UIUC-CS412 "An Introduction to Data Warehousing and Data Mining" (Fall 2010)

Midterm Exam

(Wednesday, Oct. 20, 2010, 90 minutes, 100 marks, single sheet reference, brief answers)

Name: KEY NetID: KEY Score: KEY

1. [28] Data preprocessing.

   (a) [5] It is not straightforward to visualize k-dimensional data for $k > 3$. Name 5 visualization techniques that can visualize 6-dimensional data effectively.

   [ANSWER:

   Most of the visualization methods, such as stick figure, Chernoff face, dimension stacking, parallel coordinates, multi-dimensional scatter plot, etc.

   ]

   (b) [6] For each of the following similarity measures, give one good application example.

   i. Cosine measure
   [ANSWER: measuring text similarity.]

   ii. Jaccard coefficient
   [ANSWER: computing similarity of a set of medical tests.]

   iii. Minkowski distance for $k = 1$
   [ANSWER: computing Manhattan (city-block) distance]

   (c) [9] Distinguishing the following concepts or measures

   i. Pearson correlation coefficient vs. covariance
   [ANSWER: Pearson correlation coefficient (X, Y) = covariance (X, Y) / $\sigma_x \cdot \sigma_Y$]

   ii. Principal component analysis vs. feature selection
   [ANSWER: Note: Answer could vary as long as has some idea similar to this: PCA is feature transformation. It projects from D dimension to m dimension, where $m < D$ while preserving the maximal variance of samples in lower dimensional space. Feature selection is to select the best subset of features, and it does not involve feature combination.]

   iii. Fourier transform vs. wavelet transform
   [ANSWER: Note: Answer could vary as long as has some idea similar to this: FT: transforms from time to frequency space, and the first several terms preserves most of the energy. WT: splits data into mean (smooth) and differences in different scales, preserves shapes.]

   (d) [8] For the following group of data

   $$500, 300, 100, -100$$

i. Calculate its mean and variance.
[ANSWER: mean: 200, variance: 50,000]

ii. Normalize the above group of data by min-max normalization with min = -1 and max = 1; and
[ANSWER: 1, 1/3, -1/3, -1]

iii. In z-score normalization, what should the value of 400 be transformed to?
[ANSWER: $\frac{400-200}{\sqrt{50{,}000}} = .89$]

2. [25] Data Warehousing and OLAP for Data Mining

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

$(a_1, a_2, a_3, \ldots, a_{10}), (b_1, b_2, b_3, \ldots, b_{10}),$

where $a_i = b_i$ if i is an odd number; otherwise $a_i \neq b_i$.

i. How many nonempty aggregate (*i.e.*, non-base) cells are there in this data cube?
[ANSWER: $2 \times 2^{10} - 2^5 - 2$]

ii. How many nonempty, *closed* aggregate cells are there in this data cube?
[ANSWER: 3: $(a_1, a_2, a_3, \ldots, a_{10}) : 1, (b_1, b_2, b_3, \ldots, b_{10}) : 1, (a_1, *, a_3, *, \ldots, a_9, *) : 2.$]

iii. If we set minimum support = 2, how many nonempty aggregate cells are there in the corresponding iceberg cube?
[ANSWER: $(a_1, *, a_3, *, \ldots, a_9, *) : 2$, and its further generalizations, so in total 2^5.]

(b) [10] Suppose a market shopping data warehouse consists of four dimensions: *customer, date, product*, and *store*, and two measures: *count*, and *avg_sales*, where *avg_sales* stores the real sales in dollar at the lowest level but the corresponding average sales at other levels.

i. [5] Draw a **star schema** diagram (sketch it, do not have to mark every possible level, and make your implicit assumptions on the levels of a dimension when you draw it).
[ANSWER: Assume most can draw this well.]

ii. [5] If one also wants to compute standard deviation as a measure, what other intermediate measures need to be introduced to make the computation efficient?
[ANSWER: Since $\sigma^2 = 1/n\Sigma_i^n(x_i^2) - (\Sigma_i^n(x_i)/n)^2$, we need to introduce sum of square, sum (and count) in order to compute standard deviation efficiently.]

(c) [5] Bitmap index is often used for accessing a materialized data cube. If a cuboid has 6 dimensions, each has 10 distinct values, and it has in total 3000 cells. How many bit vectors should this cuboid have? How long each bit vector should be (assuming no sophisticated compression techniques are explored)?

[ANSWER:

Each value of each dimension has its own bit vector. So we will need 6*10 = 60 bit vectors.

Each bit vector should have 3000 bits long (for 3000 cells).

]

3. [20] Data cube implementation

   (a) [5] Suppose *incremental update of a data cube* means that new data can be incrementally inserted into the base cuboid without recomputing the whole cube from scratch. Can you do this for an *iceberg cube*? If you can, state how; but if you cannot, state why not.
   [ANSWER: No. Because iceberg cube drops the count of the cells if it is below min_sup, it cannot get the correct count for cells in an iceberg cube upon incremental update.]

   (b) [5] Explain why the data cube could be quite sparse, and how one should implement such a sparse cube if one adopts an array-cube implementation.
   [ANSWER: Sparse since in most cases, the possible dimensional combination is huge but real data will not show up in most possible space, e.g., one cannot take all the courses in every semester, or one cannot buy most of the possible Walmart merchandizes in every transaction.
   Sparse array compression: Use chuck to partition the data, and use (chunk_id, offset) to store only those cells contain (nonempty) values.]

   (c) [5] 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 *iceberg cube* of low dimensionality (e.g., less than 6 dimensions),
      [ANSWER: Best: star-cubing or BUC
      Worst: multiway-array]

      (b) performing OLAP operations in a high-dimensional database (e.g., over 50 dimensions).
      [ANSWER: Best: Shell-fragment
      Worst: Any of the others since they cannot do it.]

   (d) [5] Most data cubes support OLAP operations on the whole population of data, but people may also like to support OLAP operations for sampling data. What are the major challenges to support OLAP on sampling data? Outline a method that may support such operations effectively.
   [ANSWER: Major challenges: Some drilling down cells may contain few or no data.
   Method: Intra or inter-cuboid expansion to combine with other cells whose attributes has low correlations with the dimensions interested.]

4. [24] Frequent pattern and association mining.

   (a) [8] A database has 5 transactions. Let $min\_sup = 0.6$ and $min\_conf = 0.8$.

customer	date	items_bought
100	10/15	{I, P, A, D, B, C}
200	10/15	{D, A, E, F}
300	10/16	{C, D, B, E }
400	10/18	{B, A, C, K, D}
500	10/19	{A, G, T, C}

i. List the frequent k-itemset for the largest k, and
[ANSWER: $k = 3, BCD : 3.$]

ii. all the strong association rules (with support and confidence) for the following shape of rules:
$\forall x \in transaction,\ buys(x, item_1) \wedge buys(x, item_2) \Rightarrow buys(x, item_3). \quad [s, c]$
[ANSWER:

$buys(x, B) \wedge buys(x, C) \Rightarrow buys(x, D). \quad [.6, 100\%]$
$buys(x, B) \wedge buys(x, B) \Rightarrow buys(x, C). \quad [.6, 100\%]$
$buys(x, C) \wedge buys(x, D) \Rightarrow buys(x, B). \quad [.6, 100\%]$

]

(b) [10] A research publication database like DBLP contains millions of papers authored by a set of researchers.

(i) [5] Using this dataset, explain why *null-invariance* property is important in the study which authors are "correlated".

[ANSWER: Since most authors are not coauthoring papers, null value is very high. Null-invariance is critical otherwise the "correlation value could be greatly influenced by null values.]

(ii) [5] To mine potential advisors and advisees relationships, what measures would you like to use? Explain your answer.

[ANSWER: Kulcizakii value and imbalance factor (or something like that).]

(c) [6] Apriori is an efficient method for mining frequent patterns. Briefly describe how to extend this method to mine frequent-itemset **incrementally**, i.e., incorporate newly added transactions without redoing the mining from scratch.

[ANSWER: Consider $DB + \delta DB$ and assume we know frequent itemsets (FP) in DB already. Mine FP_δ in δDB, taken union of FPs and scan DB for those frequent only in δDB and scan δDB for those only in DB and merge their counts.]

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.