

INFS4203/INFS7203
Data Mining

Lecture Notes 1: Introduction to Data Mining & Data Issues

Dr Xue Li
 University of Queensland, Brisbane Australia
<http://www.itee.uq.edu.au/~dke>
 xueli@itee.uq.edu.au

1

Instructors

- Course Coordinator: Assoc Prof Xue Li
 Phone: 3365 2379
 Email: xueli@itee.uq.edu.au
 Room: 78-650
 Consultation: Thursday 12-1pm
- Lecturer: Dr Heng Tao Shen
 Phone: 3365 8359
 Email: h.shen@uq.edu.au
 Room: 78-651
 Consultation: TBA

Tutor

- Currently... no tutor yet... ☹️
 - According to the school policy, I need to have at least 25 students in order to have a tutor... But now... ☹️

Text Book and NewsGroup

- Text Book:
 - Pang-Ning Tan, Michael Steinbach and Vipin Kumar. Introduction to Data Mining. 1st Edition. 2006.
- Newsgroup of INFS4203/INFS7203:
 - On My-UQ Website
 - Use it for the intra-class discussions for the course-related matters.

Assessment

Assessment Task	Due Date	Weighting
Exam - during Exam Period (School) Final Examination	Examination Period	60%
Work-based Assessment Individual Assignmnets	21 Aug 09 - 8 Oct 09 Assignments on Weeks 4, 7, 10, and 13	20% (5% x 4 assignments)
Exam - Mid Semester During Class Middle Semester Exam	17 Sep 09 14:00 - 17 Sep 09 15:40 Non-programmable calculator is required.	20%

5

Teaching Schedule

Week 1	Introduction to Data Mining and Data Issues (Lecture): Readings/Ref: Required Text ; Lecture Notes ;
Week 2	Association Rules Mining (Lecture): Readings/Ref: Required Text ; Lecture Notes ;
Weeks 3-4	Classification (Lecture): Readings/Ref: Required Text ; Lecture Notes ;
Weeks 5-6	Clustering (Lecture): Readings/Ref: Required Text ; Lecture Notes ;
Week 7	Revision of Previous Topics (Self Directed Learning): Read the materials that are related to the middle semester examination. Readings/Ref: Required Text ; Lecture Notes ; Reference Texts ; Reference Texts
Week 8	Middle Semester Exam (Progressive Exam): 1:30 Hrs Middle Semester Exam to be held during the lecture time. Readings/Ref: Required Text ; Lecture Notes ;
Weeks 9-10	Advanced Topic I -- Text and Web Mining (Lecture): Readings/Ref: Required Text ; Lecture Notes ;
Week 11	Advanced Topic II -- Time Series Mining (Lecture): Readings/Ref: Required Text ; Lecture Notes ; Reference Texts
Week 12	Revision of Previous Topics (Self Directed Learning): Read the materials that are related to the middle semester examination. Readings/Ref: Required Text ; Lecture Notes ; Reference Texts ; Reference Texts
Week 13	Course Revision (Lecture): Readings/Ref: Required Text ; Lecture Notes ;

6

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

INFS4203 / INFS7203 Data Mining

7

Necessity Is the Mother of Invention

- Data explosion problem
 - Automated data collection tools and mature database technology lead to tremendous amounts of data accumulated and/or to be analyzed 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
 - Mining interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

INFS4203 / INFS7203 Data Mining

8

Data Mining: How Big is the Data Set? (1)

- It is already a fact of life that data is/will be produced faster than what we can effectively process.
- In 24 hours:
 - AT&T records 275 million phone calls.
 - Google handles 100 million searches.
 - Wal-Mart records 20 million sales transactions.
- In a Second:
 - NASA's Space Shuttle operation will have 20,000 sensors telemetered once per second to Mission Control at Johnson Space Centre, Huston.

INFS4203 / INFS7203 Data Mining

9

Data Mining: How Big is the Data Set? (2)

- In a Second:
 - In United States there are about 50,000 security trading and up to 100,000 quotes and trades (ticks) are generated every second.
- In a Week:
 - In Australia there are more than 80 Million SMS messages sent a week.
- In all time:
 - In scientific data collections, such as astronomical observatories, satellites imaging, and earth sensing, data can be routinely collected in gigabytes every day.

INFS4203 / INFS7203 Data Mining

10

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.)
 - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
 - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
 - Stream data management and mining
 - Data mining with a variety of applications
 - Web technology and global information systems

INFS4203 / INFS7203 Data Mining

11

What Is Data Mining?

- 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
 - Data mining: a misnomer?
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
 - (Deductive) query processing.
 - Expert systems or small ML/statistical programs



INFS4203 / INFS7203 Data Mining

12

Why Data Mining?—Potential Applications

- Data analysis and decision support
 - Market analysis and management
 - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
 - Risk analysis and management
 - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
 - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
 - Text mining (news group, email, documents) and Web mining
 - Stream data mining
 - DNA and bio-data analysis

INFS4203 / INFS7203 Data Mining

13

Market Analysis and Management

- Where does the data come from?
 - 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
- Cross-market analysis
 - Associations/co-relations between product sales, & prediction based on such association
- Customer profiling
 - What types of customers buy what products (clustering or classification)
- Customer requirement analysis
 - identifying the best products for different customers
 - predict what factors will attract new customers
- Provision of summary information
 - multidimensional summary reports
 - statistical summary information (data central tendency and variation)

INFS4203 / INFS7203 Data Mining

14

Corporate Analysis & 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

INFS4203 / INFS7203 Data Mining

15

Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
 - Auto insurance: ring of **collusions**
 - Money laundering: suspicious monetary transactions
 - Medical insurance
 - Professional patients, ring of doctors, and ring of references
 - Unnecessary or correlated screening tests
 - Telecommunications: phone-call **fault detection**
 - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
 - Retail industry
 - Analysts estimate that 38% of retail shrink is due to dishonest employees
 - Anti-terrorism

INFS4203 / INFS7203 Data Mining

16

Other Applications

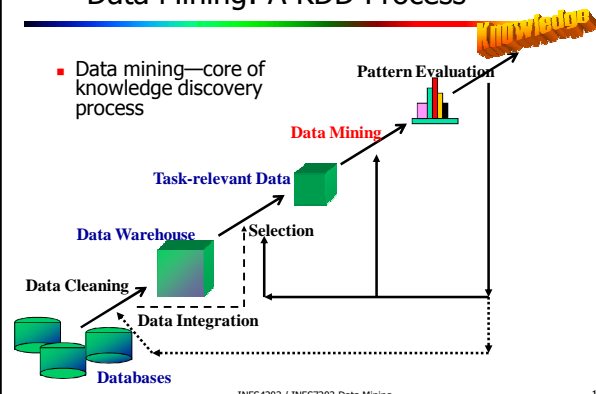
- Sports
 - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantages.
- Astronomy
 - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining.
- Internet Web Surf-Aid
 - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

INFS4203 / INFS7203 Data Mining

17

Data Mining: A KDD Process

- Data mining—core of knowledge discovery process



INFS4203 / INFS7203 Data Mining

18

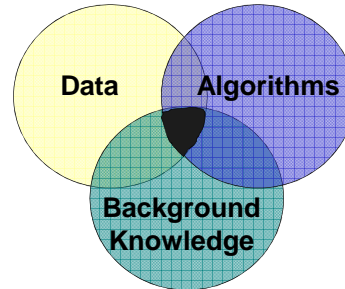
Steps of a KDD Process

- 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

INFS4203 / INFS7203 Data Mining

19

Data Mining Perspectives



INFS4203 / INFS7203 Data Mining

20

First of All: What is Data?

- A data item has two levels meaning: the **domain** and its **value**.
 - A data domain gives data structure and prescribe its possible (legal) values.
 - A data domain is associated with its domain-specific operations. For example, an *integer* is associated with arithmetic operations and a *text string* is associated with concatenation, sub-string, character padding and counting operations, etc.
 - A data value is a measurement of a real-world object or a concept.
- A data item can be either simple or complex.
 - A data item is associated to an ontology hierarchy.
 - A data item is associated to a multidimensional structure.

INFS4203 / INFS7203 Data Mining

21

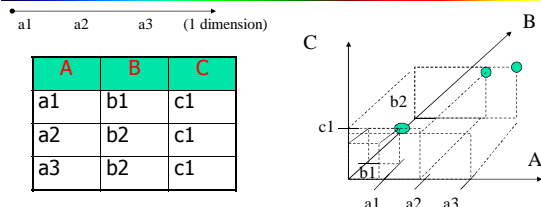
First of All: What is Data? (con)

- **Associated Patterns**: dependency, 1:m, m:n, 1:1, associations, correlations, dimensionality, etc.
- **Associated Dynamics (changes)**: monotonous changes, state transitions, etc.

INFS4203 / INFS7203 Data Mining

22

Multidimensional Data

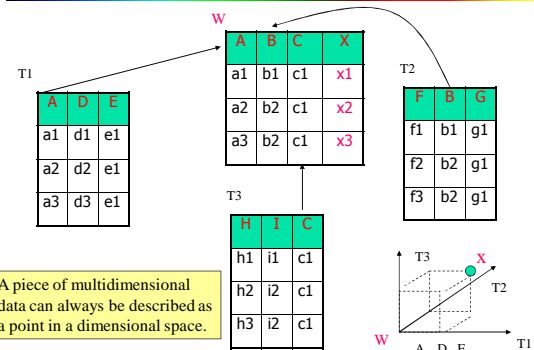


Any data record can be viewed as a point in a high dimensional data space.

INFS4203 / INFS7203 Data Mining

23

What is Multidimensional Data? – from a Relational Database Perspective



A piece of multidimensional data can always be described as a point in a dimensional space.

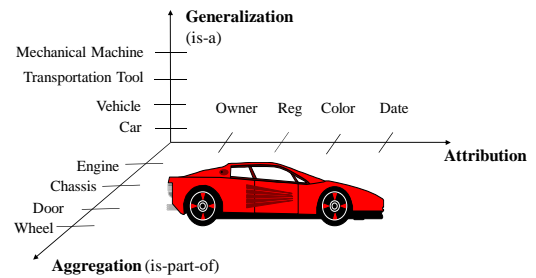
INFS4203 / INFS7203 Data Mining

24

So, for Multidimensional Data

- Each dimension is described by a **set of attributes**.
 - Each **attribute** has its unique **semantics** (different domains).
 - Each **dimension** is **structured** (different concept lattices, e.g., *is-a*, *is-part-of*, etc).
 - All dimensions are **associated** (for identifying a data item – “a container of data”).

Example: “A multidimensional car”



How are the Dimensionality associated to each other? (1)

Example of a Concept Lattice

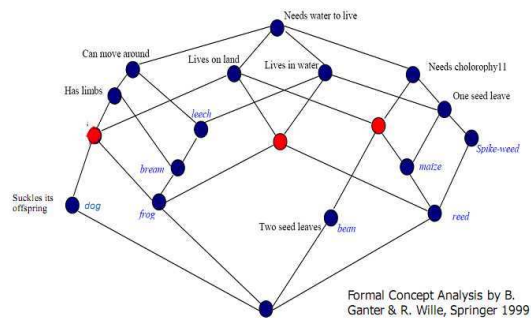
A high-dimensional data set for an educational film “Living Beings and Water”.

Creature	Needs water to live	Lives in water	Lives on land	Needs chlorophyll to produce food	Two good leaves	One good leave	Can move around	Has limbs	Suckles its offspring
Leech	X	X					X		
Bream	X	X					X	X	
Frog	X	X	X				X	X	
Dog	X		X				X	X	X
Spike-weed		X		X		X			
Wheat	X	X	X	X		X			
Bean	X		X	X	X				
Maize	X		X	X	X				

This high-dimensional data set can be represented by a lattice structure according to their subsumptions of their attributes.

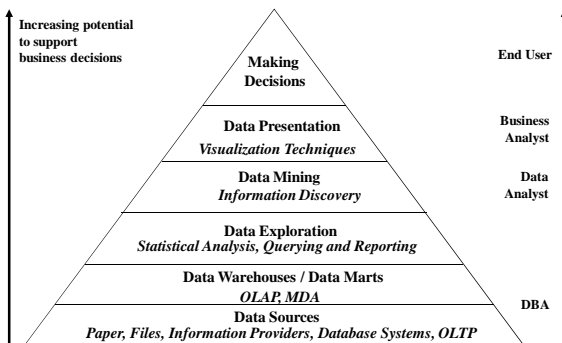
Formal Concept Analysis by B. Ganter & R. Wille, Springer 1999

How are the Dimensionality associated to each other? (2)

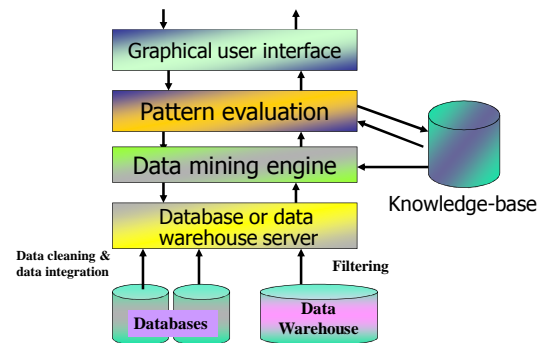


Formal Concept Analysis by B. Ganter & R. Wille, Springer 1999

Data Mining and Business Intelligence



Architecture: Typical Data Mining System



Data Mining: On What Kinds of Data?

- Relational database
- Data warehouse
- Transactional database
- Advanced database and information repository
 - Object-relational database
 - Spatial and temporal data
 - Time-series data
 - Stream data
 - Multimedia database
 - Heterogeneous and legacy database
 - Text databases & WWW

Data Mining Functionalities

- Concept description: Characterization and discrimination
 - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Association (correlation and causality)
 - Diaper → Beer [0.5%, 75%]
- Classification and Prediction
 - Construct models (functions) that describe and distinguish classes or concepts for future prediction
 - E.g., classify countries based on climate, or classify cars based on gas mileage
 - Presentation: decision-tree, classification rule, neural network
 - Predict some unknown or missing numerical values

Data Mining Functionalities (2)

- Cluster analysis
 - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
 - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
 - Outlier: a data object that does not comply with the general behavior of the data
 - Noise or exception? No! 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

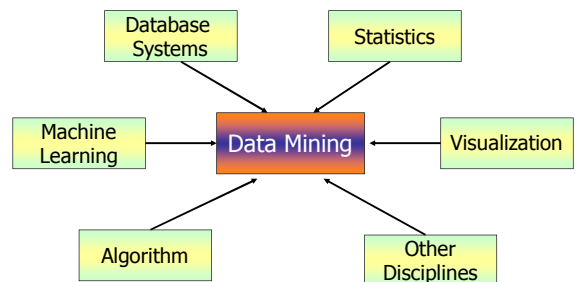
Are All the "Discovered" Patterns Interesting?

- Data mining 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?
 - Heuristic vs. exhaustive search
 - Association vs. classification vs. clustering
- Search for **only interesting patterns**: **An optimization problem**
 - Can a data mining system find **only** the interesting patterns?
 - Approaches
 - First generate all the patterns and then filter out the uninteresting ones.
 - Generate only the interesting patterns—mining query optimization

Data Mining: Confluence of Multiple Disciplines



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.
- Data mining systems and architectures
- Major issues in data mining

INFS4203 / INFS7203 Data Mining

37

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
 - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
 - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
 - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, and SIGKDD Explorations
- More conferences on data mining
 - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.

INFS4203 / INFS7203 Data Mining

38

Where to Find References?

- Data mining and KDD (SIGKDD: CDROM)
 - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
 - Journal: Data Mining and Knowledge Discovery, KDD Explorations
- Database systems (SIGMOD: CD ROM)
 - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
 - Journals: ACM-TODS, IEEE-TKDE, JIIS, J. ACM, etc.
- AI & Machine Learning
 - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), etc.
 - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics
 - Conferences: Joint Stat. Meeting, etc.
 - Journals: Annals of statistics, etc.
- Visualization
 - Conference proceedings: CHI, ACM-SIGGraph, etc.
 - Journals: IEEE Trans. visualization and computer graphics, etc.

INFS4203 / INFS7203 Data Mining

39

Recommended Reference Books

- R. Agrawal, J. Han, and H. Mannila, Readings in Data Mining: A Database Perspective, Morgan Kaufmann (in preparation)
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- S. M. Weiss and N. Indurkha, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2001

INFS4203 / INFS7203 Data Mining

40

Next Week:

Mining Association Rules

INFS4203 / INFS7203 Data Mining

41