Databases and Data Mining

Assignment 5

11-1 2010

Due: Friday 5-2 2010
Grading: This assignment will be graded from 0 to 10.
Notes:

- Every student has to make this assignment individually.
- Every question has the same weight.
- Always explain your answers carefully.
- Write down your answers for this assignment in a *.pdf* file with the following name ***"<your student number><your last name>_5.pdf"***, e.g., ***"012345jansen_5.pdf"***.
- Send this *.pdf* file as an attachment of an e-mail with subject **"DBDM_5"** to erwin@liacs.nl.

Questions

1. (From the book: 2.19) Propose an algorithm, in pseudo-code, for the following:
   a) The automatic generation of a concept hierarchy for numerical data based on the *equal-width* partitioning rule.
   b) The automatic generation of a concept hierarchy for numerical data based on the *equal-frequency* partitioning rule.

2. (From the book: 3.5) Suppose that a data warehouse consists of four dimensions, *date*, *spectator*, *location*, and *game*, and two measures, *count*, and *charge*, where *charge* is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.
   a) Draw a *star schema* diagram for the data warehouse.
   b) Starting with the base cuboid [*date, spectator, location, game*], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM_Place 2004?
   c) *Bitmap indexing* is useful in data warehousing. Taking this cube as an example, briefly discuss advantages and problems of using a bitmap index structure.

3. (From the book: 3.13 (partly)) A data cube, *C*, has *n* dimensions, and each dimension has exactly *p* distinct values in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions. What is the *maximum number of cells* possible (including both base cells and aggregate cells) in the data cube, *C*?

4. (From the book: 5.1 (partly)) The *Apriori algorithm* uses *prior knowledge* of subset support properties.
   a) Prove that all nonempty subsets of a frequent itemset must also be frequent.
   b) Given frequent itemset *l* and subset s of *l*, prove that the confidence of the rule *"s' => (l-s')"* cannot be more than the confidence of the rule *"s => (l – s)"* where *s'* is a subset of *s*.

5. An optimization in frequent item set mining is mining closed patterns, or mining max patterns instead. Describe the main differences of mining closed patterns and mining max patterns.

6. (From the book: 5.20 (partly)) The price of each item in a store is nonnegative. For each of the following cases, identify the kinds of constraint they represent (e.g. *antimonotonic, monotonic, succinct*) and briefly discuss how to mine such association rules efficiently:
   a) Containing one free item and other items the sum of whose prices is at least $190.
   b) Where the average price of all the items is between $120 and $520.

7. Suppose a city has installed hundreds of surveillance cameras at strategic locations in busy streets. All the captured video (each stream being 640 pixels x 480 pixels, 24-bits (RGB) per pixel, 25fps) is streamed to a central observation post in real time. Describe an efficient system that would allow real time data mining and continuously querying these video streams for abnormal events, such as *explosions*, *people in panic*, etc.. Also discuss the computational costs and the memory requirements of your system.

8. (From the book.) A flight data warehouse for a travel agent consists of six dimensions: *traveler, departure (city), departure_time, arrival (city), arrival_time,* and *flight*; and two measures *count*, and *avg_fare*, where *avg_fare* stores the concrete fare at the lowest level but the average fare at the other levels. Suppose the cube is fully materialized. Starting with the base cuboid [*traveler, departure, departure_time, arrival, arrival_time, flight*], what specific OLAP operations (e.g roll-up flight to airline, etc.) should one perform in order to list the average fare per month for each business traveler who flies American Airlines (AA) from Los Angeles (LA) in the year 2007?

9. In graph mining what would be the advantage of the described apriori-based approach over the pattern growth based approach (see lecture slides) and vice versa.