Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- Summary

Why Data Mining?

Huge volumes of data is Ubiquitous

The Explosive Growth of Data: from terabytes to petabytes (1 PB = 1000 TB = 1000000000000000 bytes = 10^15 bytes = 1000000 gigabytes)

Large Datasets in Scientific Research
Examples

- NASA Center for Computational Sciences in USA collects 125TB data per year.
- The Sanger Centre in UK hosted 20 TB key genomic data in 2002, increasing at a rate of 4 times per year, and now should more than 320P.
- The dataset from China State Key Lab of Remote Sensing Science for observing main land surface has size of 4TB per month.
- The dataset of astronomical information has the size of 3.65 TB per year.

Large datasets in business

Examples

- The dataset of mobile phone call records, collected by the China Net Corp., has the size of 300TB per year.
- Wal Mart has built data warehouse of Perabytes
- Best Buy has built data warehouse of Terabytes
Examples

- The size of the dataset of one airplane is more than 150GB. 1,000 airplanes will provide 150TB data.
- Industry panel discussion of VLDB 2002 shown that
  - many large enterprises have datasets with size of 100 TB, and is growing with rate of 20%
  - data warehouse is growing with the rate of 2 to 3 times of the growing rate of industry transaction datasets.

Datasets from WSN, IoT, CPS

There are 6.4 billion connected things worldwide now, rising up 30% from 2015, and will reach 20.8 billion by 2020.

Examples

- The UK funded a project
  - Analyze sensed data from aero-engines. There are around 100,000 engines in service. Each trans-Atlantic flight made by each engine generates about 1GB data per engine. The sensors generate many PB data per year.

Why Data Mining?

The world is data rich but information poor. We are drowning in data, but starving for knowledge!

Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets
Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

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
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, information harvesting, business intelligence, etc.

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process

Data Mining in Business Intelligence
KDD Process: A Typical View from ML and Statistics

Input Data → Data Pre-Processing → Data Mining → Post-Processing

Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

• This is a view from typical machine learning and statistics communities

Data Integration

Table 1
Table 2
Table 3

Antigenic Cartography

Matrix Completion

Antigenic Evolution and Genetic Evolution

Update Vaccine

Protein Sequence

Amino acid change

Antigenic distance

Matrix Completion

BE/32/92
MW/10/99
FU/411/02
NL/22/03

Antigenic Cartography

Table 1
Table 2
Table 3

Common HI values
Common HI values
Missing values
Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
  - Object-relational databases, Heterogeneous databases and legacy databases
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and information networks
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Relational Database

- A relational database for All Electronics.

```sql
CREATE TABLE customer (custID, name, address, age, occupation, annual_income, credit_information, category, ...);
CREATE TABLE item (itemID, brand, category, type, price, place_made, supplier, cost, ...);
CREATE TABLE employee (emplID, name, category, group, salary, commission, ...);
CREATE TABLE branch (branchID, name, address, ...);
CREATE TABLE purchases (transID, custID, emplID, date, time, method, paid, amount);
CREATE TABLE items_sold (transID, itemID, qty);
CREATE TABLE works_at (emplID, branchID);
```
Database Query

• Show me a list of all items that were sold in the last month.
• Show me the total sales of the last month, grouped by branch.
• Which salesperson had the highest sales?

Data Warehouse

Data Mining: On What Kinds of Data?

• Database-oriented data sets and applications
  – Relational database, data warehouse, transactional database
  – Object-relational databases, Heterogeneous databases and legacy databases
• Advanced data sets and advanced applications
  – Data streams and sensor data
  – Time-series data, temporal data, sequence data (incl. bio-sequences)
  – Structure data, graphs, social networks and information networks
  – Spatial data and spatiotemporal data
  – Multimedia database
  – Text databases
  – The World-Wide Web

Chapter 1. Introduction

• Why Data Mining?
• What Is Data Mining?
• What Kinds of Data Can Be Mined?
• What Kinds of Patterns Can Be Mined?
• What Kinds of Technologies Are Used?
• What Kinds of Applications Are Targeted?
• Major Issues in Data Mining
• A Brief History of Data Mining and Data Mining Society
• Summary
Data Mining Function: (1) Generalization

- Information integration and data warehouse construction
  - Data cleaning, transformation, integration, and multidimensional data model
- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

Characterization

- A summarization of the general characteristics or features of a target class of data.
- Example: Summarize the characteristics of customers who spend more than $5000 a year at All Electronics.
  - 40-50 years old, employed and have excellent credit rating

Discrimination

- Comparison of the general features of the target class data objects against the general features of objects from one or multiple classes.
- Example: A customer relationship manager at AllElectronics may want to compare two groups of customers—those who shop for computer products and those who rarely shop for such products

Data Mining Function: (2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper \( \rightarrow \) Beer [50%, 75%] (support, confidence)
    - \( \text{Age}(X, "20, 29") \land \text{income}(X, "40K..49K") \rightarrow \text{buys}(X, "laptop") \) [2%,60%]
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?
Data Mining Function: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Decision Trees

Application

- Credit Approve (Credit Limit)

  Age
  Personal assets
  Salary …
Data Mining Function: (4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Data Mining Function: (5) Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? — One person’s garbage could be another person’s treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis
Outlier Analysis

Application

• Credit Card Fraud

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

• Sequence, trend and evolution analysis
  – Trend, time-series, and deviation analysis: e.g., regression and value prediction
  – Sequential pattern mining
    • e.g., first buy digital camera, then buy large SD memory cards
  – Periodicity analysis
  – Motifs and biological sequence analysis
    • Approximate and consecutive motifs
  – Similarity-based analysis
• Mining data streams
  – Ordered, time-varying, potentially infinite, data streams
Structure and Network Analysis

• Graph mining
  – Finding frequent subgraphs (e.g., chemical compounds
• Information network analysis
  – Social networks: actors (objects, nodes) and relationships (edges)
    • e.g., author networks in CS
  – Multiple heterogeneous networks
    • A person could be multiple information networks: friends, family, classmates, …
  – Links carry a lot of semantic information: Link mining
• Web mining
  – Web is a big information network: from PageRank to Google
  – Analysis of Web information networks
    • Web community discovery, opinion mining, usage mining, …

Evaluation of Knowledge

• Are all mined knowledge interesting?
  – One can mine tremendous amount of “patterns”
  – Some may fit only certain dimension space (time, location, …)
  – Some may not be representative, may be transient, …
• Evaluation of mined knowledge → directly mine only interesting knowledge?
  – Descriptive vs. predictive
  – Coverage
  – Typicality vs. novelty
  – Accuracy
  – Timeliness
  – …

Interesting Patterns

• Easily understood by humans
• Valid on new or test data with some degree of certainty

An interesting pattern represents knowledge

• Novel: A pattern is also interesting if it validates a hypothesis that the user sought to confirm.

Chapter 1. Introduction

• Why Data Mining?
• What Is Data Mining?
• What Kinds of Data Can Be Mined?
• What Kinds of Patterns Can Be Mined?
• What Kinds of Technologies Are Used?
• What Kinds of Applications Are Targeted?
• Major Issues in Data Mining
• A Brief History of Data Mining and Data Mining Society
• Summary
Why Confluence of Multiple Disciplines?

• Tremendous amount of data
  – Algorithms must be scalable to handle big data
• High-dimensionality of data
  – Microarray may have tens of thousands of dimensions
• High complexity of data
  – Data streams and sensor data
  – Time-series data, temporal data, sequence data
  – Structure data, graphs, social and information networks
  – Spatial, spatiotemporal, multimedia, text and Web data
• New and sophisticated applications

Chapter 1. Introduction

• Why Data Mining?
• What Is Data Mining?
• What Kinds of Data Can Be Mined?
• What Kinds of Patterns Can Be Mined?
• What Kinds of Technologies Are Used?
• What Kinds of Applications Are Targeted?
• Major Issues in Data Mining
• A Brief History of Data Mining and Data Mining Society
• Summary

Applications of Data Mining

• Web page analysis: from web page classification, clustering to PageRank
• Collaborative analysis & recommender systems
• Basket data analysis to targeted marketing
• Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
• Data mining and software engineering
Major Issues in Data Mining (1)

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data

Major Issues in Data Mining (2)

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of science and information 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 data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, trend and outlier analysis, etc.
- Data mining technologies and applications
- Major issues in data mining

Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - IEEE Int. Conf. on Data Mining (ICDM)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - Int. Conf. on Web Search and Data Mining (WSDM)
- Other related conferences
  - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, ...
  - Web and IR conferences: WWW, SIGIR, WSDM
  - ML conferences: ICML, NIPS
  - PR conferences: CVPR, ...
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD
Where to Find References? DBLP, CiteSeer, Google

- **Data mining and KDD (SIGKDD, CDROM)**
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-SDM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, SIGKDD Explorations, ACM TDKDD

- **Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)**
  - Conferences: ACM-SIGMOD, ACM-PVDS, VLDB, IEEE-ICDE, EDBT, KDD, DASFAA

- **AI & Machine Learning**
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.

- **Web and IR**
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems, etc.

- **Statistics**
  - Conferences: Joint Stat Meeting, etc.
  - Journals: Annals of statistics, etc.

- **Visualization**
  - Conferences: Proceedings CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. on visualization and computer graphics, etc.

**Recommended Reference Books**

- E. Fayyad, G. Grinstein, and A. Wu, ed. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- Y. Sun and J. Han. Mining Heterogeneous Information Networks. Morgan & Claypool, 2012
- D-S. Tsai, M. Steinbach, and V. Kumar. Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya. Predictive Data Mining, Morgan Kaufmann, 1999