COURSE NAME:
DATA WAREHOUSING & DATA MINING

LECTURE 11
TOPICS TO BE COVERED:

- Data mining
- Motivation
- Definition
- Task

MOTIVATION: "NECESSITY IS THE MOTHER OF INVENTION"

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored 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
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

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.) and application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s—2000s:
  - Data mining and data warehousing, multimedia databases, and Web databases

WHAT IS DATA MINING?

- **Data mining** is the process of identifying valid, novel, useful and understandable patterns in data.
- Also known as **KDD** (**K**nowledge **D**iscovery in **D**atabases).
- Data Mining refers to *extracting or "mining" knowledge from large amounts of data.*

DATA MINING

- The non-trivial extraction of novel, implicit, and actionable knowledge from large datasets.
  - Extremely large datasets
  - Discovery of the non-obvious
  - Useful knowledge that can improve processes
  - Can not be done manually
- Technology to enable data exploration, data analysis, and data visualization of very large databases at a high level of abstraction, without a specific hypothesis in mind.
- Sophisticated data search capability that uses statistical algorithms to discover patterns and correlations in data.

DATA MINING

- Data Mining is a step of Knowledge Discovery in Databases (KDD) Process
  - Data Warehousing
  - Data Selection
  - Data Preprocessing
  - Data Transformation
  - Data Mining
  - Interpretation/Evaluation
- Data Mining is sometimes referred to as KDD and DM and KDD tend to be used as synonyms.

THE DATA MINING PROCESS

- Understanding domain, prior knowledge, and goals
- Data integration and selection
- Data cleaning and pre-processing
- Modeling and searching for patterns
- Interpreting results
- Consolidating and deploying discovered knowledge
- Loop

WHY DATA MINING? — POTENTIAL APPLICATIONS

- Database analysis and decision support
  - Market analysis and management
    - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and management
- Other Applications
  - Text mining (news group, email, documents) and Web analysis.
  - Intelligent query answering

MARKET ANALYSIS AND MANAGEMENT

- Where are the data sources for analysis?
  - 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
  - Conversion of single to a joint bank account: marriage, etc.
- Cross-market analysis
  - Associations/co-relations between product sales
  - Prediction based on the association information

MARKET ANALYSIS AND MANAGEMENT

- Customer profiling
  - data mining can tell you what types of customers buy what products (clustering or classification)
- Identifying customer requirements
  - identifying the best products for different customers
  - use prediction to find what factors will attract new customers
- Provides summary information
  - various multidimensional summary reports
  - statistical summary information (data central tendency and variation)

CORPORATE ANALYSIS AND 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

FRAUD DETECTION AND MANAGEMENT

- Applications
  - widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.
- Approach
  - use historical data to build models of fraudulent behavior and use data mining to help identify similar instances
- Examples
  - auto insurance: detect a group of people who stage accidents to collect on insurance
  - money laundering: detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
  - medical insurance: detect professional patients and ring of doctors and ring of references