CH1

Major sources of abundant data

- **Business:** Web, e-commerce, transactions, stocks, …
- **Science:** Remote sensing, bioinformatics, scientific simulation, …
- **Society and everyone:** news, digital cameras, YouTube

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

Knowledge Discovery (KDD) Process:

- **Data Cleaning**
- **Data Warehouse**
- **Task-relevant Data**
- **Data Mining**
- **Pattern Evaluation**

Data Mining in Business Intelligence

- **Data Sources**
- **Reporting Data, Preprocessing/Integration, Data Warehouses**
- **Data Exploration**
- **Data Mining**
- **Data Presentation**
- **Decision Making**

KDD Process: A Typical View from ML and Statistics

- **Input Data**
- **Data Pre-Processing**
  - **Data integration**
  - **Normalization**
  - **Feature selection**
  - **Dimension reduction**
- **Data Mining**
  - **Pattern discovery**
  - **Association & correlation**
  - **Classification**
  - **Clustering**
  - **Outlier analysis**
- **Post-Processing**
  - **Pattern evaluation**
  - **Pattern selection**
  - **Pattern interpretation**
  - **Pattern visualization**
- **Pattern Information Knowledge**

Multi-Dimensional View of Data Mining:

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining Function:

- **Generalization**
- **Association and Correlation Analysis**
- **Classification**
- **Cluster Analysis**
- **Outlier Analysis**

Why Confluence of Multiple Disciplines?

- Tremendous amount of data
- High-dimensionality of data
- High complexity of data
- New and sophisticated applications

Outlier analysis Outlier: A data object that does not comply with the general behavior of the data

Applications of Data Mining:

- Web page analysis
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis
- Data mining and software engineering
- From major dedicated data mining systems/tools

Major Issues in Data Mining:

- Mining Methodology
- User Interaction
- Efficiency and Scalability
- Diversity of data types
- Data mining and society

CH2

Types of Data Sets:

- **Record:** Relational records, Data matrix, Document data: text documents: term frequency vector, Transaction data
- **Graph and network:** World Wide Web, Social or information networks, Molecular Structures
- **Ordered:** Video data: sequence of images, Temporal data: time-series, Sequential Data, transaction sequences, Genetic sequence data
- **Spatial,** image and multimedia: Spatial data: maps, Image data, Video data.

Important Characteristics of Structured Data:

- **Dimensionality:** Curse of dimensionality
- **Sparsity:** Only presence counts
- **Resolution:** Patterns depend on the scale
- **Distribution:** Centrality and dispersion

Data Objects:

- Data sets are made up of data objects.
- A **data object** represents an entity.
  - Examples: sales database, medical database, university database
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes

Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.

  - *E.g., customer _ID, name, address*

Discrete vs. Continuous Attributes:

- **Discrete Attribute**
  - **Has only a finite or countably infinite set of values**
    - **E.g., zip codes, profession, or the set of words in a collection of documents**
  - **Sometimes, represented as integer variables**
- **Continuous Attribute**
  - **Has real numbers as attribute values**
  - **Continuous attributes are typically represented as floating-point variables**

Graphic Displays of Basic Statistical Descriptions:

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Histograms Often Tell More than Boxplots

Attribute Types:

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Why data visualization?

- Gain insight into an information space by mapping data onto graphical primitives
- Provide qualitative overview of large data sets
- Search for patterns, trends, structure, irregularities, relationships among data
- Help find interesting regions and suitable parameters for further quantitative analysis
- Provide a visual proof of computer representations derived

Categorization of visualization methods:

- Pixel-oriented visualization techniques
- Geometric projection visualization techniques
- Icon-based visualization techniques
- Hierarchical visualization techniques
- Visualizing complex data and relations

Dimensional Stacking:

- **Partitioning of the n-dimensional attribute space in 2-D subspaces, which are 'stacked' into each other**
- **Partitioning of the attribute value ranges into classes. The important attributes should be used on the outer levels.**

InfoCube:

- **A 3-D visualization technique where hierarchical information is displayed as nested semi-transparent cubes**
- **The outermost cubes correspond to the top level data, while the subnodes or the lower level data are represented as smaller cubes inside the outermost cubes, and so on**

Visualizing Complex Data and Relations:

- **Visualizing non-numerical data: text and social networks**
- **Tag cloud: visualizing user-generated tags**
- **Besides text data, there are also methods to visualize relationships, such as visualizing social networks**

Similarity and Dissimilarity:

- **Similarity**
  - **Numerical measure of how alike two data objects are**
  - **Value is higher when objects are more alike**
  - **Often falls in the range [0,1]**
- **Dissimilarity (e.g., distance)**
  - **Numerical measure of how different two data objects are**
  - **Lower when objects are more alike**
  - **Minimum dissimilarity is often 0**
  - **Upper limit varies**
- **Proximity refers to a similarity or dissimilarity**

CH3:

Why Preprocess the Data?

- **Measures for data quality: A multidimensional view**
  - **Accuracy: correct or wrong, accurate or not**
  - **Completeness: not recorded, unavailable, …**
  - **Consistency: some modified but some not, dangling, …**
  - **Timeliness: timely update?**
  - **Believability: how trustable the data are correct?**
  - **Interpretability: how easily the data can be understood?**

Major Tasks in Data Preprocessing :

- **Data cleaning**
  - **Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies**
- **Data integration**
  - **Integration of multiple databases, data cubes, or files**
- **Data reduction**
  - **Dimensionality reduction**
  - **Numerosity reduction**
  - **Data compression**
- **Data transformation and data discretization**
  - **Normalization**
  - **Concept hierarchy generation**

How to Handle Missing Data?

- **Ignore the tuple: usually done when class label is missing**
- **Fill in the missing value manually: tedious + infeasible?**
- **Fill in it automatically with**
  - **a global constant : e.g., “unknown”, a new class?!**
  - **the attribute mean**
  - **the attribute mean for all samples belonging to the same class**

Noise: **random error or variance in a measured variable**

Incorrect attribute values may be due to

- **faulty data collection instruments**
- **data entry problems**
- **data transmission problems**
- **technology limitation**
- **inconsistency in naming convention**

How to Handle Noisy Data?

- **Binning**
- **Regression**
- **Clustering**
- **Combined computer and human inspection**

Data integration: **Combines data from multiple sources into a coherent store**

Data reduction: **Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results**

Why data reduction?

A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Data reduction strategies:

- **Dimensionality reduction, e.g., remove unimportant attributes**
- **Numerosity reduction (some simply call it: Data Reduction)**
- **Data compression**

What Is Wavelet Transform?

- **Decomposes a signal into different frequency subbands**

Principal Component Analysis (PCA)

- **Find a projection that captures the largest amount of variation in data**
- **The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space**

Similarity and Dissimilarity

- **Similarity**
  - **Numerical measure of how alike two data objects are.**
  - **Is higher when objects are more alike.**
  - **Often falls in the range [0,1]**
- **Dissimilarity**
  - **Numerical measure of how different are two data objects**
  - **Lower when objects are more alike**
  - **Minimum dissimilarity is often 0**
  - **Upper limit varies**

Clustering: **Partition data set into clusters based on similarity, and store cluster representation.**

Sampling: **obtaining a small sample *s* to represent the whole data set *N***

Types of Sampling:

- **Simple random sampling**
- **Sampling without replacement**
- **Sampling with replacement**
- **Stratified sampling:**

Data Transformation: **A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values**

- Data Transformation Methods
  - **Smoothing: Remove noise from data**
  - **Attribute/feature construction: New attributes constructed from the given ones**
  - **Aggregation: Summarization, data cube construction**
  - **Normalization: Scaled to fall within a smaller, specified range**
    - **min-max normalization**
    - **z-score normalization**
    - **normalization by decimal scaling**
  - **Discretization: Concept hierarchy climbing**

Discretization: **Divide the range of a continuous attribute into intervals**

- Discretization methods:
  - **Binning: Top-down split, unsupervised**
  - **Histogram analysis: Top-down split, unsupervised**
  - **Clustering analysis (unsupervised, top-down split or bottom-up merge)**
  - **Decision-tree analysis (supervised, top-down split)**
  - **Correlation (e.g., χ^2) analysis (unsupervised, bottom-up merge)**

CH4:

What is a Data Warehouse?

- **A decision support database that is maintained separately from the organization's operational database**
- **Support information processing by providing a solid platform of consolidated, historical data for analysis.**
- **A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process.**

OLTP (On-line Transaction Processing) vs. OLAP(On-line Analytical Processing)

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Why a Separate Data Warehouse?

- **High performance for both systems**
- **Different functions and different data**

Three Data Warehouse Models:

- **Enterprise warehouse: collects all of the information about subjects spanning the entire organization**
- **Data Mart: a subset of corporate-wide data that is of value to a specific group of users. Its scope is confined to specific, selected groups, such as marketing data mart**
- **Virtual warehouse: A set of views over operational databases**

Extraction, Transformation, and Loading (ETL):

- **Data extraction: get data from multiple, heterogeneous, and external sources**
- **Data cleaning: detect errors in the data and rectify them when possible**
- **Data transformation: convert data from legacy or host format to warehouse format**
- **Load: sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions**
- **Refresh: propagate the updates from the data sources to the warehouse**

Meta data is the data defining warehouse objects.

A data warehouse is based on a multidimensional data model which views data in the form of a data cube

Conceptual Modeling of Data Warehouses:

- **Star schema: A fact table in the middle connected to a set of dimension tables**
- **Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake**
- **Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation**

Data Cube Measures: Three Categories:

- **Distributive: E.g., count(), sum(), min(), max()**
- **Algebraic: E.g., avg(), min_N(), standard_deviation()**
- **Holistic: E.g., median(), mode(), rank()**

Typical OLAP Operations:

- **Roll up (drill-up): summarize data**
  - ***by climbing up hierarchy or by dimension reduction***
- **Drill down (roll down): reverse of roll-up**
  - ***from higher level summary to lower level summary or detailed data, or introducing new dimensions***
- **Slice and dice: *project and select***
- **Pivot (rotate):**
  - ***reorient the cube, visualization, 3D to series of 2D planes***
- **Other operations**
  - ***drill across: involving (across) more than one fact table***
  - ***drill through: through the bottom level of the cube to its back-end relational tables (using SQL)***

Four views regarding the design of a data warehouse:

- **Top-down view: allows selection of the relevant information necessary for the data warehouse**
- **Data source view: exposes the information being captured, stored, and managed by operational systems**
- **Data warehouse view: consists of fact tables and dimension tables**
- **Business query view: sees the perspectives of data in the warehouse from the view of end-user**

Typical data warehouse design process

- **Choose a business process to model, e.g., orders, invoices, etc.**
- **Choose the *grain* (*atomic level of data*) of the business process**
- **Choose the dimensions that will apply to each fact table record**
- **Choose the measure that will populate each fact table record**

Data Warehouse Usage (applications):

- **Information processing: supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs**
- **Analytical processing: multidimensional analysis of data warehouse data**
- **Data mining: knowledge discovery from hidden patterns**

Why online analytical mining?

- **High quality of data in data warehouses**
- **Available information processing structure surrounding data warehouses**
- **OLAP-based exploratory data analysis**
- **On-line selection of data mining functions**

OLAP tools

ODBC, OLEDB, Web accessing, service facilities, reporting and OLAP tools

CH5:

Properties of Proposed Method:

- **Partitions the data vertically**
- **Reduces high-dimensional cube into a set of lower dimensional cubes**
- **Online re-construction of original high-dimensional space**
- **Lossless reduction**
- **Offers tradeoffs between the amount of pre-processing and the speed of online computation**

Intra-Cuboid Expansion

- **Combine other cells' data into own to "boost" confidence**
- **Cell segment similarity**
- **Cell value similarity**

Four ways to interact OLAP-styled analysis and data mining

- **Using cube space to define data space for mining**
- **Using OLAP queries to generate features and targets for mining, e.g., multi-feature cube**
- **Using data-mining models as building blocks in a multi-step mining process, e.g., prediction cube**
- **Using data-cube computation techniques to speed up repeated model construction**

CH6:

- support, ***s*, probability that a transaction contains X ∪ Y**
- confidence, ***c*, conditional probability that a transaction having X also contains *Y***

Frequent pattern: **a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set**

- **Applications :Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.**

Apriori pruning principle**: If there is any itemset which is infrequent, its superset should not be generated/tested!**

Closed pattern: **is a lossless compression of freq. patterns**

An itemset X is closed if: **X is *frequent* and there exists *no super-pattern* Y ɔ X, *with the same support* as X**

An itemset X is a max-pattern if: **X is frequent and there exists no frequent super-pattern Y ɔ X (proposed by**

When minsup is low:, **there exist potentially an exponential number of frequent itemsets**

The downward closure: **property of frequent patterns**

- **Any subset of a frequent itemset must be frequent**

Benefits of the FP-tree Structure:

- **Completeness**
  - **Preserve complete information for frequent pattern mining**
  - **Never break a long pattern of any transaction**
- **Compactness**
  - **Reduce irrelevant info—infrequent items are gone**
  - **Items in frequency descending order: the more frequently occurring, the more likely to be shared**
  - **Never be larger than the original database (not count node-links and the *count* field)**

Parallel projection vs. partition projection techniques

- **Parallel projection**
  - **Project the DB in parallel for each frequent item**
  - **Parallel projection is space costly**
  - **All the partitions can be processed in parallel**
- **Partition projection**
  - **Partition the DB based on the ordered frequent items**
  - **Passing the unprocessed parts to the subsequent partitions**

CLOSET**: An Efficient Algorithm for Mining Frequent Closed Itemsets (تحتاج مراجعه!!)**

دعواتكم لمن ساهم فى هذا العمل