CS377: Database Systems
Data Warehouse and Data Mining

Li Xiong
Department of Mathematics and Computer Science
Emory University
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
Knowledge Discovery (KDD) Process

KDD Process:
- Databases
- Data Integration
- Data Cleaning
- Task-relevant Data
- Data Mining
- Pattern Evaluation

Diagram:
- Knowledge
- Data Warehouses
- Selection and transformation
What is a Data Warehouse?

- A data warehouse is a database used for reporting and analysis
- “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”—W. H. Inmon

- Key aspects
  - A decision support database that is maintained separately from the organization’s operational database
  - Support OLAP (vs. OLTP)

- Data warehousing:
  - The process of constructing and using data warehouses
Data Warehouse Approach

Query & Analysis

Metadata

Warehouse

Extract, transform and load (ETL)

Source

Source

Source
Multi Dimensional View: From Tables to Data Cubes

- **Data warehouse model**
  - *Multidimensional data model* views data in the form of a data (hyper) cube
  - Each cell corresponds to an aggregated value, such as total sales

- **Data warehouse operations**
  - Drill down
  - Roll up
  - Slide and dice
Orders database - relational schema
Data warehouse for the orders database – star schema

- Dimensional tables
- Fact tables – each fact corresponds to a data cube
<table>
<thead>
<tr>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mostly updates</td>
<td>Mostly reads</td>
</tr>
<tr>
<td>Many small transactions</td>
<td>Queries long, complex</td>
</tr>
<tr>
<td>Mb-Tb of data</td>
<td>Gb-Tb of data</td>
</tr>
<tr>
<td>Raw data</td>
<td>Summarized, consolidated data</td>
</tr>
<tr>
<td>Clerical users</td>
<td>Decision-makers, analysts</td>
</tr>
<tr>
<td>Up-to-date data</td>
<td>Historical data</td>
</tr>
<tr>
<td>Consistency, recoverability critical</td>
<td></td>
</tr>
</tbody>
</table>
Data warehouse construction

- Semantic data integration – reconciling semantic heterogeneity of information sources

- Levels
  - Schema matching (schema mapping)
  - Data matching (data deduplication, record linkage, entity/object matching)
Schema Matching

- Techniques
  - Rule based
  - Learning based

- Type of matches
  - 1-1 matches vs. complex matches (e.g. list-price = price *(1+tax_rate))

- Information used
  - Schema information: element names, data types, structures, number of sub-elements, integrity constraints
  - Data information: value distributions, frequency of words
  - External evidence: past matches, corpora of schemas
  - Ontologies

- Multi-matcher architecture
Record Linkage

- Rule based
  - pair-wise similarity comparison
  - Varying weights for different fields, e.g. name, SSN, birthdate

- Machine learning based
Knowledge Discovery (KDD) Process

Data Cleaning

Data Integration

Data Warehouse

Task-relevant Data

Selection and transformation

Data Mining

Pattern Evaluation

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

- Predictive: predict the value of a particular attribute based on the values of other attributes
  - Classification
  - Regression

- Descriptive: derive patterns that summarize the underlying relationships in data
  - Cluster analysis
  - Association analysis
## Classification – Fruit Identification

<table>
<thead>
<tr>
<th>Skin</th>
<th>Color</th>
<th>Size</th>
<th>Flesh</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hairy</td>
<td>Brown</td>
<td>Large</td>
<td>Hard</td>
<td>Safe</td>
</tr>
<tr>
<td>Hairy</td>
<td>Green</td>
<td>Large</td>
<td>Hard</td>
<td>Safe</td>
</tr>
<tr>
<td>Smooth</td>
<td>Red</td>
<td>Large</td>
<td>Soft</td>
<td>Dangerous</td>
</tr>
<tr>
<td>Hairy</td>
<td>Green</td>
<td>Large</td>
<td>Soft</td>
<td>Safe</td>
</tr>
<tr>
<td>Smooth</td>
<td>Red</td>
<td>Small</td>
<td>Hard</td>
<td>Dangerous</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
Classification and prediction

- Classification: construct models (functions) that describe and distinguish classes for future prediction
- Prediction/regression: predict unknown or missing numerical values
- Derived models can be represented as rules, mathematical formulas, etc.
  - Classification: Decision tree, Bayesian classification, Neural networks, Support vector machines, kNN
  - Regression: linear and non-linear regression
Classification

Training Data

Training/Learning

Classifier (Model)

Unseen Data

Classification

Classified Data

<table>
<thead>
<tr>
<th>Skin</th>
<th>Color</th>
<th>Size</th>
<th>Flesh</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hairy</td>
<td>Brown</td>
<td>Large</td>
<td>Hard</td>
<td>Safe</td>
</tr>
<tr>
<td>Hairy</td>
<td>Green</td>
<td>Large</td>
<td>Hard</td>
<td>Safe</td>
</tr>
<tr>
<td>Smooth</td>
<td>Red</td>
<td>Large</td>
<td>Soft</td>
<td>Dangerous</td>
</tr>
<tr>
<td>Hairy</td>
<td>Green</td>
<td>Large</td>
<td>Soft</td>
<td>Safe</td>
</tr>
<tr>
<td>Smooth</td>
<td>Red</td>
<td>Small</td>
<td>Hard</td>
<td>Dangerous</td>
</tr>
</tbody>
</table>

Rule: if skin = smooth and color = red, then dangerous
Classification example

Diagram showing a classification process with decision points for attributes such as attention span, hardworking, attitude, and specific medical fields.
Frequent pattern mining and association analysis

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

- **Applications**
  - Basket data analysis — Beer and diapers
  - Web log (click stream) analysis

- **Challenge**: efficient algorithms to handle exponential size of the search space
Cluster and outlier analysis

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Unsupervised learning (vs. supervised learning)
  - Maximizing intra-class similarity & minimizing interclass similarity

- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
    - E.g. Extreme large purchase

![Cluster and Outlier Analysis Diagram](image-url)
For more

- CS570: Introduction to Data Mining