Introduction

Instructor: Jen-Wei Huang

Office: 92528 in the EE building
jwhuang@mail.ncku

What is Data Mining

- Knowledge Discovery from Data
- Extraction of interesting, non-trivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amount of data

Why Data Mining

- The explosive growth of data
- Data collection and data availability
- Many abundant data
  - Web, transactions, stocks, sensor networks, bioinformatics, scientific simulation, news, digital cameras, …
- We are drowning in data, but starving for knowledge!

Data Mining Processes

Data collection Data cleaning Data integration	Preprocessing
Data warehousing Data selection	Exploration
Pattern evaluation Knowledge discovery	Data Mining
Information presentation Decision making	Post processing

What Kinds of Data

- Database-oriented data sets
  - Relational database, data warehouse
- Advanced data sets
  - Data streams
  - Time-series data, sequence data, bio-sequences
  - Graphs, social networks and multi-linked data
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text
  - The World-Wide Web

Applications of Data Mining

- Marketing
- Web page analysis
- Collaborative analysis
- Recommender systems
- Biological and medical data analysis
- Software engineering
- Other dedicated knowledge discovery

Performance Measurement

- Efficiency
- Effectiveness (interestingness)
  - Objective measures; based on statistics & structures of patterns
    e.g. support, confidence
  - Subjective: based on user's beliefs in data
    e.g. unexpectedness, novelty

Interestingness

- A pattern is interesting if it is
  - Easily understood by humans
  - Valid on new or test data with some degree of certainty
  - Potentially useful
  - Validates some hypothesis that a user seeks to confirm

Techniques to Be Utilized

- Database
- Machine learning
- Neural network
- Fuzzy set
- Statistics
- Visualization
- Domain knowledge

Features & Challenges of KDD

- Handling of different types of data
- Efficiency & scalability of data mining algorithm
- Usefulness, certainly & expressiveness of results
- Interactive mining at multiple abstraction levels
- Parallel & distributed data mining
- Domain specific data mining
- Protection of privacy & data security

Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
- 1991-1994 Workshops on Knowledge Discovery in Databases
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
- Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

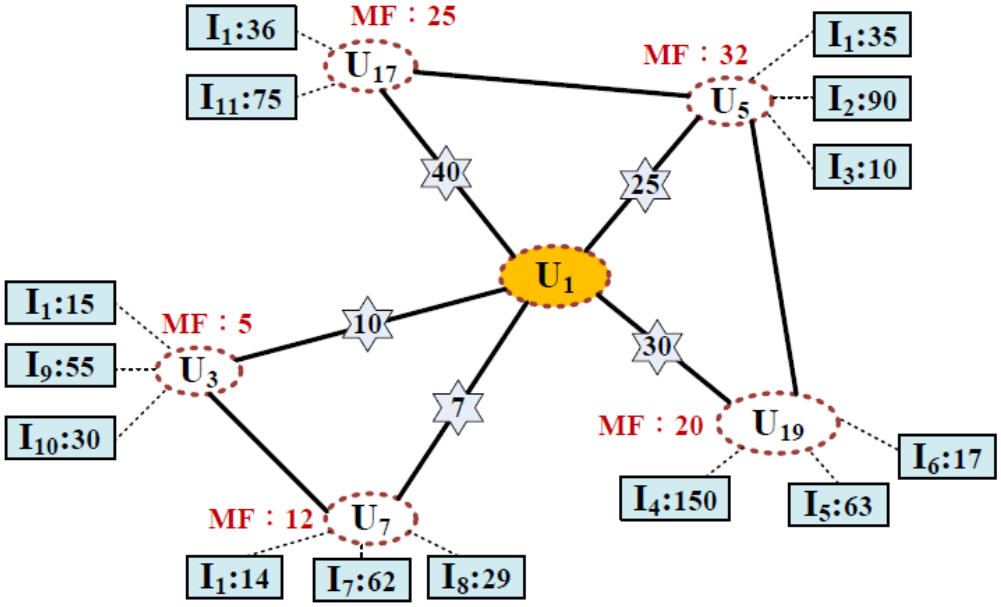
Social Network

- Social network is a concept to represent relationships between people
- It can be modeled as a graph via nodes, which represent individuals, and edges, which represent relationships

http://www.facebook.com/

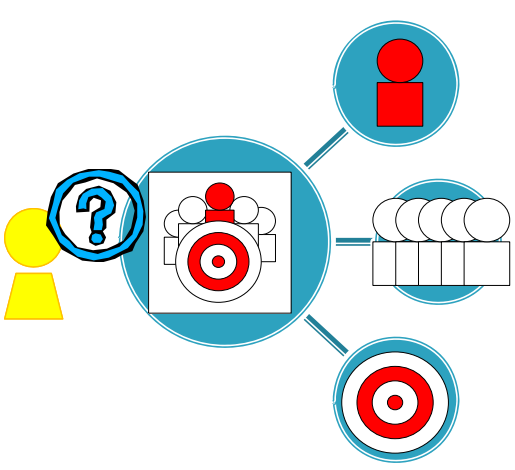
http://twitter.com/

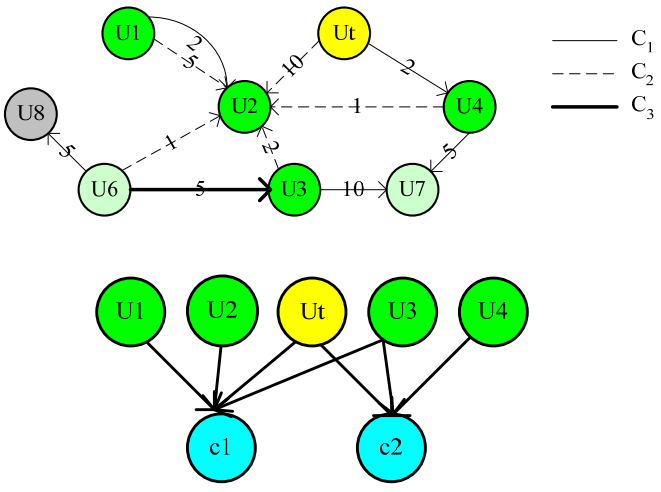
Example Social Network



Social Network Analysis

- SNA is to systematically understand the network structure and the user behaviors.





SNA Challenges

- Collect social network data
- Model the social behavior
- Calculate properties of the network
- Find interesting information
- Present and explain the knowledge

Reference Books

- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2nd ed., 2006
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005

Reference Books

- J. Scott, Social Network Analysis, SAGE Publications Ltd, 3rd ed., 2012
- J. Scott, Social Network Analysis: A Handbook, SAGE Publications Ltd, 2nd ed., 2000
- S. Chakrabarti, Mining the Web – Discovering Knowledge from Hypertext Data, Morgan Kaufmann Publisher, 2002
- M.E.J. Newman, Networks – An introduction, Oxford University Press, 2010

References

- Slides from Prof. J.-W. Han, UIUC
- Slides from Prof. W.-Z. Peng, NCTU
- Slides from Prof. S.-D. Lin, NTU

How Search Works

- http://www.google.com/insidesearch/howsearchworks/thestory/index.html

Facebook Love Love Report

- http://www.inside.com.tw/2014/02/17/the-formation-of-love

How I Teach Kids to Love Science

- Cesar Harada on TED Talks
- https://www.youtube.com/watch?v=jAemh_JxgOk