Arsi University

Dep't of Information Technology

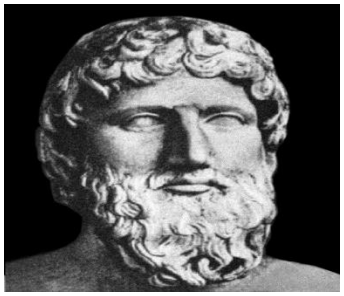
Introduction to Data Warehousing and Data Mining

Contents of the course

- ✓ **Chapter 1: Introduction to Dm. and data warehousing**
- ✓ **Chapter 2: Data Preprocessing**
- ✓ **Chapter 3: Classification**
- ✓ **Chapter 4: Clustering**
- ✓ **Chapter 5: Associations**

Introduction

- *In this chapter we will cover the following issues in brief*
  - *Motivation: Why data mining?*
  - *What is data mining?*
  - *Data mining vs Statistics*
  - *Challenges in Data Mining*
  - *Application of data mining*
  - *Data mining functionality*
  - *Are all the patterns interesting?*
  - *Classification of data mining systems*



Motivation: "Necessity is the Mother of Invention"

- *Our capacity of generating and collecting data have been increased rapidly in the last several decades*
- *Huge amount of data is available at the tip of our hand*
- *It is predicted that more data will be produced in the next 2 years than has been generated during the entire existence of humankind!*

Motivation:

"Necessity is the Mother of Invention"

- ***Contributing factors include***
  - *Widespread use of bar code for most commercial products,*
  - *40 billion RFID tags world wide*
  - *Billions of telephone calls are recorded daily worldwide*
  - *Billions of customers are using face book and other social network applications*
  - *10 billions of content are shared on face book per month*
  - *Computerization of many business, scientific, and governmental transactions,*
  - *Advances in data collection tools (audio, video, satellite, remote sensing, scanning, image capturing tools)*
  - *Usage of WWW as a global information system*
  - *comprehensive application software,*
  - *new computing and storage technologies*

Motivation:
"Necessity is the Mother of Invention"

- *All this have made it easier to create, collect, and store all types of data.*
- *As a result it creates a problem what is called data exposition.*
- *Data explosion is the problem of having huge amount of data in an enterprise stored in databases, data warehouses and other information repositories generated by automated*
- *data collection tools and mature database technology in large databases which has to be processed to make a decision.*
- *As the size of **data get larger, analyzing** the data becomes very difficult*

Motivation:
"Necessity is the Mother of Invention"

- ***Data can be managed and stored in***
  - *Data warehouse*
  - *structured databases;*
  - *in semi-structured file systems, such as e-mail;*
  - *unstructured fixed content, like documents and graphic files.*

Motivation:
"Necessity is the Mother of Invention"

- *Companies rely on this enterprise data to improve **decision-making and to gain a competitive advantage**;*
- *Data has indeed become a highly valued **business asset**.*
- *The huge amount of data exceeds our human ability to make comprehension on the data and to put the best decision without tools*
- *Generating and storing of large volumes of data has reached a critical mass and appropriate tools for comprehend the data becomes vital*.

Motivation:
"Necessity is the Mother of Invention"

- *We are drowning in data, but starving for knowledge!*
- ***The Solution:*** ***Data warehousing and data mining***
- *Data mining can be viewed as a result of the natural evolution of information technology.*
- *This can be more explained if we look at the evolution of database technology since 19th century.*

Motivation:
"Necessity is the Mother of Invention"

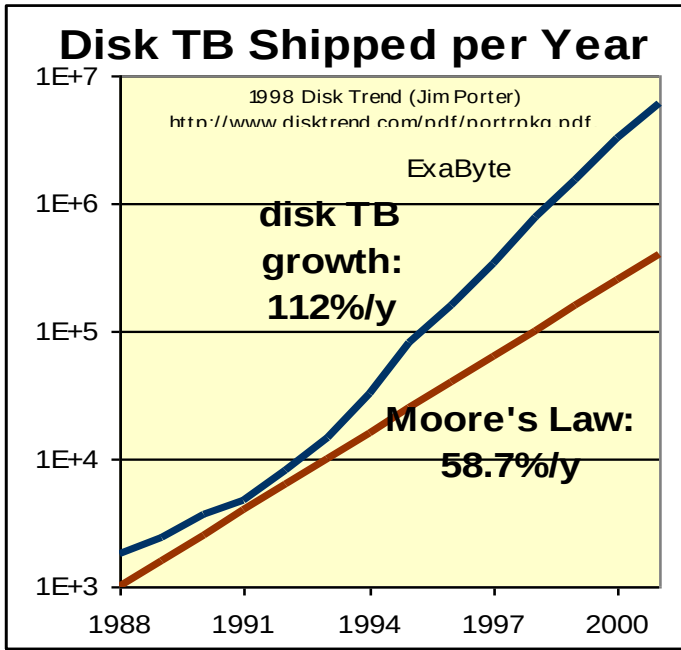
- *1960s:*
  - *Known to be the era of primitive file processing*
  - *There were activities such as*
    - *Data collection,*
    - *database creation,*
    - *Information management system (IMS), mainly using COBOL*
- *1970s:*
  - *Relational data model, relational DBMS implementation*
  - *Data modeling tools like ER diagram*
  - *Indexing and data organization techniques such as B+ tree, hashing, etc*
  - *Query language such as SQL*
  - *User interfaces, forms and reports*
  - *Query processing and optimization techniques*
  - *Transaction management: recovery, concurrency control, etc*
  - *Online Transaction processing (OLTP)*

Motivation:
"Necessity is the Mother of Invention"

- *1980s:*
  - *Period of advanced DB Systems*
    - *advanced data models*
      - ***extended-relational, Object Oriented, Object-Relational, deductive, etc.)***
    - *application-oriented DBMS*
      - ***spatial, temporal, multimedia, active, scientific, engineering, Knowledgebase, etc.)***
- *1990s—2000s:*
  - ***Data mining and data warehousing, Knowledge discovery, OLAP and Web based databases***

Data Mining Enablers

- ***Explosion of data***
- ***Fast and cheap computation and storage***
  - *Moore's Law: processing doubles every 19 months*
  - *Disk storage doubles every 9 months*
  - *Database technology*
- ***Competitive pressure in business***
  - *Data has value!*
- ***New, successful models***
- ***Commercial products***
    - *SAS, SPSS, Insightful, IBM, Oracle*
  - *Open Source products*
    - *Weka*



What is Data Mining?

- *Data mining is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data source(Han and Kamber)*
- *The process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and*
- *by using pattern recognition technologies as well as statistical and mathematical techniques (The Gartner Group)*
- *The exploration and analysis of large quantities of data in order to discover meaningful patterns and rules (Berry and Linoff)*
- *The nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Frawley, Paitestsky-Shapiro and Mathews)*
- *The non-trivial discovery of novel, valid , comprehensible and potentially useful patterns from data (Fayyad et. al).*
- *Focused on hypothesis generation, not on hypothesis testing*

What is Data Mining?

- *The term Data mining is a misnomer as it doesn't directly related to what is does.*
- *For exampling mining gold from rock is called Gold mining but not rock mining.*
- *Similarly oil mining is mining oil from the ground.*
- *Data mining should best describe as knowledge mining from* ***data rather that data mining***
- *Any way, we will use the term with this understanding*

What is Data Mining?

- ***Alternative names***
  - *Knowledge discovery(mining) from databases (KDD),*
  - *knowledge extraction,*
  - *data/pattern analysis,*
  - *data archeology,*
  - *data dredging,*
  - *information harvesting,*
  - *business intelligence, etc.*
- *Note that:*
  - *query processing systems, Expert statistical data analysis or Information retrieval systems are not data mining tasks*

What is Data Mining?

- *Sample pattern you might find*
  - *Supermarket data*
    - *On Thursday nights people who buy diapers also tend to buy beer*
  - *Insurance company data*
    - *People with good credit ratings are less likely to have accidents*
  - *Telecom data*
    - *Government lines are busy than private line*

Statistics vs. Data Mining

Statistics	Data Mining
Confirmative	Explorative
Small data sets/ File-based	Large data sets/ Databases
Small number of variables	Large number of variables
Deductive	Inductive
Numeric data	Numeric and non-numeric (including txt, networks)
Clean data	Data cleaning

Data Mining vs. Statistics

- *Statistics is known for:*
  - *well defined hypotheses used to learn about a topic*
  - *Work on specifically chosen population*
  - *Require carefully collected data for inferences well known properties.*
- *Data mining isn't that careful. It is:*
  - *data driven discovery of pattern*
  - *observational data sets is needed (data collected as side issue of other operations)*

Data Mining vs. Statistics

- ***Traditional statistics***
  - *first hypothesize, then collect data, then analyze*
  - *often model-oriented (strong parametric models)*
- ***Data mining:***
  - *few if any a priori hypotheses*
  - *data is usually already collected a priori*
  - *analysis is typically data-driven not hypothesis-driven*
  - *Often algorithm-oriented rather than model-oriented*

Challenges in Data Mining

- *Efficiency and scalability of data mining algorithms*
- *Parallel, distributed, stream, and incremental mining methods*
- *Handling high-dimensionality*
- *Handling noise, uncertainty, and incompleteness of data*
- *Incorporation of constraints, expert knowledge, and background knowledge*
- *Pattern evaluation and knowledge integration*
- *Mining diverse and heterogeneous kinds of data: e.g., bioinformatics, Web,*
- *Application-oriented and domain-specific data mining*
- *Invisible data mining (embedded in other functional modules)*
- *Protection of security, integrity, and privacy in data mining*

Potential Applications of Data Mining

- *Market analysis and management*
  - *target marketing analysis*
  - *Market basket analysis*
  - *Customer cross selling Analysis*
  - *customer purchase pattern analysis*
  - *Market segmentation*
- *Fraud detection and management*
- *etc*

Data source for DM applications

- *Where are the data sources for analysis?*
  - *Credit card transactions,*
  - *loyalty cards,*
  - *discount coupons,*
  - *customer complaint calls,*
  - *Customer calls*
  - *Log files*
  - *Transaction files etc…*

Market Basket Analysis

- *It is a processes of modeling item-set that consumers will put into his/her basket in one shopping*
- *This permits seller to arrange item-set so that consumers will find them easily*

Customer Cross-Selling Analysis

- *It is a processes of modeling item-set that consumers will purchase them at different time so that if customer buys item X them the business will recommend item Y which goes together*
- *This permits seller to maximize their profit, motivate their customers and improve their business strategy*

Target Market Analysis

- It is the process of identifying cluster of customers who will buy your service
- These customers share the same characteristics
- Target market analysis is the process of identifying (modeling) such groups of individuals

Market Segmentation

- It is the process of dividing the market into different homogeneous groups of consumers
- This better satisfy customers as they can choose the appropriate market for their need

Customer purchase pattern Analysis

- *It is the process of identifying the behaviors of consumers on their purchase pattern which includes*
  - *Why consumers make purchase and when?*
  - *What factors influence their purchase behavior*
- *This allows business to make selective promotion of good*

Fraud Detection and Management

- *Applications*
  - *widely used in health care, retail, credit card services, banking, insurance company, 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*
  - *medical insurance: detect professional patients and ring of doctors and ring of references*

Fraud Detection and Management

- *Detecting inappropriate medical treatment*
  - *Australian Health Insurance Commission identifies that in many cases blanket screening tests were requested (save Australian $1m/yr).*
- *Detecting telephone fraud*
  - *Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.*
  - *British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.*
- *Retail*
  - *Analysts estimate that 38% of retail shrink is due to dishonest employees.*

Data Mining: On What Kind of Data?

- *Relational databases*
- *Data warehouses*
- *Transactional databases*
- *Advanced DB and information repositories*
  - *Object-oriented and object-relational databases*
  - *Spatial databases*
  - *Time-series data and temporal data*
  - *Text databases and multimedia databases*
  - *Heterogeneous and legacy databases*
  - *WWW*

Data Mining Functionalities

- *Data mining can be performed on various types of data stores and Databases*
- *Data mining functionalities are used to specify the kind of patterns to be found in data mining task*
- *Data mining task can be broadly classified into two as*
  - *Descriptive*
  - *Predictive*

Data Mining Functionalities

- *Descriptive data mining task characterize the general properties of the data in a database.*
  - *For example one can say*
    - *Ethiopia's weather is selected to leave in for many birds*
    - *The past 10 years rainfall of Ethiopia is appropriate for the agriculturalist in southern Shewa*
    - *All mobile callers make few calls to wired lines than mobile receipents*

Data Mining Functionalities

- *Predictive data mining task perform inference on the current data in order to make prediction to the future reference*
- *For example one can say*
  - *A person loves to leave in Ethiopia if he/she was in ASIA for the last two years*
  - *It will rain in Addis with in two days if there is a wind from Mediterranean see in west - east direction and average current temperature at Addis is bellow 20^{o}c*

Data Mining Functionalities

- *The kind of pattern to be mined form a given data is not known for the user (hence it is hypothesis generation not hypothesis proving)*
- *Techniques should be implemented to extract various pattern from the available data so that user can choose what they need to use.*
- *There are different kinds of data mining functionalities that can be used to extract various types of pattern from data*

Data Mining Functionalities

- *This are*
  - *Concept /class description: Characterization and discrimination*
  - *Association Analysis*
  - *Classification and prediction*
  - *Clustering analysis*
  - *Outlier analysis*
  - *Evolution analysis*

Are All the "Discovered" Patterns Interesting?

- A data mining system/query may generate thousands of patterns, not all of them are interesting.
- Questions
  1. What makes a pattern interesting?
  2. Can a data mining system generate all of the interesting patterns?
  3. Can a data mining system generate only interesting patterns?

Question 1

1. What makes a pattern interesting?

- A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- An interesting pattern represents knowledge
- **Measure of Interestingness measures**
  - **Two types (Objective vs. subjective)**
    - Objective: based on statistics and structures of patterns, e.g., support, confidence, FP, FN, TN, TP, Recall, Precision, etc.
    - Subjective: based on user's belief in the data, e.g., unexpectedness (contradicting a user's belief), novelty, actionability, etc.

Question 2

2. Can a data mining system generate all of the interesting patterns?

- Referred as Completeness of the data mining algorithm
- No single data mining system is complete but users can set a constraint on the type of pattern they are looking for in which the data mining function generate all the pattern with the specified constraints
- Association algorithms don’t find classification pattern and others for example

Question 3

3. Can a data mining system generate only interesting patterns?

- This is an Optimization problem in data mining system
- it remain an challenging issue
- Usually data mining system generate pattern from the data set which may or may not relevant at the point
- So first generate all the patterns and then filter out the uninteresting ones.

Data Mining: Classification Schemes

- *Different views, different classifications*
  - *Kinds of databases to be mined*
  - *Kinds of knowledge to be discovered*
  - *Kinds of techniques utilized*
  - *Kinds of applications adapted*

A Multi-Dimensional View of Data Mining Classification

- ***Databases to be mined***
  - *Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.*
- ***Knowledge to be mined***
  - *Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.*
  - *Multiple/integrated functions and mining at multiple levels*
- ***Techniques utilized***
  - *Database-oriented, data warehouse (OLAP) oriented, machine learning, statistics, visualization, neural network, etc.*
- ***Applications adapted***
  - *Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.*