UIUC-CS412 "An Introduction to Data Warehousing and Data Mining" (Fall 2011)

Midterm Exam

(Wednesday, Oct. 19, 2011, 90 minutes, 100 marks, single sheet reference, brief answers)

Name: NetID: Score:

1. [30] Data preprocessing.

   (a) [10] For data visualization, we have learned four classes of techniques: (i) pixel-oriented techniques, (ii) geometric projection techniques, (iii) hierarchical techniques, and (iv) icon-based techniques. For each of the following methods, state which category it belongs to.

      i. stick figure

      ii. parallel coordinates

      iii. dimensional stacking

      iv. scatterplot matrices

      v. 3-D cone trees

(b) [6] What are the best distance measure for each of the following applications:

i. Find whether two text documents are similar

ii. Find the maximum difference between any attribute of two vectors

iii. Find whether a numerical value is likely an outlier within a set of 100 numerical values

(c) [6] What are the value ranges of the following measures, respectively?

i. Jaccard coefficient

ii. Lift

iii. covariance

(d) [8] Briefly state the key difference between each of the following pairs of concepts?

i. *correlation analysis by χ^2 test* vs. *correlation analysis with Pearson correlation coefficient*

ii. *histograms* vs. *boxplot*

iii. *Fourier transform* vs. *wavelet transform*

iv. *stratified sampling* vs. *simple sampling*

2. [20] Data Warehousing and OLAP for Data Mining

(a) [10] Suppose the base cuboid of a data cube contains three cells

$(a_1, a_2, a_3, a_4, \ldots, a_{20})$, $(a_1, b_2, a_3, b_4, \ldots, b_{20})$, $(c_1, a_2, c_3, a_4, \ldots, a_{20})$

where $a_i \neq b_i \neq c_i$ if i for any i

i. How many nonempty cuboids are there in this data cube?

ii. How many (nonempty) aggregate closed cells are there in this data cube?

iii. How many (nonempty) aggregate cells are there in this data cube?

iv. If we set minimum support = 2, how many (nonempty) aggregate cells are there in the corresponding iceberg cube?

(b) [5] Suppose an Amazon data cube uses *standard deviation* to measure the dispersion of the sales of its commodities. Explain how this measure in the cube can be incrementally updated, when a new batch of base data set D is added in.

Hint: The **standard deviation** of n observations $x_1, x_2, \ldots, x_n$ is defined as

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2]}. \quad (0.1)$$

where $\bar{x}$ is the average (*i.e.*, mean) value of $x_1, \ldots, x_n$.

(c) [5] Suppose a large data relation about products contains 50 attributes and 10^6 records. What is the minimum number of database scans in order to derive a generalized relation by *attribute-oriented induction* if one wants to compare all the products made in two countries? and why?

3. [20] Data cube implementation

(a) [7] Suppose a data relation has 99 attributes and 10^6 tuples. Each attribute has 100 distinct values. If each cell takes 16 bytes of space, what is the total size (in bytes) of the pre-computed shell-fragments of size 3?

(b) [6] Which of the following three algorithms: (i) BUC; (ii) Multiway array cubing; and (iii) StarCubing, cannot support iceberg cube computation efficiently? Why not?

(c) [7] Suppose a car dealer implements a ranking cube to help customers search desired cars, such as Toyota cars with around 50K mileage and around $5000. Outline how the cube should be structured and how it supports such kind of queries efficiently. (Suppose the dealer has the following information about cars: make, model, year, price, and mileage).

4. [27] Frequent pattern and association mining.

   (a) [9] A database with 150 transactions has its FP-tree shown in Fig. 1. Let $min\_sup = 0.5$ and $min\_conf = 0.8$.

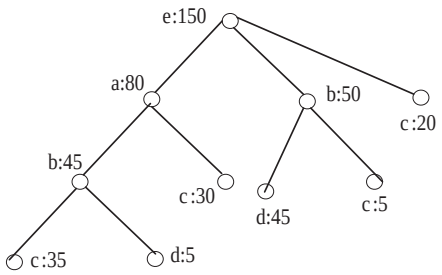

Figure 1: FP tree of a transaction DB

   i. Show c's conditional (i.e., projected) database:

   ii. Present all the frequent k-itemsets for the largest k:

   iii. Present two strong association rules (with support and confidence) containing the k items (for the largest k only):

(b) [6] Suppose one chain-store has n stores in a country. Describe an efficient Apriori-based **distributed** frequent-itemset mining method which mines global frequent itemsets, without transmitting all the transactional data to one site.

(c) [6] Suppose a transaction database contains N transactions, ct transactions contain both coffee and tea, $c\bar{t}$ transactions contain coffee but not tea, $\overline{c}t$ transactions contain tea but not coffee, and $\overline{c}\bar{t}$ transactions contain neither tea nor coffee. At what condition that Lift may not be a good measure to indicate correlations between coffee and tea? and why?

(d) [6] Suppose a WalMart manager is interested in only the *frequent patterns* (i.e., *itemsets*) that satisfy certain constraints. For the following cases, state the characteristics (*i.e.*, categories) of *every constraint* in each case and how to mine such patterns most efficiently.

i. The average price of all the items in each pattern is greater than \$40.

ii. The sum of the price of all the items with profit over \$5 in each pattern is at least \$100.

5. [3] (Opinion).

(a) I ☐ like ☐ dislike the exams in this style.

(b) In general, the exam questions are ☐ too hard ☐ too easy ☐ just right.

(c) I ☐ have plenty of time ☐ have just enough time ☐ do not have enough time to finish the exam questions.