CS4038/CS5012
Data Mining and Visualization

Wei Pang

Time table

- Lectures
  - 2 lectures
    - 11:00 -12:00 Tuesdays in Meston 6
    - 9:00 – 10:00 Fridays in Cruickshank G08
    - No lectures in Week 17 (5 November, 8 November)
- Practicals
  - 1 two hour practical on Mondays
  - 15:00-17:00, Meston 311

Assessment

- Course is worth 15 credits
- Two components
  - 25% continuous assessment
  - 75% end of term exam
- Continuous assessment
  - Issued in Week 6/7
  - Due on the Friday of Week 10/11

Lecturers

- Course Organiser: Dr. Wei Pang

  Introduction, Data, Clustering, Time Series, Anomaly Detection, Sequence Data

- Lecturer: Dr. Chenghua Lin

  Classsification, Associate Rule, Feature Selection, Visualisation, Case Study

Reading

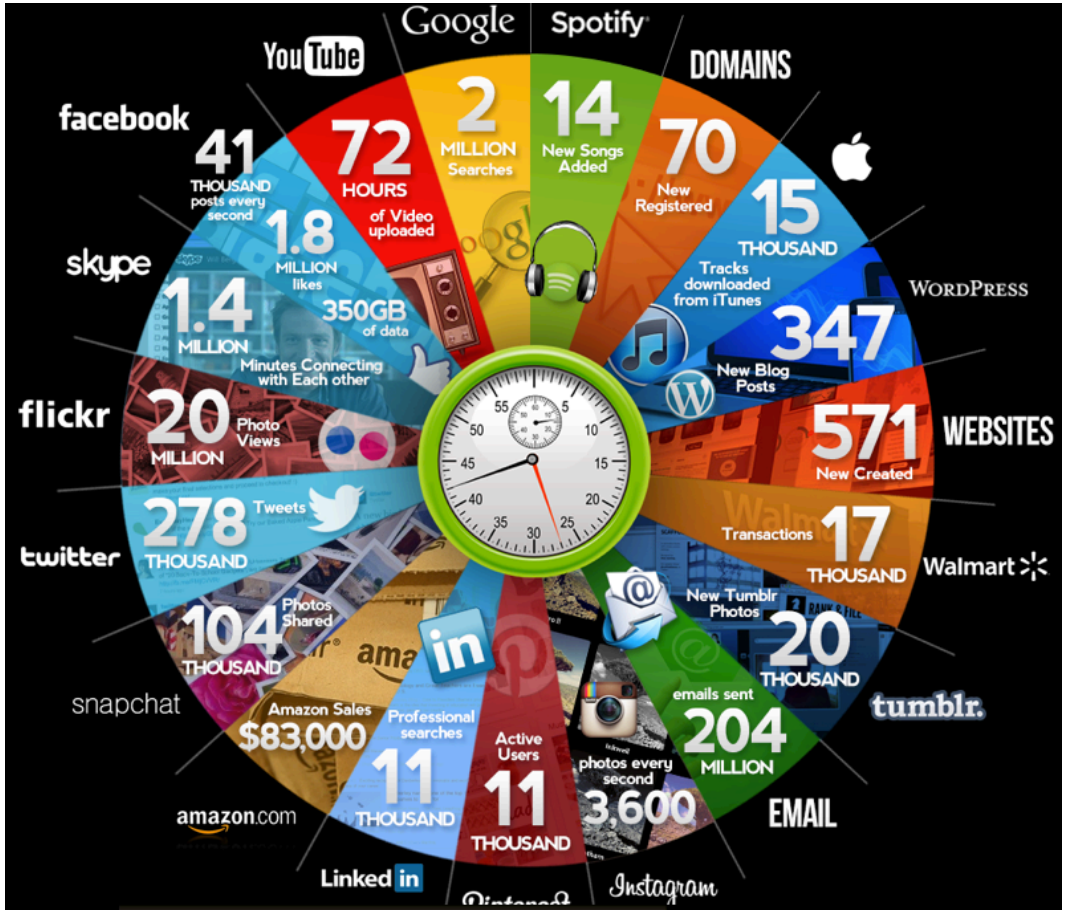
- Mostly lecture notes and some research papers
- Online reading material
- Textbook:
- Introduction to Data Mining by Tan, Steinbach, and Kumar. (free chapters online.)

Introduction

Overgrowth of Data

- Humans accumulate large volumes of data in many domains
  - Business
    - Transactional data
  - Scientific
    - Complete sequence data from Human Genome Project
      - of 3 billion DNA units
  - Engineering
    - 100s of sensors on a gas turbine taking measurements every second
  - And many more?

Why Data Mining



Picture from QMEE

Career Prospects

- Data Analyst
- Business Analyst
- Data Scientist
  - Junior: £35,000
  - Big Data Scientist Contract: £450-650 a day. (from *Indeed.co.uk, September 2013*)

Information Hidden in Data

- Data are raw facts
- Humans routinely 'dig' useful abstractions from raw data
  - An example abstraction 'mined' from past exam results
  - No coursework submitted => will fail the exam as well
- For small data sets (a few hundred bytes)
  - Simple and manual data analysis OK (Even preferred!!!)
  - Statistics
- For large data sets (a few Gigabytes or more)
  - Manual analysis is impossible
  - Computer Assistance needed

What is Data Mining, and What is not?

- **Process of automatically (or semi-automatically) discovering useful, novel and meaningful patterns from substantial quantities of data.**
  - Sorting a customer database based on customer ID number.
  - Computing the total sales of a company
  - Predicting the future profit of a company based on sales records from previous years.
  - Detect abnormal weather conditions based on historical weather information.

Data Mining Tasks

- Predictive Tasks
  - Use some variables to predict unknown or future values of other variables.
- Descriptive Tasks
  - Find human-interpretable patterns that describe the data.

DM Task Examples

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]

Classification

- Descriptive: distinguish objects from different classes
- Predictive: predict the class label of previously unseen records.

Examples of Classification

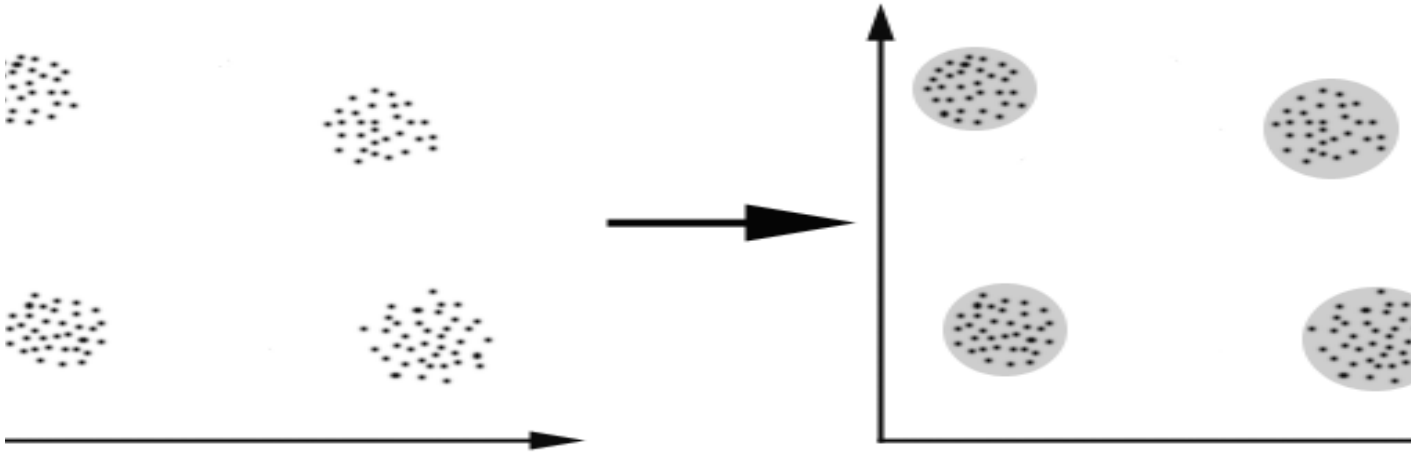
- Classify a newly found species as mammal, reptile, fish, or bird based on the attributes:
  - Skin Cover
  - Body Temperature
  - Hibernate
  - ……

Any other example?

Clustering

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustration of Clustering

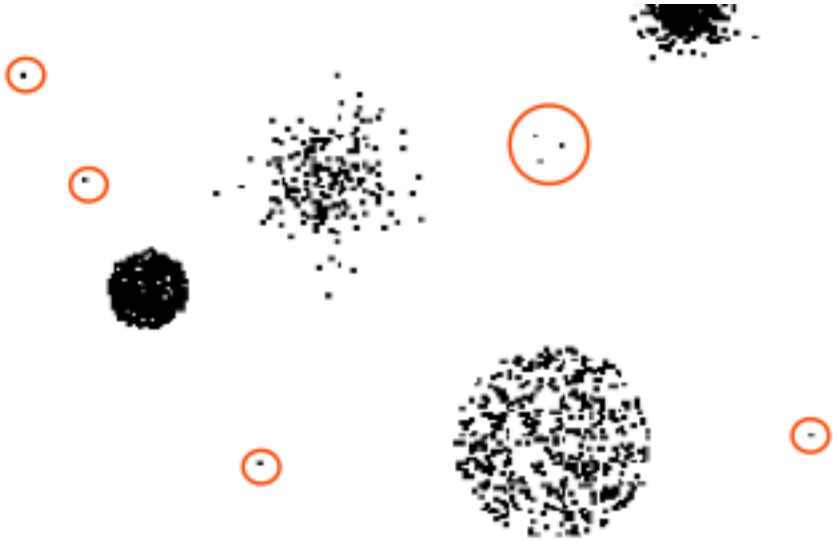


Associate Rule Mining

The Story of Beers and Diapers

Anomaly Detection

Fraud Detection
Intrusion Detection



Two views of computer assistance

- Data Mining View
  - Machines can automatically (or semi-automatically) extract meaningful and useful information from heaps of raw data
- Information Visualization (InfoVis) View
  - Humans themselves can make sense of data if data are presented visually
- We learn both these views in this course

Typical Applications of DM

- Customer Relationship Management (CRM)
- Linking gene variations among individuals to common illnesses (e.g. Cancer)
- Identifying abnormal conditions in an operational gas turbine
- More ?

Information Visualization

- Process of representing data in such a way (usually involves visual presentations) that enable users to gain useful insights into the data
- Focus is on designing a data representation scheme that makes in underlying 'information' visible to the user
- For rendering the representation scheme
  - Computer graphics technology is exploited
- Good InfoVis techniques are based on
  - Good understanding of the information structures underlying the data
  - Good understanding of the human perception and cognition
  - Good graphics

Summary

- All modern organizations
  - possess large volumes of data and
  - Users want to understand these data
- You learn technologies to
  - Extract and/Or present information from large data sets
    - Analytical methods
    - Visualization methods

Next Lecture

- What is Data?
- Chapter 2 of the Kumar Book
  - Introduction to Data Mining

Acknowledgement

- Some of the slides are based on the course slides provided by
  - Tan, Steinbach and Kumar (Introduction to Data Mining)

- Some pictures are taken from various online resources.