickets on planes, hotel rooms, etc.
- "Diapers and beer". Placement of items in supermarket shelves to increase their sales.
- Skycat and Sloan Sky Survey: clustering sky objects by their radiation levels in different bands allowed astronomers to distinguish between galaxies, nearby stars, and many other kinds of celestial objects.
- Comparison of genotype of people with/without a condition allowed discovery of sets of genes that together account for many cases of diabetes.

Application: Sky Survey Cataloging

- Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory)
  - 3000 images with 23,040 x 23,040 pixels per image
- Approach:
  - Segment the image
  - Measure image attributes (features) - 40 of them per object
  - Model the class based on these features
  - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classifying Galaxies



Class:
- **Stages of Formation**

Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

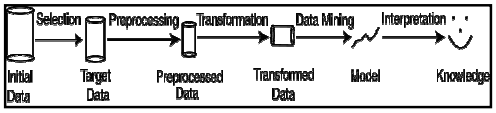
Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

Data Mining vs. KDD

Knowledge Discovery in Databases (KDD): process of finding useful information and patterns in data.

Data Mining: Use of algorithms to extract the information and patterns derived by the KDD process.

Data Mining: A part of the KDD Process



Selection: Obtain data from various sources.

Preprocessing: Cleanse data.

Transformation: Convert to common format. Transform to new format.

Data Mining: Obtain desired results.

Interpretation/Evaluation: Present results to user in meaningful manner.

Steps of the KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

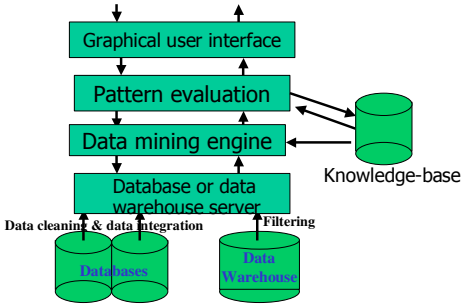
KDD Process Example: Web Log Analysis

- *Selection:*
  - Select log data (dates and locations) to use
- *Preprocessing:*
  - Remove identifying URLs
  - Remove error logs
- *Transformation:*
  - Sessionize (sort and group)
- *Data Mining:*
  - Identify and count patterns
  - Construct data structure
- *Interpretation/Evaluation:*
  - Identify and display frequently accessed sequences.
- *Potential User Applications:*
  - Cache prediction
  - Personalization

KDD Issues

- Human Interaction
- Overfitting
- Outliers
- Interpretation
- Visualization
- Large Datasets
- High Dimensionality
- Multimedia Data
- Missing Data
- Irrelevant Data
- Noisy Data
- Changing Data
- Integration
- Application

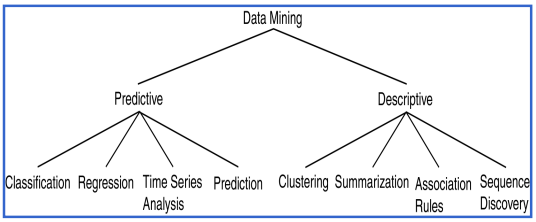
Architecture of a Typical Data Mining System



Data Mining: On What Kind of Data?

- Relational databases
- Data warehouses
- Transactional databases
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial and temporal data
  - Time-series data and stream data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW

Data Mining Models and Tasks



Basic Data Mining Tasks

Classification maps data into predefined groups or classes
- Supervised learning
- Pattern recognition
- Prediction

Regression is used to map a data item to a real valued prediction variable.

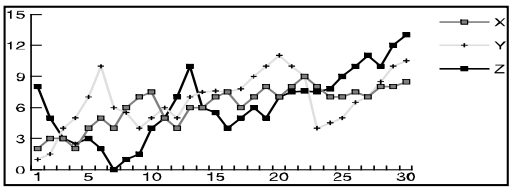
Clustering groups similar data together into clusters.
- Unsupervised learning
- Segmentation
- Partitioning

Summarization maps data into subsets with associated simple descriptions.
- Characterization
- Generalization

Link Analysis uncovers relationships among data.
- Affinity analysis
- Association rules
- Sequential analysis determines sequential patterns.

Example: Time Series Analysis

- Example: Stock Market
- Predict future values
- Determine similar patterns over time
- Classify behavior



Data Mining Functionalities (1)

- Concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Association (correlation and causality)
  - Multi-dimensional vs. single-dimensional association
  - age(X, "20..29") & salary(X, "20..29K") → buys(X, "PlasmaTV") [support = 2%, confidence = 60%]
  - buys(T, "DVD") → buys(T, "DVD movies") [support=1%, confidence=75%]

Data Mining Functionalities (2)

- Classification and Prediction
  - Finding models (functions) that describe and distinguish classes or concepts for future prediction
  - E.g., classify countries based on climate, or classify cars based on gas mileage
  - Presentation: decision-tree, classification rule, neural network
  - Prediction: Predict some unknown or missing numerical values
- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity

Data Mining Functionalities (3)

- Outlier analysis
  - 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
- Trend and evolution analysis
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses

Are All the "Discovered" Patterns Interesting?

- A data mining system/query may generate thousands of patterns, not all of them are interesting.
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**: 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
- **Objective vs. subjective interestingness measures:**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Can We Find All and Only Interesting Patterns?

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns?
  - Association vs. classification vs. clustering
- Search for only interesting patterns: Optimization
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First generate all the patterns and then filter out the uninteresting ones.
    - Generate only the interesting patterns—mining query optimization

Data Mining Communities

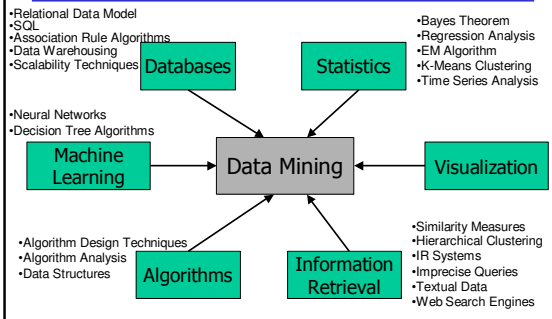
- Statistics
- AI, where it is called "machine learning"
- Researchers in clustering algorithms and in neural networks
- Visualization
- Databases

In this course, we will examine data mining (mostly) from a database perspective.

Database Perspective on Data Mining

- Scalability
- Real World Data
- Updates
- Ease of Use

Data Mining: Confluence of Multiple Disciplines



Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views, different classifications
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted

A Multi-Dimensional View of Data Mining Classification

- **Databases to be mined**
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Web-log analysis, etc.

Major Issues in Data Mining (1)

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

Major Issues in Data Mining (2)

- 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)
- 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

Other Data Mining Issues

- Social Implications of Data Mining
  - Privacy
  - Profiling
  - Unauthorized use
- Data Mining Success Metrics
  - Usefulness
  - Return on Investment (ROI)
  - Accuracy
  - Complexity measures

  We will only consider Space/Time success measures

Summary

- Data mining: efficiently discovering interesting patterns from large amounts of data
- A natural evolution of database technology, in great demand, with wide range of applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in 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