Data Mining Concepts, Techniques and Applications

Lecturer and Contact

- **Lead Lecturer: Dr Maria Indrawan, H7.32, Phone: 990 31916 email: maria.indrawan@infotech.monash.edu.au**
- **Other Lecturers are: Professor B Srinivasan and Dr Campbell Wilson.**

Unit Outline

- **Data Mining Concepts**
- **Data Mining Techniques**
- **Data Mining Applications**
- **Unit Guide:**
  - http://www.infotech.monash.edu.au/units/cse3212/
- **Unit Website:**
  - MUSO

Assessments

- Assignments
  - Assignment 1
    - 15%
    - Performing manual data mining.
    - Monday – in the lecture, week 7
  - Assignment 2
    - 15% :
    - Software assisted data mining
    - Monday – in the lecture, week 11
- Unit test
  - 10%
  - Monday – in the lecture, week 8
- Examination
  - 60%
  - 2 hours closed book exam.
- In order to pass the unit you need to get:
  - At least 40% of the total marks available in the exam.
  - At least 40% of the total of assignments and unit test.
  - At least 50% of the total marks in the unit.

Textbook

- **No prescribed text book.**
- **Recommended reading:**
  - Data Mining Tutorial Based Primer, Roiger R & Geatz M
  - Data Mining: Concepts and Techniques, Han J & Kamber M (2001 or 2006 editions).
- **The recommended books are available:**
  - For purchase from the bookshop
  - For loan in the library (reserve and short term loan)

Tutorials

- **Students do not need to register to any particular tutorial class.**
- **Tutors**
  - Mr. Samar Zutshi (samar.zutshi@infotech.monash.edu.au)
  - Mr. Manoj Kathpalia (manoj.kathpalia@infotech.monash.edu.au)
- **Tutorial classes:**
  - Monday 8-10 PM, Room: HB.36 (Samar)
  - Tuesday 10 AM – 12 PM, Room: HB.40 (Manoj)

Communication

- **Refer all enquiries regarding the administration of the unit (eg assignment extension, assessing MUSO, etc) to Maria Indrawan.**
- **Need clarification on the content?:**
  - Discussion board in MUSO.
    - The discussion board will be created based on each lecture topic.
    - Students are welcome to answer other students' question.
  - Helpdesk sessions
    - Room: H6.42, schedule: TBA, starts in week 3.
    - Schedule will be published in MUSO.
  - Contact the lecturer who delivers the lecture for that particular topic.

Semester Plan

Week	Lecture	Tutorial	Assessment
1	Introduction	No tutorial	
2	Data Mining Approaches	Data Mining Approaches	
3	Data Preparation	Data Preparation	
4	Association Mining	Association Mining	
5	Classification	Classification	
6	Decision Tree	Decision Tree	
7	Clustering	Clustering	Assignment 1
8	Unit Test	iDA / WEKA as data mining software	Unit Test
9	Neural Networks	Discussion on unit tests	
10	Self Organising Map	iDA/WEKA as data mining software	
11	Visualisation	Visualisation	Assignment 2
12	Web Mining	Web Mining	

Introduction

- **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**
- **Link to Data Warehousing**

Motivation: "Necessity is the Mother of Invention"

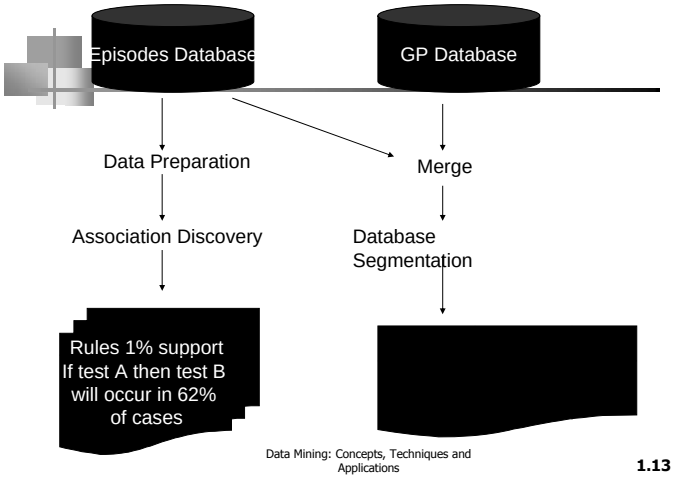
- **Data explosion problem**
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- **We are drowning in data, but starving for knowledge!**
- **Solution: Data warehousing and data mining**
  - Data warehousing and on-line analytical processing
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Evolution of Database Technology

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

What Is Data Mining?

- **Data mining (knowledge discovery in databases - KDD):**
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- **Alternative names and their "inside stories":**
  - Data mining: a misnomer?
  - 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



Why Data Mining? — Potential Applications

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

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: interest, 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 tell you what types of customers buy what products (clustering or classification)
- **Identifying customer requirements**
  - identifying the best products for different customers
  - use prediction to find what factors will attract new customers
- **Provides summary information**
  - various multidimensional 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 use data mining to help identify similar instances
- **Examples**
  - auto insurance: detect a group of people who stage accidents to collect on insurance
  - money laundering: detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
  - medical insurance: detect professional patients and ring of doctors and ring of references

Fraud Detection and Management (2)

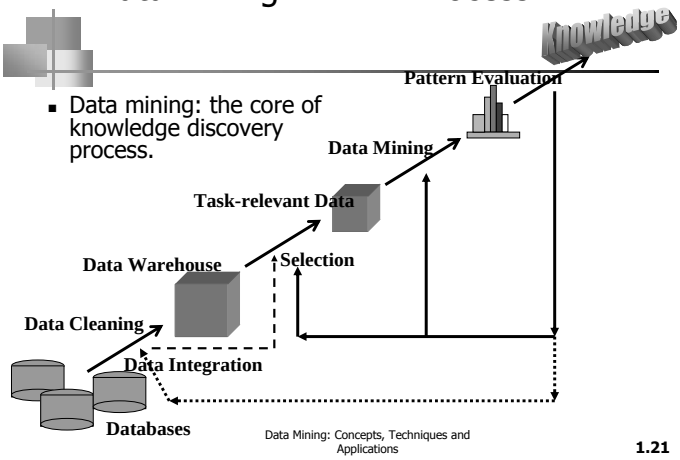
- **Detecting inappropriate medical treatment**
  - Health Insurance Commission identifies that in many cases blanket screening tests might have been requested (can save $$).
- **Detecting telephone fraud**
  - Telephone call model: destination of the call, duration, time of day or week. Analyze patterns 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 Applications

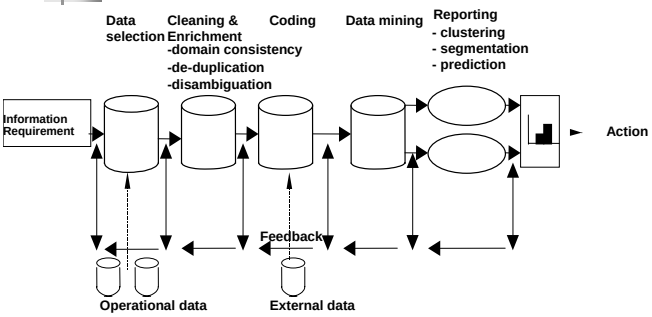
- **Sports**
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat
- **Astronomy**
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining
- **Internet Web Surf-Aid**
  - 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.

Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.



The Process of Knowledge Discovery

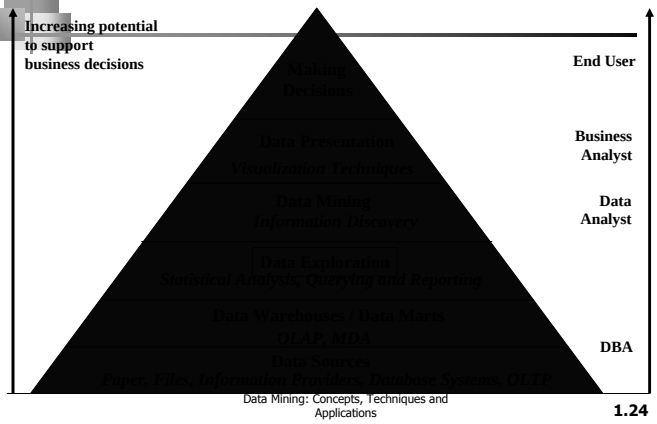


The Knowledge Discovery in Databases (KDD) process (Adriens/Zantinge)

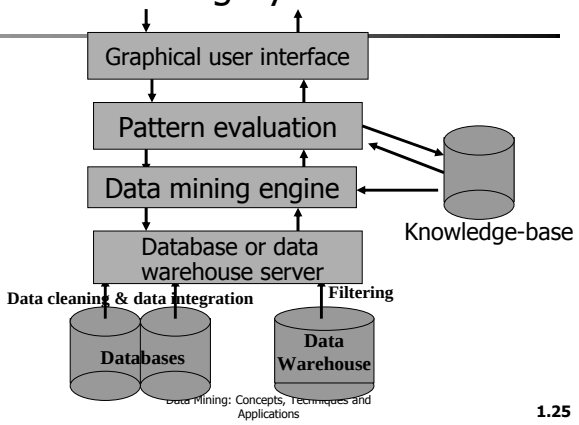
Steps of a KDD Process

- **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**

Data Mining and Business Intelligence



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

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") ^ income(X, "20..29K") → buys(X, "PC") [support = 2% ➔are 20..29 years old with an income of 20..29k, confidence = 60%➔that a customer in this age and income group will buy a PC]
  - contains(T, "computer") → contains(x, "software") [1%, 75%]
    computer ➔ software [1%, 75%]

An Association Rule

- **The** Rule
  - When a customer buys a shirt, in 70% of cases, he or she will also buy a tie
  - The **Confidence Factor** is 70%
- **The** Support Factor
  - This occurs in 13.5% of all purchases
  - The Support Factor is 13.5%
- **More formally**

Support and Confidence

- **Support:**
  - Percentage of transactions from a transaction database that the given rule satisfies.
  - This can be taken as the probability $P(X \cup Y)$ where $X \cup Y$ indicates that a transaction contains both X and Y, that is union of item sets X and Y.
- **Confidence:**
  - Which assess the degree of certainty of the detected association.
  - This can be taken as the conditional probability P(Y|X), that is, the probability that a transaction containing X also contains Y.
- **More formally**
  - Support $(X \Rightarrow Y)$ = $P(X \cup Y)$
  - Confidence $(X \Rightarrow Y)$ = $P(Y|X)$

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 petrol 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**

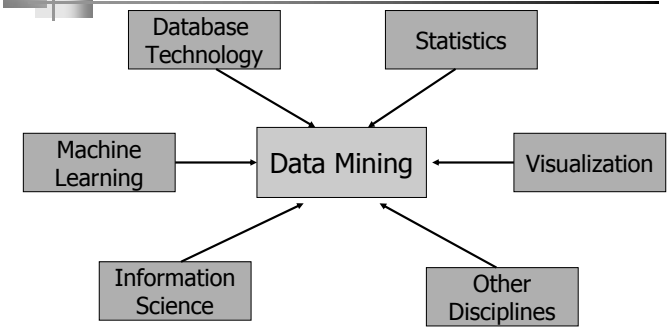
Are All the "Discovered" Patterns Interesting?

- **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 general all the patterns and then filter out the uninteresting ones.
    - Generate only the interesting patterns—mining query optimization

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, Weblog analysis, etc.

Data Mining and the Data Warehouse

- **Organizations realized that they had large amounts of data stored (especially of transactions) but it was not easily accessible**
- **The data warehouse provides a convenient data source for data mining. Some data cleaning has usually occurred. It exists independently of the operational systems**
  - Data is retrieved rather than updated
  - Indexed for efficient retrieval
  - Data will often cover 5 to 10 years
- **A data warehouse is not a pre-requisite for data mining**

Data Mining and OLAP

- **Online Analytic Processing (OLAP)**
- **Tools that allow a powerful and efficient representation of the data**
- **Makes use of a representation known as a cube**
- **A cube can be sliced and diced**
- **OLAP provide reporting with aggregation and summary information but does not reveal patterns, which is the purpose of data mining**

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

Summary

- **Data mining: discovering interesting patterns from large amounts of data**
- **A natural evolution of database technology, in great demand, with wide 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**

References

- **U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.**
- **J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2006.**
- **T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.**
- **G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.**
- **G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.**