

Computer Science COMP3420 (2010) – Tutorial One

Question 1. Among the three popular databases, the hierarchical database, the network database and the relational database, why has the relational database become the dominant database in the commercial world?

Answer: Relational algebra and the entity-relationship model (ER model) are the theoretical foundation of the relational database, while the simplicity and powerful expression ability of the relational query language SQL makes the relational database become a popular database.

Question 2. How is a data warehouse different from a database? How are they similar?

Answer: The data in a data warehouse is historical, consolidated data (e.g., past 5-10 years), the operations on the data are summarizations, aggregations, and multidimensional data analysis. The queries in DW are usually ad hoc nature and very complicated. DW serves for data analysis and decision making support by providing aggregated, summarized data. Thus, data warehousing and OLAP are an essential step toward the knowledge discovery process.

Operational databases deal with the current data, on which the frequent operation is the OLTP. An operational database is designed and tuned for known tasks and workloads such as indexing and hashing using primary keys, searching for particular records, and optimizing “canned” queries. It also supports the concurrent processing of multiple transactions. Concurrency control and recovery mechanisms are required to ensure the consistency and robustness of transactions.

There are some similarities between them. They both are used for data management that means that they both perform data analysis and query processing. The operational databases provide source data for data warehouses, while the latter can carry some high-level complicated summarization and analysis on the huge volume of data.

Question 3. What is the difference between discrimination and classification? between characterization and clustering? Between classification and prediction? For each of these pairs of tasks, how are they similar?

Answer: Data 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, while data classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

Data characterization is a summarization of the general characteristics of features of a target class of data. Unlike classification and prediction which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label.

Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process. These attributes can then be excluded.

Question 4. What are the major challenges of data mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set)?

Answer: The major challenges are high scalability and efficiency of the mining techniques.

Question 5. What is a data warehouse? What is data mining?

Answer: A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management decision-making process. Data mining is automatically extraction of interesting (non-trivial, implicit, previously unknown, and potentially useful) patterns or knowledge from huge amount of data.

Question 6. State why, for the integration of multiple heterogeneous information sources, many companies in industry prefer the *update-driven approach* which constructs and uses data warehouses, rather than the *query-driven approach* which applies wrappers and integrator. Describe a situation where the query-driven approach is preferable over the update-driven approach.

Answer: Data integration takes enormous workloads and time, the data in a data warehouse is historical and consolidated data. Rare updating happens for such data. If there is such data updating, the updated data volume is significantly small, compared with the entire data volume. Thus, the update driven approach (refreshment) is adopted during the data warehouse construction and usage. For operational database, the data in it is dynamically changed daily. Thus, the query driven approach is preferable to get a query result from such a data base.

Question 7. Can you briefly describe the four stages of knowledge discovery (KDD)? Can you describe the multi-tiered data warehouse architecture?

Answer:

- Stage one: Identify the heterogeneous data sources, data integration and transformation.
- Stage two: Consolidated, summarized or aggregated data have been stored in data warehouse.
- Stage three: Perform OLAP and OLMP operations to mine interesting knowledge.
- Stage four: Present the results to the decision-makers (consumers).

The multi-tiered architecture of DW consists of data source layer, data storage layer, OLAP engine layer, and the front-end tools layer.

Question 8. Briefly describe the five major tasks in data preprocessing. What are the main objectives of each of these tasks?

Answer:

- Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration: integration of multiple databases, data cubes, or files
- Data transformation: normalization and aggregation
- Data reduction: obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization: part of data reduction but with particular importance, especially for numerical data.

Question 9. Can you list the three measures used for measuring data central tendency? Give an example for each of the measures.

Answer: Measure of central tendency includes mean, median and mode.

Let x_1, x_2, \dots, x_n be a set of n values or observations. The **mean** of this set of values is $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$.

If n is odd, the **median** is the middle value of the ordered set; otherwise, the median is the average of the middle two values.

The **mode** for a set of data is the value that occurs frequently in the set.

Question 10. Given a set of data $\{12, 8, 45, 23, 6, 7, 8, 25, 34, 22, 19, 22, 34, 45, 41\}$,

(a) Can you employ the binning method to allocate the data in the set into 3 bins by equal-frequency, smoothing by bin means and smoothing by bin boundaries respectively?

(b) Can you use the histogram method to allocate the data in the set to 4 equal intervals?

Answer: We sort the data in increasing order first. The result is:

6, 7, 8, 8, 12, 19, 22, 22, 23, 25, 34, 34, 41, 45, 45.

We allocate the data into 3 bins as follows.

- Equal-frequency
 - Bin 1: 6, 7, 8, 8, 12
 - Bin 2: 19, 22, 22, 23, 25
 - Bin 3: 34, 34, 41, 45, 45
- Smoothing by bin means
 - Bin 1: 8, 8, 8, 8, 8
 - Bin 2: 22, 22, 22, 22, 22
 - Bin 3: 40, 40, 40, 40, 40
- Smoothing by bin boundaries
 - Bin 1: 6, 6, 6, 6, 12
 - Bin 2: 19, 19, 19, 25, 25
 - Bin 3: 34, 34, 45, 45, 45

(b) The interval width is $W = \frac{\text{maximum} - \text{minimum}}{4} = \frac{45 - 6}{4} \approx 9$. Thus, we have 4 intervals, which are $[6, 15]$, $[16, 25]$, $[26, 35]$, $[36, 45]$. The frequencies of data in these intervals are 5, 5, 2, 3. The means of data in the intervals are 8, 22, 34, and 43 respectively.

Question 11. What are the main metrics for measuring data dispersion? What are the main tasks of data cleaning? Briefly describe several approaches to remove data noise.

Answer:

- Quartiles: Q1 (25th percentile), Q3 (75th percentile),
- Inter-Quartile Range: $IQR = Q3 - Q1$
- Five number summary: min, Q1, Median, Q3, max
- Boxplot: Ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
- Outlier: usually, a value higher/lower than $1.5 \times IQR$, etc.

Data cleaning tasks consist of filling in missing values, identifying outliers and smoothing out noisy data, correcting inconsistent data, and resolving redundancy caused by data integration.

The approaches for removing data noise are as follows.

- Binning: first sort data and partition into (equal-frequency) bins then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression: smooth by fitting the data into regression functions
- Clustering: detect and remove outliers
- Combined computer and human inspection: detect suspicious values and check by human (e.g., deal with possible outliers).

Question 12. Can you point out how to perform data transformation during building a data warehouse?

Answer: The typical tools used for data transformations are as follows.

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
 - min-max normalization
 - z -score normalization
 - normalization by decimal scaling
- Attribute/feature construction.

Question 13. Outlines the major research challenges of data mining in one specific applications domain such as stream/sensor data analysis, or bioinformatics.

Answer: Consider within a data stream/sensory data environment, there are the huge amounts of data to be processed, not all the data can be stored and processed, there is a time dimension, the data can be scanned only once. Thus, effective and efficient management and analysis of stream data pose great challenges to researchers, for example, which data structure can be used to keep the unpredicted patterns or frequent items when scanning the original data. Mining data streams involves the efficient discovery of general patterns and dynamic changes within stream data.

Question 14. What is the difference between OLAP (data warehouses) and OLTP (operational databases)?

Answer: OLTP (on-line transaction processing) is one of the major operations in traditional relational DBMS, it can be used to deal with day-to-day operations, purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.

OLAP (on-line analytical processing) is the major analysis engine in data warehousing. It deals with multidimensional data analysis and provides the support to enterprise strategic decision making.

Distinct features (OLTP vs. OLAP):

- User and system orientation: customer vs. market data contents; current, detailed vs. historical, consolidated
- Database design: ER + application vs. star + subject
- View: current, local vs. evolutionary, integrated
- Access patterns: update vs. read-only but complex queries.

Question 15. What are the differences between the three main types of data warehouse usages: *information processing*, *analytical processing* and *data mining*? Discuss the motivation behind OLAP mining (OLAM).

Answer: Information processing supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts, or graphs. A current trend in data warehouses information processing is to construct low-cost Web-based accessing tools that are then integrated with Web browsers.

Analytical processing supports basic OLAP operations, including slide-and-dice drill-down, roll-up, and pivoting. It generally operates on historical data in both summarized and detailed forms. The major strength of on-line analytical processing over information processing is the multidimensional data analysis of data warehouse data.

Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

The motivation behind of OLMA is in the following.

- High quality of data in data warehouses
- Available information processing infrastructure surrounding data warehouses
- OLAP-based exploratory data analysis
- On-line selection of data mining functions.