Data Warehousing
&
Data Mining

IT434

Lab Instructors

- Ms. Wejdan Alkaldi<walkaldi@ksu.edu.sa>
- Ms. Sumayah Al-Rabiaah <salrabiaah@ksu.edu.sa>
- Ms. Weam AlRashed

- Note: when you email me, please insert [IT434] in the subject line and include your REAL name, otherwise I will not answer your email.

Grading

Lab Work (30%)

- Quizzes - 10%
- Assignments - 5%
- Project - 15% :
  - 1- software - 15%
  - 2- or paper- %10 + small software- 5%

Question 1.1 page 40

1.1. What is *data mining? In your answer, address the following:*

(a) Is it another hype?

(b) Is it a simple transformation of technology developed from databases, statistics, and machine learning?

(c) Explain how the evolution of database technology led to data mining.

(d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

Answer of Question 1.1 page 40

Data mining refers to the process or method that extracts or \mines" interesting knowledge or patterns from large amounts of data.

(a) Is it another hype?

Data mining is not another hype. Instead, the need for data mining has arisen due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Thus, data mining can be viewed as the result of the natural evolution of information technology.

Answer of Question 1.1 page 40 *(cont.)*

(b) Is it a simple transformation of technology developed from databases, statistics, and machine learning?

No. Data mining is more than a simple transformation of technology developed from databases, statistics, and machine learning. Instead, data mining involves an integration, rather than a simple transformation, of techniques from multiple disciplines such as database technology, statistics, ma chine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis.

Answer of Question 1.1 page 40 *(cont.)*

(c) Explain how the evolution of database technology led to data mining.
Database technology began with the development of data collection and database creation mechanisms that led to the development of effective mechanisms for data management including data storage and retrieval, and query and transaction processing. The large number of database systems offering query and transaction processing eventually and naturally led to the need for data analysis and understanding.
Hence, data mining began its development out of this necessity.

Answer of Question 1.1 page 40 *(cont.)*

(d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

The steps involved in data mining when viewed as a process of knowledge discovery are as follows:

-**Data cleaning**, a process that removes or transforms noise and inconsistent data

-**Data integration**, where multiple data sources may be combined

-**Data selection**, where data relevant to the analysis task are retrieved from the database

-**Data transformation**, where data are transformed or consolidated into forms appropriate for mining.

- **Data mining**, an essential process where intelligent and efficient methods are applied in order to extract patterns

- **Pattern evaluation**, a process that identifies the truly interesting patterns representing knowledge based on some interestingness measures

-**Knowledge presentation**, where visualization and knowledge representation techniques are used to present the mined knowledge to the user

-See Figure 1.4 (Data mining as a step in the process of knowledge discovery.) page#6

Question 1.2 page 41

1.2. Present an example where data mining is crucial to the success of a business. What *data mining functions* does this business need? Can they be performed alternatively by data query processing or simple statistical analysis?

Answer of Question 1.2 page 41

1.2. Present an example where data mining is crucial to the success of a business. What *data mining functions* does this business need? Can they be performed alternatively by data query processing or simple statistical analysis?

Answer:

A department store, for example, can use data mining to assist with its target marketing mail campaign.

Using data mining functions such as association, the store can use the mined strong association rules to determine which products bought by one group of customers are likely to lead to the buying of certain other products. With this information, the store can then mail marketing materials only to those kinds of customers who exhibit a high likelihood of purchasing additional products. Data query processing is used for data or information retrieval and does not have the means for finding association rules. Similarly, simple statistical analysis cannot handle large amounts of data such as those of customer records in a department store.

Question 1.4 page 41

1.4 How is a *data warehouse different from a database? How are they similar?*

Answer of Question 1.4 page 41

1.4. How is a *data warehouse different from a database? How are they similar?*

Answer:

- **Differences between a data warehouse and a database:** *A data warehouse is a repository of informa*tion collected from multiple sources, over a history of time, stored under a unified schema, and used for data analysis and decision support; whereas a database, is a collection of interrelated data that represents the current status of the stored data. There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another. A database system supports ad-hoc query and on-line transaction processing. Additional differences are detailed in Section 3.1.1 (Differences between Operational Databases Systems and Data Warehouses).

- **Similarities between a data warehouse and a database:** *Both are repositories of information, storing* huge amounts of persistent data.

Question 1.5 page 41

1.5 Briefly describe the following advanced database systems and applications: object relational databases, spatial databases, text databases, multimedia databases, stream data,
The World Wide Web.

Answer of Question 1.5 page 41

1.5. Briefly describe the following *advanced database systems and applications: object relational databases,* spatial databases, text databases, multimedia databases, the World Wide Web.

Answer:

*-**An objected-oriented database** is designed based on the object-oriented programming paradigm* where data are a large number of objects organized into classes and class hierarchies. Each entity in the database is considered as an object. The object contains a set of variables that describe the object, a set of messages that the object can use to communicate with other objects or with the rest of the database system, and a set of methods where each method holds the code to implement a message.

*- **A spatial database** contains spatial-related data, which may be represented in the form of raster* or vector data. Raster data consists of *n-dimensional bit maps or pixel maps, and vector data are* represented by lines, points, polygons or other kinds of processed primitives, Some examples of spatial databases include geographical (map) databases, VLSI chip designs, and medical and satellite images databases.

Answer of Question 1.5 page 41 *(cont.)*

-***A text database*** *is a database that contains text documents or other word descriptions in the form of* long sentences or paragraphs, such as product specifications, error or bug reports, warning messages, summary reports, notes, or other documents.

-A multimedia database *stores images, audio, and video data, and is used in applications such as* picture content-based retrieval, voice-mail systems, video-on-demand systems, the World Wide Web, and speech-based user interfaces.

-The World Wide Web *provides rich, world-wide, on-line information services, where data objects* are linked together to facilitate interactive access. Some examples of distributed information services associated with the World Wide Web include America Online, Yahoo!, AltaVista, and Prodigy.

Question 1.6 page 41

1.6 Define each of the following data mining functionalities:
characterization, discrimination, association and correlation analysis, classification, prediction, clustering, and evolution analysis.
Give examples of each data mining functionality, using a real-life database with which you are familiar.

Answer of Question 1.6 page 41

Answer:

- **Characterization** is a summarization of the general characteristics or features of a target class of data. For example, the characteristics of students can be produced, generating a profile of all the University first year computing science students, which may include such information as a high GPA and large number of courses taken.
- **Discrimination** is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. For example, the general features of students with high GPA's may be compared with the general features of students with low GPA's. The resulting description could be a general comparative profile of the students such as 75% of the students with high GPA's are fourth-year computing science students while 65% of the students with low GPA's are not.
-**Association** is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. For example, a data mining system may find association rules like

$$major(X, \text{“computing science””}) \Rightarrow owns(X, \text{“personal computer”}) \quad [support = 12\%, confidence = 98\%]$$

where *X is a variable representing a student. The rule indicates that of the students under study,* 12% (**support**) major in computing science and own a personal computer. There is a 98% probability (**confidenc**e, or certainty) that a student in this group owns a personal computer.

Answer of Question 1.6 page 41 *(cont.)*

-**Classification** differs from **prediction** in that the former constructs a set of models (or functions) that describe and distinguish data classes or concepts, whereas the latter builds a model to predict some missing or unavailable, and often numerical, data values. Their similarity is that they are both tools for prediction: Classification is used for predicting the class label of data objects and prediction is typically used for predicting missing numerical data values.

-**Clustering** analyzes data objects without consulting a known class label. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. Each cluster that is formed can be viewed as a class of objects. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

-Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Question 1.14 page 42

1.14 Describe three challenges to data mining regarding data mining methodology and user interaction issues.

Answer of Question 1.14 page 42

Answer:

Challenges to data mining regarding data mining methodology and user interaction issues include the following: mining different kinds of knowledge in databases, interactive mining of knowledge at multiple levels of abstraction, incorporation of background knowledge, data mining query languages and ad hoc data mining, presentation and visualization of data mining results, handling noisy or incomplete data, and pattern evaluation. Below are the descriptions of the first three challenges mentioned:

- ***Mining different kinds of knowledge in databases****: Different users are interested in different kinds* of knowledge and will require a wide range of data analysis and knowledge discovery tasks such as data characterization, discrimination, association, classification, clustering, trend and deviation analysis, and similarity analysis. Each of these tasks will use the same database in different ways and will require different data mining techniques.

Answer of Question 1.14 page 42 *(cont.)*

-Interactive mining of knowledge at multiple levels of abstraction*: Interactive mining, with* the use of OLAP operations on a data cube, allows users to focus the search for patterns, providing and refining data mining requests based on returned results. The user can then interactively view the data and discover patterns at multiple granularities and from different angles.

- Incorporation of background knowledge*: Background knowledge, or information regarding the* domain under study such as integrity constraints and deduction rules, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction. This helps to focus and speed up a data mining process or judge the interestingness of discovered patterns.