Data Mining: Foundation, Techniques and Applications

Lesson 1a: Introduction

Li Cuiping(李翠平)
School of Information
Renmin University of China

Anthony Tung(鄧锦浩)
School of Computing
National University of Singapore

Main objectives of this course:

- Data mining is a diverse field which draw its foundation from many research areas like databases, machine learning, AI, statistic etc. The aim of this course is to highlight concepts from these areas which are fundamental and often used in building data mining tools.
- At the end of the course, students should:
  - Have a good knowledge of the fundamental concepts that provide the foundation of data mining.
  - Understand how these concepts are engineered to provide some of the basic data mining tools.
  - Able to adopt these concepts to develop new data mining tools for new application

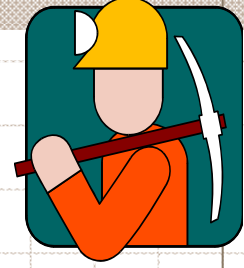
Secondary objective

- Indirectly introduce to students some basic principles of doing research including:
  - critical thinking: judging work through science rather than by superstition (eg. believing that work by foreign, famous researchers is always good and correct)
  - thoughts organization: how do research idea come about?
- Train up talents who can eventually run their own data mining courses:
  - 本地人办本地事, 本地辣姜多的是
  - It does not make sense if we are still needed to run the course after say 5 years because the course essentially does not help to bring out new talents

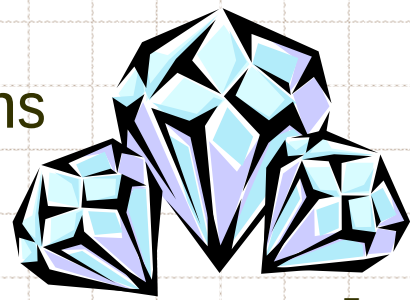
Necessity Is the Mother of Invention

- 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

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



Data Mining: A Generalized Framework

Techniques

- Association rules discovery
- Sequential Pattern Discovery
- Cluster analysis
- Outlier Detection
- Classifier Building
- Data Cube/Data Warehouse Construction
- Visualization ...

Applications

- Customer Relationship Management (CRM)
- Web pages Searches and Analysis
- Network Security
- Geographical Data Analysis
- Genomic Database ...

Principles

- **Database Technology:**
  - **Indexing, Compression, Data Structure**
- **AI / Machine Learning**
- **Statistics**
- **Information Theory**
- **Theoretical CS :**
  - **Approximate, Random, Online Algorithms**
- **Mathematical Programming**
- **Computational Geometry ...**

Applications

Applications (CRM)

- Customer Relationship Management deal with an important aspect of the new economic, personalization.
- Given a customer, can we recommend the correct products to him/her without having him/her asking about it ?
- Can we anticipate his/her need ?
- Can we offer a personalized discount package to each customer ?

Applications (Web Analysis)

- Web pages:
  - search: the web search engine Google is developed through data mining research in Stanford
  - Web pages clustering: finding similar pages in the web
- Web log
  - if customer visit page A and page B, then he/she is likely to go to page C and then buy product E
  - sequences clustering, finding group of customers who have very similar page visit sequences

Applications (Security)

- Network Security
  - used to detect network intrusion
  - analyze the commands issued
  - analyze the flow of network traffic
- Security Camera
  - analyze abnormal movement in a room or in a air plane
  - fast detection of weapons during X-ray scan

Applications (Geographical Data)

- Data that are presented on a map
- Weather prediction
- Finding pollution sources
- Analyze crime patterns
- Location planning
- Traffic analysis

Application (Performance Optimization)

- cache prefetching
- semantic compression using data mining
- better indexes through clustering
- supply-chain management

Applications (Bioinformatics)

- DNA sequences analysis:
  - indexing
  - clustering
  - compression
- Gene expression analysis
  - function prediction
  - visualization
  - clustering

Applications (Others)

- Sports
  - analyze the strategy of players and teams
- 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.

Application Affect The 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

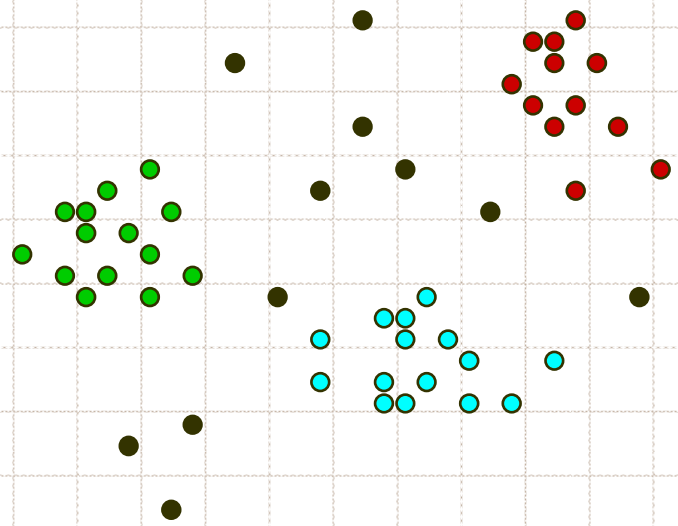
Techniques

Techniques(Association Rules Discovery)

- Association rule mining:
  - Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.
- Applications:
  - Basket data analysis, cross-marketing, catalog design, loss-leader analysis, clustering, classification, etc.
- Examples.
  - Rule form: "Body $\rightarrow$ Head [support, confidence]".
  - buys(x, "diapers") $\rightarrow$ buys(x, "beers") [0.5%, 60%]
  - major(x, "CS") ^ takes(x, "DB") $\rightarrow$ grade(x, "A") [1%, 75%]

Techniques (Cluster Analysis I)

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters

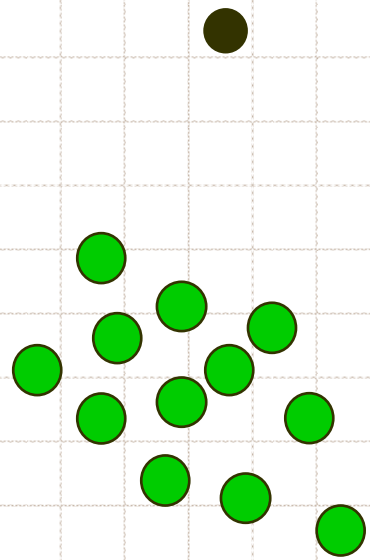


Techniques (Cluster Analysis II)

- Issues to be consider in clustering include
  - Types of distance functions
  - Objective measures
  - Handling high dimensionalities
  - Scalability
  - Selecting relevant dimensions ….
- Clustering had been studied in established field like statistic but continue to be an important research topic in data mining as new applications and data types emerge.

Techniques (Outlier Detection)

- What are outliers?
  - The set of objects are considerably dissimilar from the remainder of the data
  - Example: Sports: Michael Jordon, ...
- Naïve way to find outliers: cluster objects and objects that are very far from the clusters are outliers
- Challenge is to find outliers without clustering



Techniques (Classifier Building)

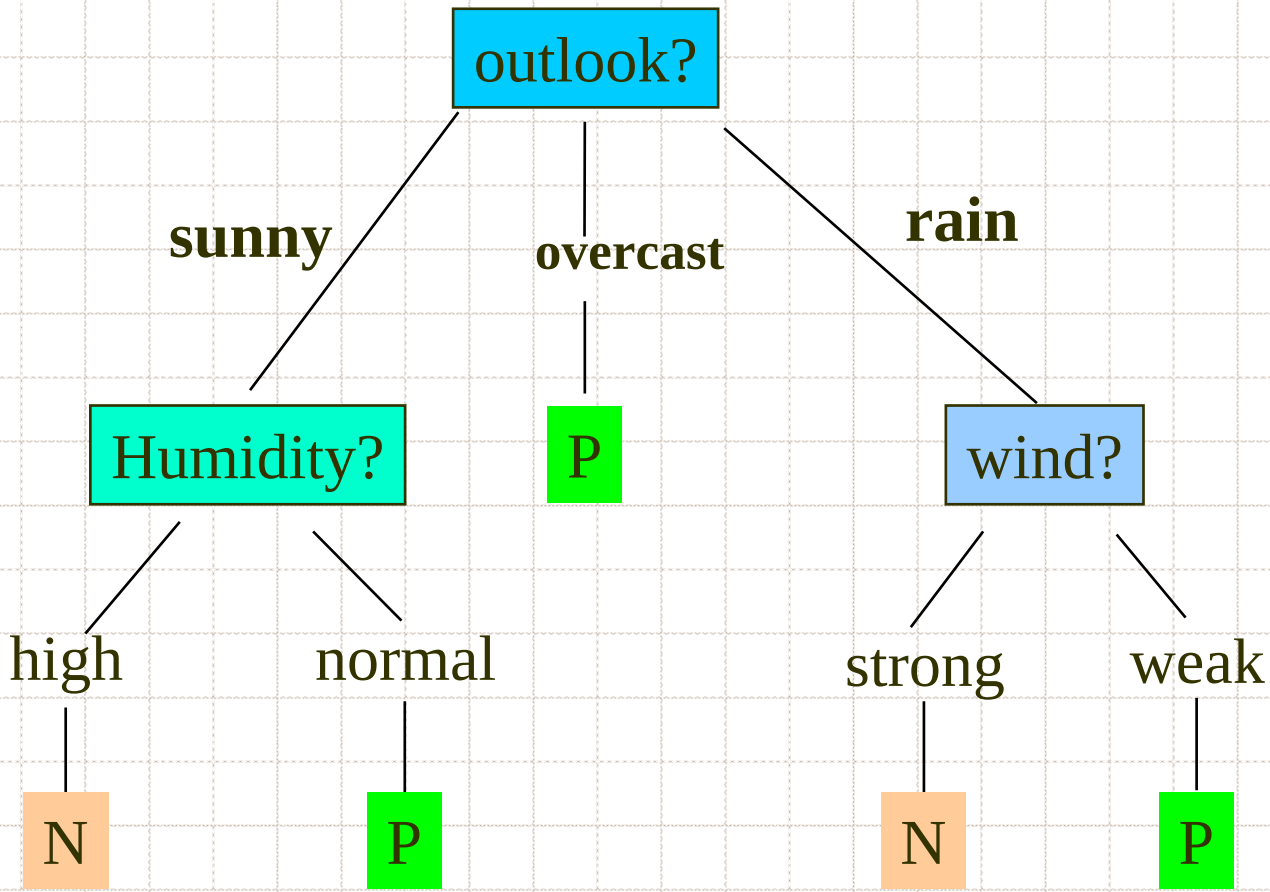
- Classification:
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Model construction: describing a set of predetermined classes
  - Each tuple is assumed to have a class label attribute
  - The set of tuples used for model construction: training set
  - The model is represented as classification rules ...
- Model usage: for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model

Classification: Training Dataset

This follows an example from Quinlan's ID3

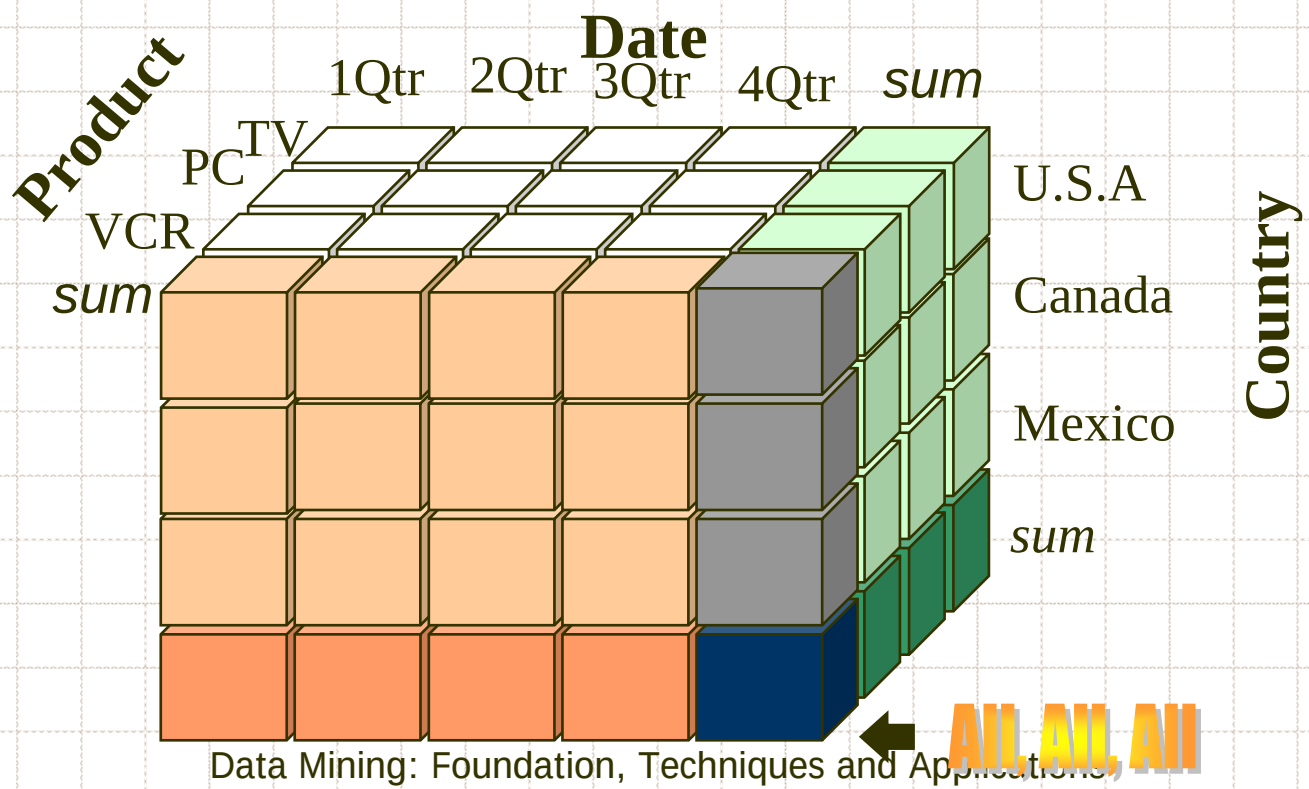
Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

Classification: Decision Tree Model



Techniques(Data Cube Computation)

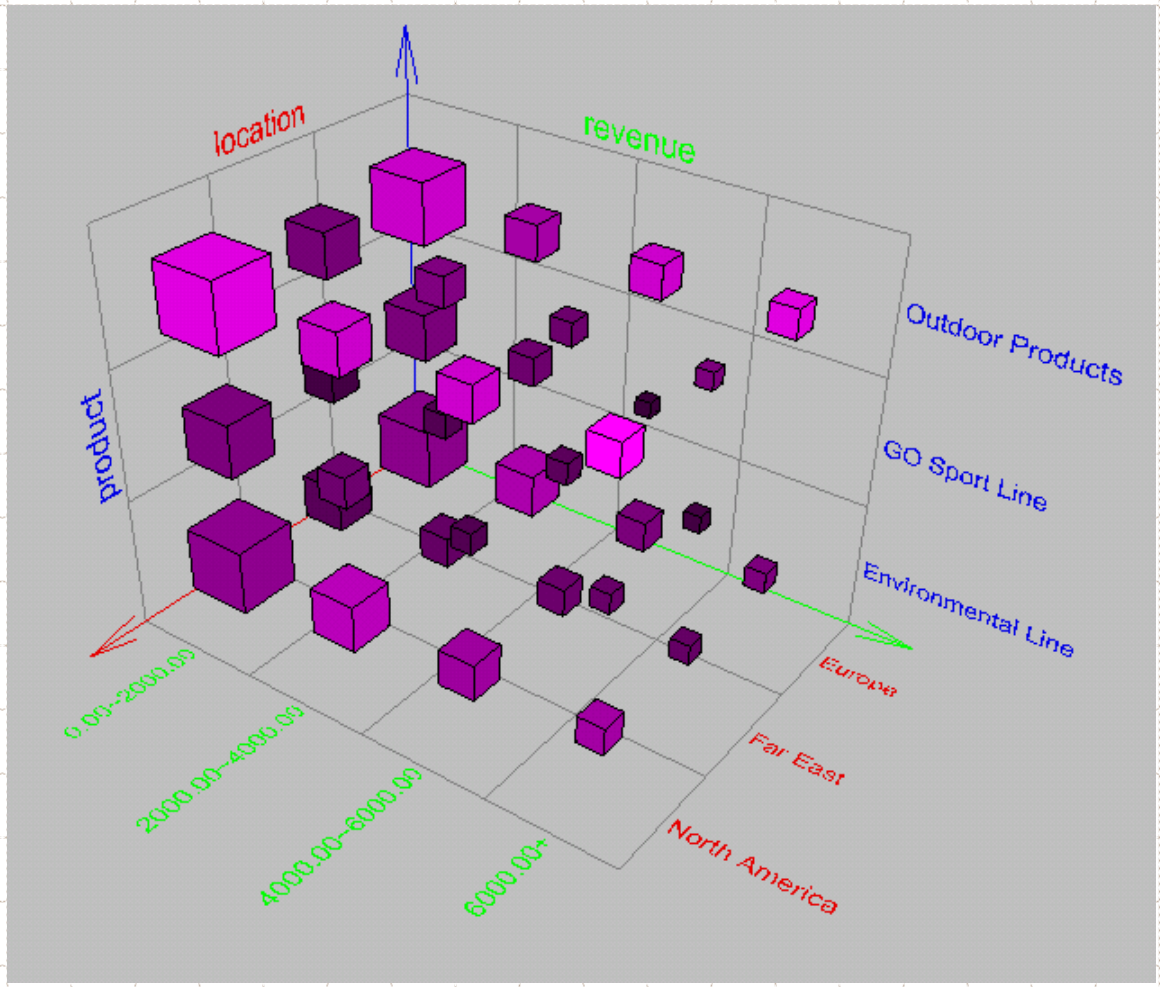
- Provide a multidimensional view of data for easier data analysis
- Eg: Sales volume as a function of product, month, and country



Techniques(Visualization I)

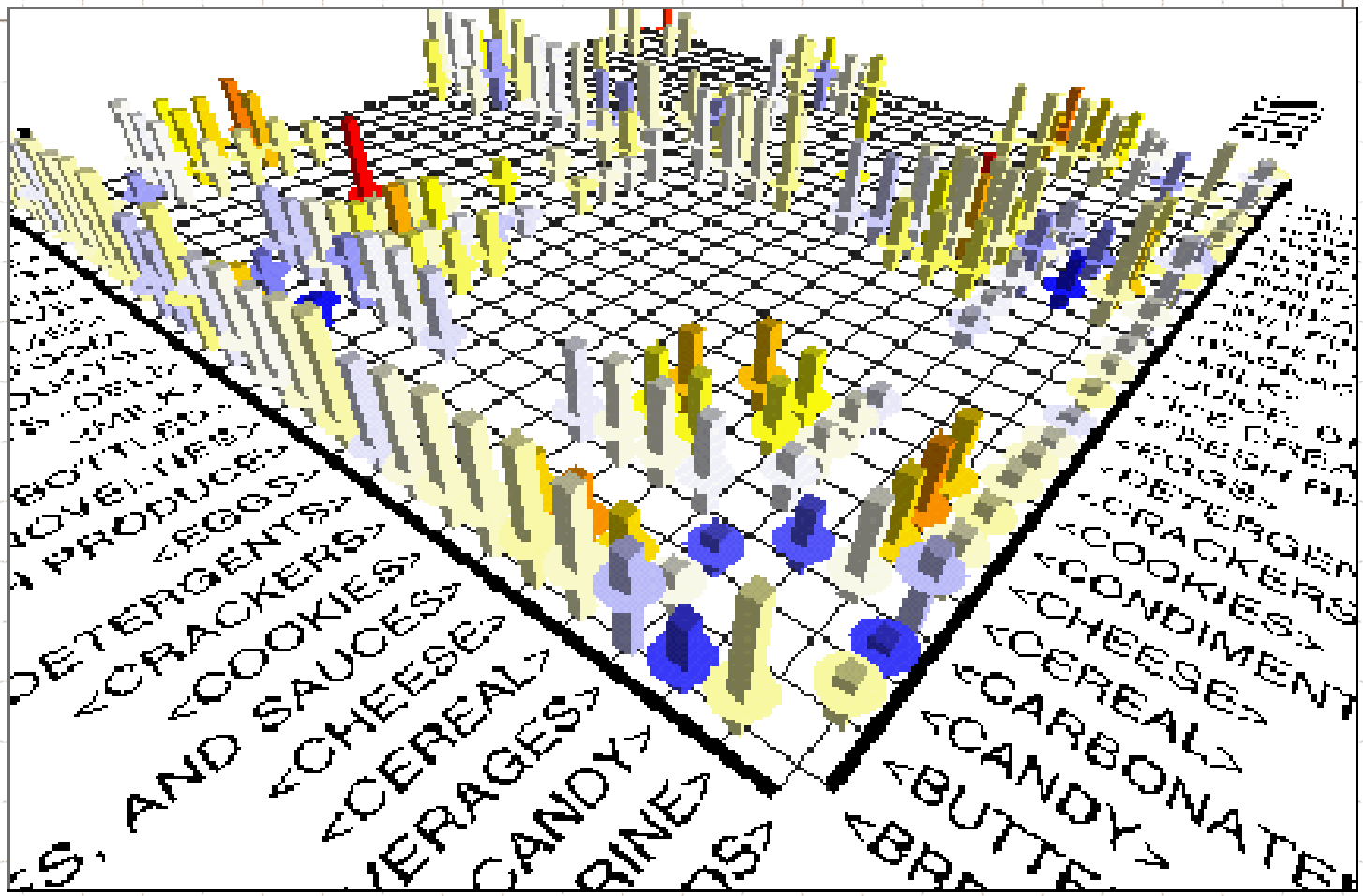
- One Picture May Worth 1000 Words!

Browsing a Data Cube



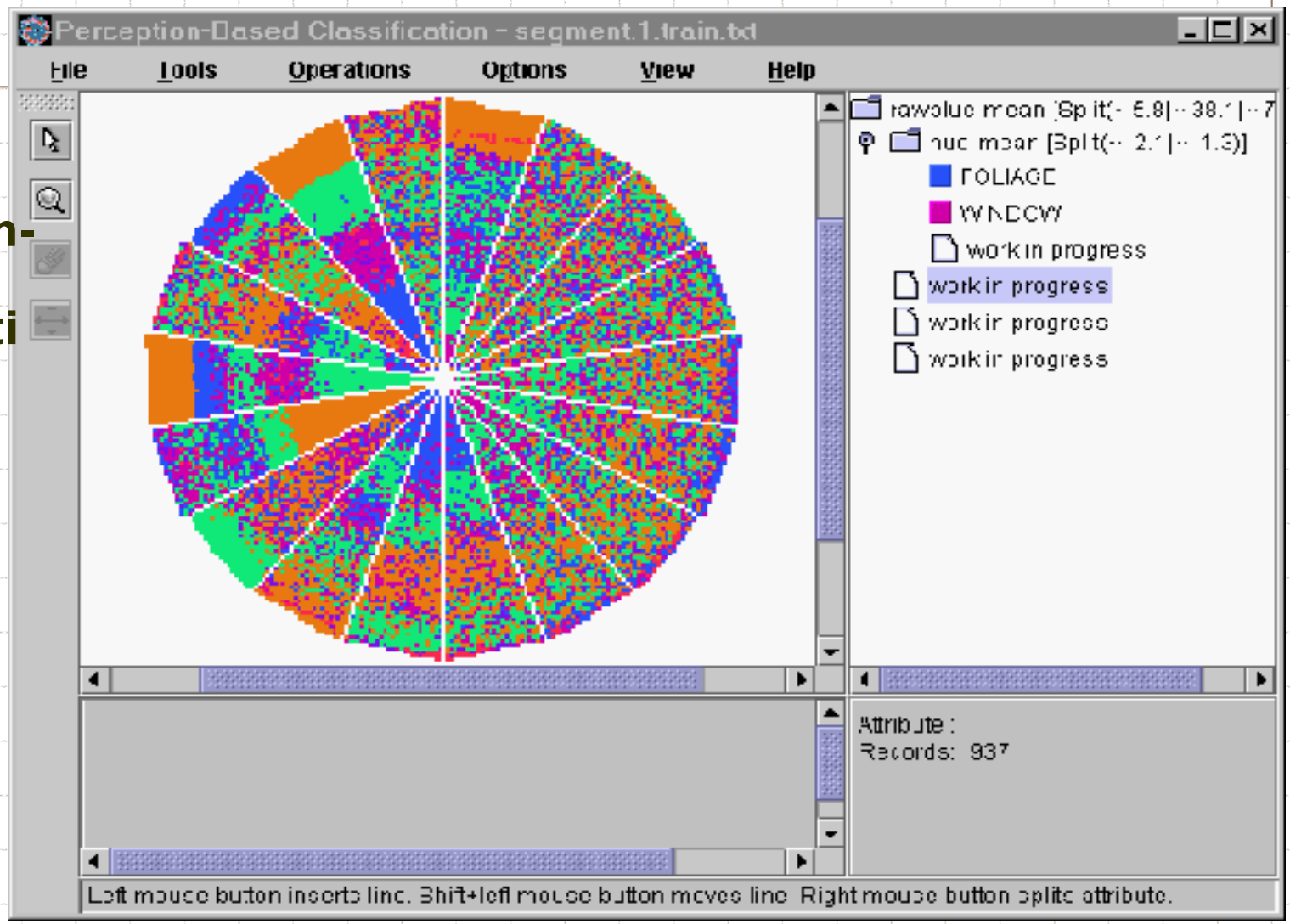
Techniques(Visualization II)

- **Visualization of Association Rules**



Techniques(Visualization III)

- **Perception-Based Classification (PBC)**



Foundation

Principles of DM (Statistic)

- Statistics
  - A discipline dedicated to data analysis. See data as a samples derived from real world model
  - Try to fit the data to a mathematical model with various form of statistical testing. Provide basic tools for data exploration.
  - Data mining, dredging, snooping and fishing are negative terms in the field of statistics in the 1960s.
    - If you look hard enough, you will always find a model that fitted the data!
  - Often used in data mining for rules verification etc.
- Example
  - Association rules are found based on **support** and **confidence**. However these two quantities tend to be high for items that occur frequently. For example, if two items A and B occur in the database 80% and 90% of the time, then the rules A->B will have expected support of 72% and confidence of 90% even if A and B occur together by random. To filter off such rules, we used the chi-squared test which test the independency between items.

Principles of DM (Machine Learning)

- Systematic search for statistical models and parameters over data sets
- Many techniques from data mining are in fact machine learning algorithms which are made scalable using database technologies, approximation algorithms etc. However, data mining is NOT only machine learning
- Example:
  - Classification is well studied in machine learning. Decision tree induction algorithms like C4.5, ID3 are first developed by the machine learning researchers and make efficiency by the database community

Principles of DM (Database Technology)

- Techniques like indexing, compression and query optimization help to provide quick access to relevant data for the mining algorithm
- **Examples**
  - Many clustering algorithms require k-nearest neighbors searches which is support by index structure like R* tree etc. With an index, time complexity is O(n log n), otherwise O(n^2)
  - After being compressed, databases can be stored in the main memory for fast mining operation

Principles of DM (Information Theory)

- Information Theory
  - is originally developed by Claude Shannon and applied in the area of communication
  - quantitative measurement of information using entropy or the minimum number of bits needed to encode a dataset
  - dataset is separated into a model and noise
  - minimize the total number of bits that is needed to represent the model and noise
- Example
  - Often used in decision tree building to avoid building over-complex decision tree
  - Often used to find deviants (which are noise)

Principles of DM (Others)

- Theoretical CS
  - Many rules discovery problem in data mining are NP-hard
  - Solution: Used Approximate, Randomized, Online Algorithms
- Mathematical Programming
  - Lots of optimization in data mining too
  - Algorithms like neural net training, k-means clustering, SVMs are in fact mathematical programming
- Computational Geometry ...

Schedule of this course

Date	Session	Topics
4th Dec 2007	Morning	**Introduction/A Quick Overview**
	Afternoon	**Machine Learning and Statistic**
	Night	**Database Techniques I (Indexing)**
5th Dec 2007	Morning	**Database Techniques II (Pre-Computation)**
	Afternoon	**Association Rule, Frequent Pattern (I)**
6th Dec 2007	Morning	**Association Rule, Frequent Pattern (II), Classification & Regression (I)**
	Afternoon	**Classification & Regression (II), Clustering (I)**
	Night	**Clustering (II),**
7th Dec 2007	Morning	**Skyline/Dominance Relationship Analysis**
	Afternoon	**Searching and Mining High Dimensional Data**
	Night	**Searching and Mining Sequences**
8th Dec 2007	Morning	**Searching and Mining Trees**
	Afternoon	**Searching and Mining Graphs**

Textbooks and References

- David Hand, Heikki Mannila, and Padhraic Smyth, "*Principles of Data Mining* ", MIT Press, August 2001
- Jiawei Han and Micheline Kamber "*Data Mining: Concepts and Techniques*", 2nd Edition, Morgan Kaufmann Publishers, August 2006
- Various research papers

Sponsorship

- 数据工程与知识工程教育部重点实验室
- NUS, School of Computing, Graduate Division