CS412 "An Introduction to Data Warehousing and Data Mining" (Fall 20)**

Midterm Exam

(Wednesday, Oct. **, 20**, 90 minutes, 100 marks, single sheet reference, brief answers)

Name: NetID: Score:

1. [35] Data and data preprocessing.

   (a) [6] For data visualization, there are three classes of techniques: (i) geometric techniques, (ii) hierarchical techniques, and (iii) icon-based techniques. Give names of two methods for each of these techniques.

   (b) [6] What are the value ranges of the following measures, respectively?
   i. χ^2:
   ii. Jaccard coefficient:
   iii. *covariance*:

   (c) [8] Name four methods that perform effective *dimensionality reduction* and four methods that perform effective *numerosity reduction*.

   (d) [6] What are the best distance measure for each of the following applications:
   (i) compare similar diseases with a set of medical tests
   (ii) find whether two text documents are similar
   (iii) the maximum difference between any attribute of two vectors

   (e) [9] For the following group of data

   $$100, 400, 1000, 500, 2000$$

   i. Calculate its mean and variance.
   ii. Normalize the above group of data by min-max normalization with min = 1 and max = 10; and
   iii. In z-score normalization, what value should the first number 100 be transformed to?

2. [17] Data Warehousing and OLAP for Data Mining

   (a) [7] Suppose a base cuboid has D dimensions but contains only p (where $p > 1$) nonempty cells
   (i) *how many cuboids* does this cube contain (including base and apex cuboids)?
   (ii) what is the *maximum number of nonempty cells possible* in such a materialized cube?

   (b) [5] Suppose a WalMart data cube takes sum, mean, and standard deviation to measure the sales of its commodities. Explain how the three measures of 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 disk-based large relation contains 30 attributes. What is the minimum number of database scans in order to derive a generalized relation by *attribute-oriented induction*?

3. [22] Data cube technology

(a) [10] Given the following four methods: *multiway array cubing* (Zhao, et al. SIGMOD'1997), *BUC* (bottom-up computation) (Beyer and Ramakrishnan, SIGMOD'2001), *StarCubing* (Xin et al., VLDB'2003), and *shell-fragment* approach (Li et al, VLDB'2004), list one method which is the best and another which is the worst (or not working) to implement one of the following:

(a) computing a dense full cube of low dimensionality (e.g., less than 6 dimensions),

(b) computing a large iceberg cube of around 8 dimensions, and

(c) performing OLAP operations in a high-dimensional database (e.g., over 60 dimensions).

(b) [6] Suppose a (sampling) survey data sets contains 100 dimensions (variables), but people would still like to perform multidimensional drilling into the cells containing no or few data and examine the statistics (measures) of the cell. Outline a method that may implement such a mechanism effectively.

(c) [6] Suppose one would like to implement a web-based search engine to return top-k best deals for used cars with user selected multidimensional features, such as model, year, price-range, etc., where the best deal is a user-defined function of book-price and sales-price. Outline the design of a cube structure to support such a search engine.

4. [23] Frequent pattern and association mining.

(a) [10] A database with 100 transactions has its FP-tree shown in Fig. 1. Let $min\_sup = 0.5$ and $min\_conf = 0.8$. Show

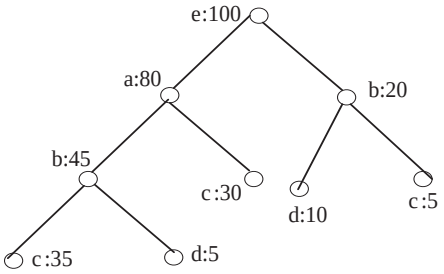

Figure 1: FP tree of a transaction DB

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

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

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

(b) [6] Briefly describe one efficient **distributed** pattern growth mining method which can mine enterprise-wide (*i.e.*, global) frequent itemsets for a chain store like Sears, without shipping data to one site.

(c) [7] It is important to use a good measure to check whether two items in a large transaction dataset are strongly correlated.

(i) Give one example to show that *lift* may not be a good measure for such a purpose.

(ii) Give a good measure for this purpose and reason why it is a good measure.

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.