Data Mining

Dr. Muhammad Abulaish

Assoc. Professor, Dept. of Computer Science
South Asian University, New Delhi
Email: abulaish@sau.ac.in

Data Mining

- **Part One:** Intuitive Introduction and DM Overview
- **Part Two:** Textbook chapters
- **Part Three:** Students Presentations
- **Course Textbook:**

J. Han, M. Kamber
DATA MINING
Concepts and Techniques
Morgan Kaufmann, 2003/2006

Course Outline

Click here to see the course outline

Data

- Data is the Latin plural of datum
- Used to represent unprocessed facts and figures without any added interpretation or analysis.
- Generally associated with some entity and often viewed as the lowest level of abstraction from which information and knowledge are derived.
- Data may be unstructured, semi-structured, and structured
- Example: The price of petrol is Rs. 48 per liter

Information

- Information is interpreted (processed) data so that it has meaning for the user.
- "The price of petrol has risen from Rs. 64 to Rs. 69 per liter" – is information for a person who tracks petrol prices.
- Data becomes information when it is processed for some purpose and adds value for the recipient.
- A set of raw sales figures – Data
- Sales report (chart plotting, trend analysis) – Information

Knowledge

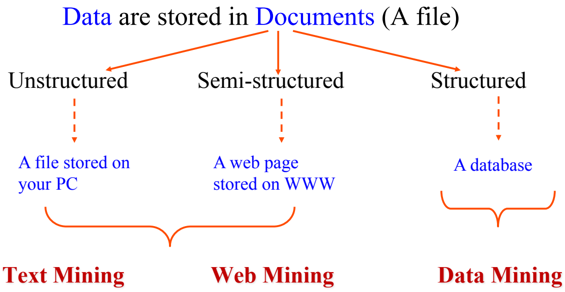
- Knowledge is a fluid mix of information, experience and insight that may benefit the individual or the organization.
- "When petrol prices go up by Rs. 5 per liter, it is likely that bus fare will rise by 10%" is knowledge.
- The boundaries between data, information, and knowledge is fuzzy
- What is data to one person is information to someone else.

Summarized View

- Data – as in databases
- Information – Processed data
- knowledge is a meta information about the patterns hidden in the data

The patterns must be discovered automatically

Data Mining, Text Mining and Web Mining



Data Mining Main Objectives

- Identification of data as a source of useful information
- Use of discovered information for competitive advantages when working in business enviroment

Why Data Mining?

- Data explosion problem
  - The Explosive Growth of Data: from terabytes to petabytes
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, datawarehouses and other information repositories

Why Data Mining? (cont…)

- Data explosion problem (cont…)
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation
    - Society and everyone: news, digital cameras,

Why Data Mining? (cont…)

- Data explosion problem (cont…)
- We are drowning in data, but starving for knowledge!
- **Solution:** Data warehousing and Data Mining
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

The Huber Taxonomy of Data Set Sizes

Descriptor	Data Set Size in Bytes	Storage Mode
Tiny	10^2	Piece of Paper
Small	10^4	A Few Pieces of Paper
Medium	10^6	A Floppy Disk
Large	10^8	Hard Disk
Huge	10^{10}	Multiple Hard Disks, e.g. RAID Storage
Massive	10^{12}	Robotic Magnetic Tape, Storage Silos

Algorithmic Complexity

Algorithm	Complexity
Plot a scatterplot	$O(n^{1/2})$
Calculate means, variances, kernel density estimates	O(n)
Calculate fast Fourier transforms	O(n log(n))
Calculate singular value decomposition of an rc matrix; solve a multiple linear regression	O(nc)
Solve most clustering algorithms	$O(n^2)$

No. of Operations for Algorithms of Various Computational Complexities and various Data Set Sizes

n	$n^{1/2}$	n	n log(n)	$n^{3/2}$	n^2
tiny	10	10^2	$2x10^2$	10^3	10^4
small	10^2	10^4	$4x10^4$	10^6	10^8
medium	10^3	10^6	$6x10^6$	10^9	10^{12}
large	10^4	10^8	$8x10^8$	10^{12}	10^{16}
huge	10^5	10^{10}	10^{11}	10^{15}	10^{20}

Computational Feasibility on a Pentium PC 10 MegaFLOPs Performance Assumed

n	$n^{1/2}$	n	n log(n)	$n^{3/2}$	n^2
tiny	10^{-6} seconds	10^{-5} seconds	$2x10^{-5}$ seconds	.0001 seconds	.001 seconds
small	10^{-5} seconds	.001 seconds	.004 seconds	.1 seconds	10 seconds
medium	.0001 seconds	.1 seconds	.6 seconds	1.67 minutes	1.16 days
large	.001 seconds	10 seconds	1.3 minutes	1.16 days	31.7 years
huge	.01 seconds	16.7 minutes	2.78 hours	3.17 years	317,000 years

Computational Feasibility on a Silican Graphics Onyx Workstation 300 MegaFLOPs Performance Assumed

n	$n^{1/2}$	n	n log(n)	$n^{3/2}$	n^2
tiny	$3.3x10^{-8}$ seconds	$3.3x10^{-7}$ seconds	$6.7x10^{-7}$ seconds	$3.3x10^{-6}$ seconds	$3.3x10^{-5}$ seconds
small	$3.3x10^{-7}$ seconds	$3.3x10^{-5}$ seconds	$1.3x10^{-4}$ seconds	$3.3x10^{-3}$ seconds	.33 seconds
medium	$3.3x10^{-6}$ seconds	$3.3x10^{-3}$ seconds	.02 seconds	3.3 seconds	55 minutes
large	$3.3x10^{-5}$ seconds	.33 seconds	2.7 seconds	55 minutes	1.04 years
huge	$3.3x10^{-4}$ seconds	33 seconds	5.5 minutes	38.2 days	10,464 years

Computational Feasibility on an Intel Paragon XP/S A4 4.2 GigaFLOPs Performance Assumed

n	$n^{1/2}$	n	n log(n)	$n^{3/2}$	n^2
tiny	$2.4x10^{-9}$ seconds	$2.4x10^{-8}$ seconds	$4.8x10^{-8}$ seconds	$2.4x10^{-7}$ seconds	$2.4x10^{-6}$ seconds
small	$2.4x10^{-8}$ seconds	$2.4x10^{-6}$ seconds	$9.5x10^{-6}$ seconds	$2.4x10^{-4}$ seconds	.024 seconds
medium	$2.4x10^{-7}$ seconds	$2.4x10^{-4}$ seconds	.0014 seconds	.24 seconds	4.0 minutes
large	$2.4x10^{-6}$ seconds	.024 seconds	.19 seconds	4.0 minutes	27.8 days
huge	$2.4x10^{-5}$ seconds	2.4 seconds	24 seconds	66.7 hours	761 years

Computational Feasibility on a TeraFLOP Grand Challenge Computer 1000 GigaFLOPs Performance Assumed

n	$n^{1/2}$	*n*	*n log(n)*	$n^{3/2}$	n^2
tiny	10^{-11} seconds	10^{-10} seconds	$2x10^{-10}$ seconds	10^{-9} seconds	10^{-8} seconds
small	10^{-10} seconds	10^{-8} seconds	$4x10^{-8}$ seconds	10^{-6} seconds	10^{-4} seconds
medium	10^{-9} seconds	10^{-6} seconds	$6x10^{-6}$ seconds	.001 seconds	1 second
large	10^{-8} seconds	10^{-4} seconds	$8x10^{-4}$ seconds	1 second	2.8 hours
huge	10^{-7} seconds	.01 seconds	.1 seconds	16.7 minutes	3.2 years

Types of Computers for Interactive Feasibility Response Time < 1 Second

n	$n^{1/2}$	*n*	*n log(n)*	$n^{3/2}$	n^2
tiny	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Personal Compute
small	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Super Compute
medium	Personal Computer	Personal Computer	Personal Computer	Super Computer	Teraflop Compute
large	Personal Computer	Workstation	Super Computer	Teraflop Computer	---
huge	Personal Computer	Super Computer	Teraflop Computer	---	---

Types of Computers for Feasibility Response Time < 1 Week

n	$n^{1/2}$	*n*	*n log(n)*	$n^{3/2}$	n^2
tiny	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Personal Compute
small	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Personal Compute
medium	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Personal Compute
large	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Teraflop Compute
huge	Personal Computer	Personal Computer	Personal Computer	Super Computer	---

DM: Intuitive Definition

- Process to extract previously unknown knowledge from large volumes of data
- Requires both new technologies and methods

Data Mining

- DM creates models (algorithms):
- Classification
- Clustering
- Association
- Prediction
- DM often presents the knowledge as a set of rules of the form:
  IF.... THEN...
- Finds other relationships in data
- Detects deviations

DM Some Applications

- Target marketing, customer relation management, market basket analysis, cross selling, market segmentation
- Forecasting, customer retention, quality control, competitive analysis

DM Other Applications

- **Other Applications**
  - Text mining (news group, email, documents) and Web analysis.
  - Intelligent query answering
  - Scientific Applications

DM: Business Advantages

- Data Mining uses gathered data to
- **Predicts** tendencies and waves
- **Classifies** new data
- Find previously unknown patterns
- Discover unknown relationships

DM: Technologies

- Many commercially available tools
- Many methods (models, algorithms) for the same task
- TOOLS ALONE ARE NOT THE SOLUTION
- The user must be able to interpret the results; one of the requirements of DM is:

"the results must be easily comprehensible to the user"

- Most often,especially when dealing with statistical methods analysts are needed to interpret the knowledge – weakness of statistical methods.

Data Mining vs Statistics

- Some statistical methods are considered as a part of Data Mining i.e. they are used as Data Mining algorithms, or as a part of Data Mining algorithms
- Some, like statistical prediction methods of different types of regression and clustering methods are now considered as an integral part of Data Mining research and applications

Bussiness Applications

- Buying patterns
- Fraud detection
- Decision support
- Medical aplications
- Marketing
- and more

Fraud Detection and Management (B1)

- **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 use data mining to help identify similar instances**

Fraud Detection and Management (B2)

- Examples
  - **auto insurance**: detect characteristics of group of people who stage accidents to collect on insurance
  - **money laundering**: detect characteristics of suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
  - **medical insurance**: detect characteristics of fraudulent patients and doctors

Fraud Detection and Management (B3)

- **Detecting inappropriate medical treatment**
  - Australian Health Insurance Commission detected that in many cases blanket screening tests were requested (save Australian $1m/yr).
- **Detecting telephone fraud**
  - DM builds telephone call model: destination of the call, duration, time of day or week. Detects 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.

Fraud Detection and Management (B4)

- **Retail**
  - Analysts used Data Mining techniques to estimate that 38% of retail shrink is due to dishonest employees
  - and more….

Data Mining vs Data Marketing

- Data Mining methods apply to many domains
- Applications of Data Mining methods in which the goal is to find buying patterns in Transactional Data Bases has been named: Data Marketing

Market Analysis and Management (MA1)

- Where are the data sources for analysis?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - DM finds clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.

Market Analysis and Management (MA2)

- Determine customer purchasing patterns over time
  - Conversion of single to a joint bank account: when marriage occurs, etc.
- Cross-market analysis
  - Associations/co-relations between product sales
  - Prediction based on the association information

Market Analysis and Management (MA3)

- Customer profiling
  - data mining can tell you what types of customers buy what products (clustering or classification)
- Identifying customer requirements
- identifying the best products for different customers

Corporate Analysis and Risk Management (CA1)

- 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

Corporate Analysis and Risk Management (CA2)

- Competition:
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

Scientific Applications

- Networks failure detection
- Controllers
- Geographic Information Systems
- Genome- Bioinformatics
- Intelligent robots
- etc… etc ….

Other 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
  - And more …..

What is NOT Data Mining

- Once the patterns are found Data Mining process is finished
- The use of the patterns is not Data Mining
- Queries to the database are not DM

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation

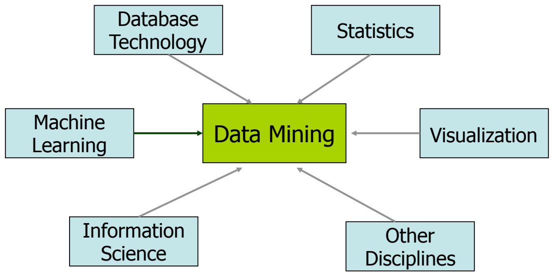
Evolution of Database Technology c.d.

- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s—2000s:
  - Data mining and data warehousing, multimedia databases, and Web databases

Short History of Data Mining

- **1989 - KDD term (Knowledge Discovery in Databases)** appears in (IJCAI Workshop)
- **1991 -** a collection of research papers edited by Piatetsky-Shapiro and Frawley
- **1993 – Association Rule Mining Algorithm APRIORI** proposed by Agrawal, Imielinski and Swami.
- **1996 – present:** KDD evolves as a conjuction of different knowledge areas (data bases, machine learning, statistics, artificial intelligence) and the term **Data Mining** becomes popular

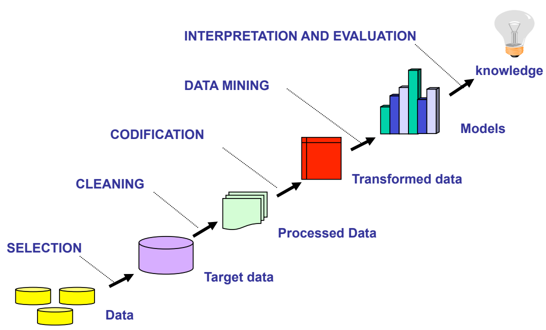
Data Mining: Confluence of Multiple Disciplines



KDD process: Definition [Piatetsky-Shapiro 97]

- KDD is a non trivial process for identification of :
  - Valid
  - New
  - Potentially useful, and
  - Understable patterns in data

The KDD process



Steps of the KDD process

- **Preprocessing:** includes all the operations that have to be performed before a data mining algorithm is applied
- **Data Mining:** knowledge discovery algorithms are applied in order to obtain the patterns
- Interpretation: discovered patterns are presented in a proper format and the user decides if it is neccesary to re-iterate the algorthms

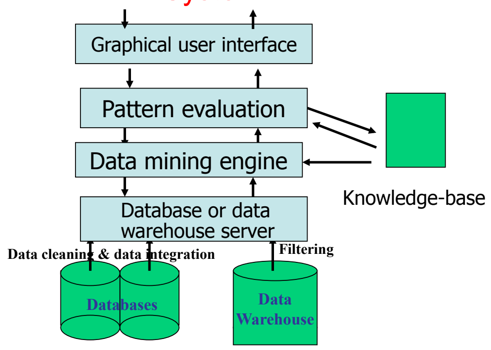
DM: Data Mining

- **DM** is a step of the KDD process in which algorithms are applied to look for patterns in data
- It is necessary to apply first the preprocessing operation to clean and preprocess the data in order to obtain significant patterns

KDD vs DM

- **KDD** is a term used by Academia
- **DM** is a commercial term
- DM term is also being used in Academia, as it has become a "brand name" for both KDD process and its DM sub-process
- The important point is to see Data Mining as a process

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 databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW

Concept, class, description

- **Concept** – is defined semantically as any subset of records.
- We often define the concept by attribute **c** and its value **v**
- **In this case the concept description** is syntactically written as : **c=v** and we define:
- ***CONCEPT={records: c=v}***
- For example: *climate=wet (description of the concept)*
- *CONCEPT={records: climate=wet}*
- **We use word: CLASS, class attribute for Concept, concept attribute**

Concept Characteristics

- ***Concept C characteristics*** *is a set of attributes a1, a2, … ak, and their respective values v1, v2, …. vk that are characteristic for a given concept c , i.e.*
- *{records: a1=v1 & a2=v2&…..ak=vk}*
- ***Characteristics description*** *is then syntactically written as*

  a1=v1 & a2=v2&…..ak=vk

Data Mining Functionalities Characterization (1)

- Describes the process which aim is to find rules that describe properties of a concept. They take the form

If *concept* ***then*** *characteristics*

- C=1 → A=1 & B=3 **25%** (support: there are 25% o the records for which the rule is true)
- C=1 → A=1 & B=4 **17%**
- C=1 → A=0 & B=2 **16%**

Discrimination (2)

- It is the process which aims is to find rules that allow us to discriminate the objects (records) belonging to a given concept (one class) from the rest of records (classes)

If characteristics then concept

- **A=0 & B=1 → C=1 33% 83% (support, confidence: the conditional probability of the concept given the characteristics)**
- **A=2 & B=0 → C=1 27% 80%**
- **A=1 & B=1 → C=1 12% 76%**
- Discriminant rule can be good even if it has a low support (and high confidence)

Classification and Prediction (3)

- **Classification and Prediction - Supervised learning**
  - Finding models (rules) that describe (characterize) or/ and distinguish (discriminate) classes or concepts for future prediction
  - **Example:** classify countries based on climate (characteristics), or classify cars based on gas mileage and use it to predict classification of a new car
  - **Presentation:** decision-tree, classification rules, neural network, Bayes Network

Data Mining Functionalities (4)

- **Prediction (statistical)**

  - predict some unknown or missing numerical values
- **Cluster analysis**
  - Class label is unknown: Group data to form new classes- **unsupervised learning**
  - For example: cluster houses to find distribution patterns
  - Clustering is based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity

Data Mining Functionalities (5)

- **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

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

Major Issues in Data Mining (2)

- 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 (3)

- 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