Data Mining

Prof. Tahar Kechadi

UCD School of Computer Science
Insight Centre for Data Analytics
Science East, 3rd Floor, E3.42

Important Orientation Information

- **Applications for Extenuating Circumstances**
  - These refer to very grave issues that occasionally arise such as
    - Serious illness, hospitalisation, an accident
    - Family bereavement (parent, sibling)
    - Ongoing serious personal or emotional circumstances
  - Extenuating Circumstances **do not** cover events which are **foreseen** (e.g. 21st party, Debs ball, wedding etc.)
- **Minor Circumstances (absent for a few days)**
  - These situations should be handled locally by making direct contact with the lecturer/relevant School. Extenuating Circumstances do NOT apply in these cases

- **Missing a Lecture, Lab session or Tutorial**
  - …
- **Late submission of Coursework**
  - Where coursework is submitted late due to unanticipated exceptional or extenuating circumstances, students should follow procedures under the **Policy on Late Submission of Coursework**: http://www.ucd.ie/registry/academicsecretariat/docs/latesub_po.pdf
  - **An application for Extenuating Circumstances is not appropriate in this case.**

Plagiarism & UCD Computer Science

- **Plagiarism is a serious academic offence**
  - [Student Code, section 6.2] or [UCD Registry Plagiarism Policy] or [CS Plagiarism policy and procedures]
- Our staff and demonstrators are **proactive** in looking for possible plagiarism in all submitted work
- Suspected plagiarism is reported to the CS Plagiarism subcommittee for investigation
  - Usually includes an interview with student(s) involved
  - 1st offence: **usually** 0 or NG in the affected components
  - 2nd offence: referred to the **University disciplinary committee**
- Student who enables plagiarism is equally responsible

http://www.ucd.ie/registry/academicsecretariat/docs/plagiarism_po.pdf

http://www.ucd.ie/registry/academicsecretariat/docs/student_code.pdf

http://libguides.ucd.ie/academicintegrity

Organisation

- Lectures
  - Tuesday: 12:00 – 13:00
  - Thursday: 11:00 – 12:00
- Practical/Tutorial
  - Tuesday: 16:00 – 18:00
  - Course material: Blackboard (Virtual Learning Environment)
- Tutorials & Practical work
  - Tutorial and practical sessions will be organised every week. We will ask you to submit some of them.
  - Exercises will be posted on Blackboard prior to the tutorial/practical session
  - The work will be done as home work or in tutorial sessions depending on the topic covered or on the questions asked

Assessment

- Continuous Examination
  - Worth 40% of the overall mark
  - 3 MCQs of 30 minutes each
  - MCQ 1 : will take place on October 02rd (in Practical Session)
  - MCQ 2 : will take place on October 30th (in Practical Session)
  - MCQ 3 : will take place on November 20th (in Practical Session)
- Tutorial and Practical Work
  - Worth 20% of the overall mark
- Final Exam
  - Worth 40%
- Marking Scheme
  - The standard UCD Grading Scheme is used for this module.

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
- Major issues in data mining

References

- Data Mining: Concepts and Techniques
  - J. Han & M. Kamber, Morgan Kaufmann, 2nd Edition, 2006
- Introduction to Data Mining
  - P-N Tan, M. Steinbach & V. Kumar,Addison Wesley, 2006
- Data Mining: A Tutorial-Based Primer,
  - R. Roiger & M.W. Geatz, Addison Wesley, 2003
- Advances in Data Mining: Theoretical Aspects and Applications
  - Petra Pertner (Ed.), July 2007
- Data Mining: Concepts, Models, Methods, and Algorithms,
  - M. Kantardzic, Wiley-International, 2001
- Discovering Knowledge in Data: An introduction to Data Mining
  - D.T. Larose, Wiley, 2005

Why Data Mining?

- Data explosion problem
  - Data collection: Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- Advances in ICT
  - Proliferation of large storage devices and high communication networks
  - We are actually living in the data age
- We are drowning in data, but starving for knowledge!

What is Data Mining?

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

DM 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

- 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 or “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/correlations between product sales
  - Prediction based on the association information

Market Management

- Customer profiling
  - DM 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
- Providing summary information
  - various multidimensional summary reports
  - statistical summary information (data central tendency and variation)

Corporations

- 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

- Applications
  - widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.
- Approach
  - use historical data to build models of fraudulent behaviour 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

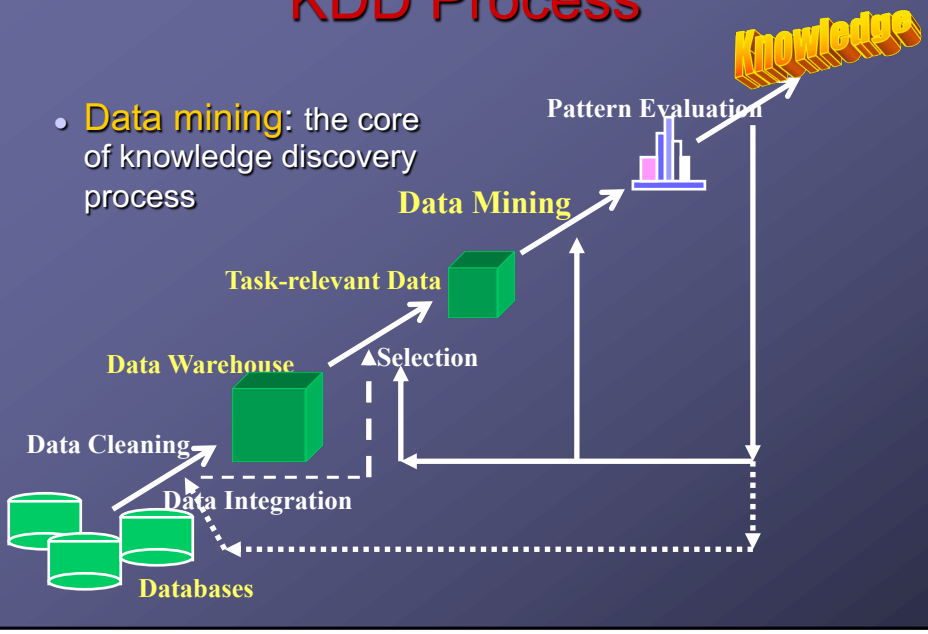
- Detecting inappropriate medical treatment
  - Australian Health Insurance Commission identifies that in many cases blanket screening tests were requested (save Australian €1m/yr)
- Detecting telephone fraud
  - Telephone call model: destination of the call, duration, time of day or week. Analyse patterns 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 analysed 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 behaviour pages, analysing effectiveness of Web marketing, improving Web site organisation, etc.

KDD Process

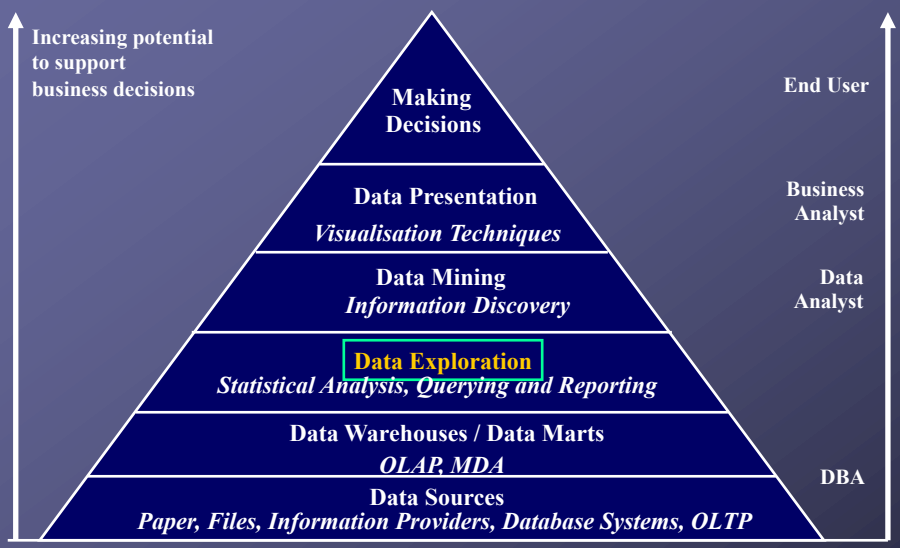
- Data mining: the core of knowledge discovery process



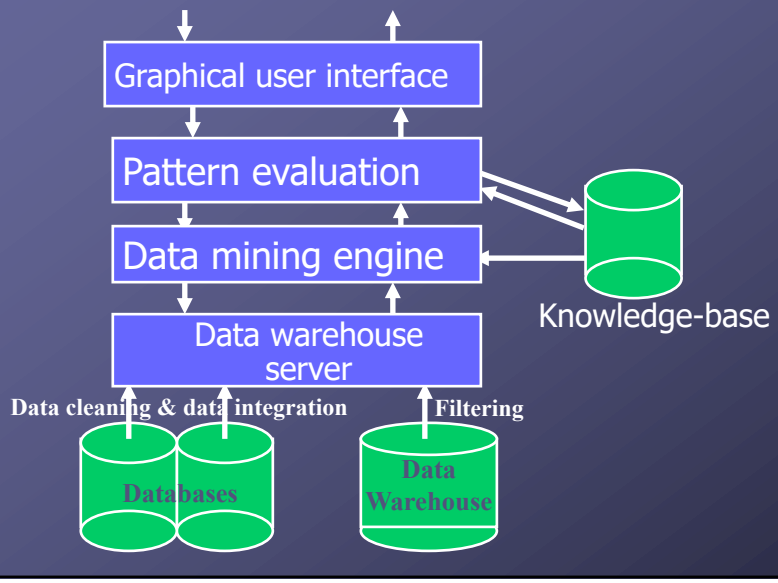
KDD Process

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

Business Intelligence



Architecture of DM System



DM Software Tools

- DM Software Tools
  - We can find several Data Mining tools
  - Commercial and free (open source) software tools
  - KXEN Modeler, IBM SPSS Modeler, Oracle Data Mining, Angoss KnowledgeSTUDIO, etc.
  - RapidMiner, Weka, KNIME, SCaViS, Tanagra, R, …
- Orange
  - Free version