# Business Intelligence (BI) and Data Warehousing (DW)

## Data Mining

Jackline Ssanyu

February 28, 2020

# Definition

- The purpose of data analysis is to discover previously unknown data characteristics, relationships, dependencies, or trends.
- Such discoveries then become part of the information framework on which decisions are built.
- Data mining refers to the **activities that analyze the data**, **uncover problems or opportunities hidden in the data relationships**, **form computer models** based on their findings, and then **use the models to predict** business behavior—requiring minimal end-user intervention.
- Therefore, the end user is able to use the system's findings to gain knowledge that might yield competitive advantages.

# Why Data Mining?

Why Data mining

- OLAP can only provide shallow data analysis which is not sufficient to support business decisions.
- E.g OLAP can't answer the following queries:
  - how to boost sales of other products
  - when people buy product say 6 what other products do they or are likely to buy? cross selling
- OLAP is based on SQL
  - E.g SELECT PRODUCTS.PNAME, SUM(SALESFACTS.SALES AMT) FROM DBSR.PRODUCTS PRODUCTS, DBSR.SALESFACTS SALESFACTS WHERE ( ( PRODUCTS.PRODUCT KEY = SALESFACTS.PRODUCT KEY ) ) GROUP BY PRODUCTS.PNAME;
  - The nature of SQL decides that complicated algorithm cannot be implemented with SQL.
  - Complicated algorithms need to be developed to support deep data analysis - i.e. data mining

# Why data mining....

- OLAP results generated from data sets with large number of attributes are difficult to be interpreted. E.g Pick three attributes related to a customer: income level, education level and sales amount

# OLAP Vs Data mining

Both OLAP and data mining are important analytical technologies in the business intelligence family.

- OLAP is good at aggregating of large amount of transaction data based on the dimension definitions. The typical questions answered by OLAP are:
  - What is the total sales amount of beverage products in the past three months in the Northwest region?
  - What are the top 10 products sold in all stores last month?
  - What are the store sales for male and female customers, respectively?
  - What is the daily sales difference during a promotional period versus a normal period?

# OLAP Vs Data mining....

Data mining is good at finding the hidden patterns of a dataset by analyzing correlations among attribute values. The following are typical questions answered by data mining:

- What is the profile of customers who like to buy the newest model digital cameras?
- What are the products to recommend to this particular customer?
- What's the estimated sales amount for digital cameras in the next three months?
- How should I segment the customer base?

# General phases of data mining

- Data preparation - the main data sets to be used by the data-mining operation are identified and cleansed of any data impurities. Because the data in the data warehouse are already integrated and filtered, the data warehouse usually is the target set for data-mining operations.
- Data analysis and classification - studies the data to identify common data characteristics or patterns. During this phase, the data-mining tool applies specific algorithms to find:
  - Data groupings, classifications, clusters, or sequences.
  - Data dependencies, links, or relationships.
  - Data patterns, trends, and deviations.

# General phases of data mining...

- Knowledge acquisition - uses the results of the data analysis and classification phase. In this phase a data mining tool uses a suitable algorithm to generate a computer model that reflects the behavior of the target data set.
- Prognosis - the data-mining findings are used to predict future behavior and forecast business outcomes. Examples of data-mining findings can be:
  - Sixty-five percent of customers who did not use a particular credit card in the last six months are 88 percent likely to cancel that account.
  - Eighty-two percent of customers who bought a 42-inch or larger LCD TV are 90 percent likely to buy an entertainment center within the next four weeks.
  - If age $< 30$ and income $<= 25{,}000$ and credit rating $< 3$ and credit amount $> 25{,}000$, then the minimum loan term is 10 years.

# Data Mining Tasks

Data Mining is based on algorithms that can discover hidden patterns. It is interactive, not fully automated. Major data mining tasks include:

- Summarization
  - Summarization is the generalization of data. A set of relevant data is summarized which result in a smaller set that gives aggregated information of the data.
  - For example, the shopping done by a customer can be summarized into total products, total spending, offers used, etc.
  - Such high level summarized information can be useful for sales or customer relationship team for detailed customer and purchase behavior analysis. Data can be summarized in different abstraction levels and from different angles.

# Data Mining Tasks....

- Classification
  - Classification derives a model to determine the class of an object based on its attributes.
  - Classification can be used in direct marketing, that is to reduce marketing costs by targeting a set of customers who are likely to buy a new product.
  - Using the available data, it is possible to know which customers purchased similar products and who did not purchase in the past.
  - Hence, purchase, don't purchase decision forms the class attribute in this case.
  - Once the class attribute is assigned, demographic and lifestyle information of customers who purchased similar products can be collected and promotion mails can be sent to them directly.

# Data Mining Tasks....

- Association rule mining
  - Association discovers the association or connection among a set of items.
  - Association identifies the relationships between objects.
  - Association analysis is used for commodity management, advertising, catalog design, direct marketing etc.
  - A retailer can identify the products that normally customers purchase together or even find the customers who respond to the promotion of same kind of products.
  - If a retailer finds that soft drinks and chips are bought together mostly, he can put chips on sale to promote the sale of soft drinks.

# Data Mining Tasks....

- Clustering
  - Clustering is used to identify data objects that are similar to one another.
  - The similarity can be decided based on a number of factors like purchase behavior, responsiveness to certain actions, geographical locations and so on.
  - For example, an insurance company can cluster its customers based on age, residence, income etc.
  - This group information will be helpful to understand the customers better and hence provide better customized services.

# Data Mining Tasks....

- Prediction
  - Prediction task predicts the possible values of missing or future data.
  - Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest.
  - For example, a model can predict the income of an employee based on education, experience and other demographic factors like place of stay, gender etc.
  - Also prediction analysis is used in different areas including medical diagnosis, fraud detection etc.

# Data Mining Tasks....

- Time Series Analysis
  - Consists of sequences of values or events obtained over repeated measurements of time (weekly, hourly).
  - E.g Stock market analysis, economic and sales forecasting, scientific and engineering experiments, medical treatments etc.

# Applications of data mining

Some of the applications of data mining are:

- Market basket analysis
  - is a modeling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items.
  - This technique may allow the retailer to understand the purchase behavior of a buyer.
  - This information may help the retailer to know the buyers needs and change the stores layout accordingly.
  - Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

# Applications of data mining.....

- Educational Data Mining (EDM)
  - Is concerned with developing methods that discover knowledge from data originating from educational Environments.
  - The goals of EDM are identified as predicting students future learning behavior, studying the effects of educational support, and advancing scientific knowledge about learning.
  - Data mining can be used by an institution to take accurate decisions and also to predict the results of the student.
  - With the results the institution can focus on what to teach and how to teach.
  - Learning pattern of the students can be captured and used to develop techniques to teach them.

# Applications of data mining.....

- Intrusion Detection
  - Any action that will compromise the integrity and confidentiality of a resource is an intrusion.
  - The defensive measures to avoid an intrusion includes user authentication, avoid programming errors, and information protection.
  - Data mining can help improve intrusion detection by adding a level of focus to anomaly detection.
  - It helps an analyst to distinguish an intrusion activity from common everyday network activity.
  - Data mining also helps extract data which is more relevant to the intrusion problem.

# Applications of data mining.....

- Lie Detection
  - Apprehending a criminal is easy whereas bringing out the truth from him is difficult.
  - Law enforcement can use mining techniques to investigate crimes, monitor communication of suspected terrorists.
  - This field includes text mining also. This process seeks to find meaningful patterns in data which is usually unstructured text. The data sample collected from previous investigations are compared and a model for lie detection is created.

# Applications of data mining.....

- Customer Segmentation
  - Traditional market research may help us to segment customers but data mining goes in deep and increases market effectiveness.
  - Data mining aids in aligning the customers into a distinct segment and can tailor the needs according to the customers.
  - Market is always about retaining the customers. Data mining allows to find a segment of customers based on vulnerability and the business could offer them with special offers and enhance satisfaction.

# Questions

- Explain why data mining is important.
- Describe four applications of data mining other than those presented in the slides.
- Briefly discuss the major difference between Classification and Clustering. List one real application for each of them respectively.