

Name: _____ , _____
(Family name) (Given name)

Student ID: _____

THE UNIVERSITY OF NEW SOUTH WALES
Final Exam

COMP9318
Data Warehousing and Data Mining

SESSION 1, 2008

-
- Time allowed: **10 minutes** reading time + **3 hours**
 - Total number of questions: **7 + 1**
 - Total number of marks: **100 + 5 (Bonus)**
 - Only UNSW exam calculators are allowed in this exam.
 - Answer **all** questions.
 - You can answer the questions in any order.
 - Start each question on a **new page**.
 - Answers must be written in ink.
 - Answer these questions in the script book provided.
 - Do **not** write your answer in this exam paper.
 - Start each questions on a new page.
 - If you use more than one script book, fill in your details on the front of *each* book.
 - You may **not** take this question paper out of the exam.
-

SECTION A: Potpourri

Question 1

(20 marks)

Briefly answer the following questions in your script book:

- (a) List at least three differences between OLAP and OLTP.
- (b) List at least two algorithms we have discussed in the course that follow the divide-and-conquer paradigm.
- (c) What is the confidence for the rule $\emptyset \rightarrow A$? (\emptyset stands for the empty set)
- (d) What is the confidence for the rule $A \rightarrow \emptyset$? (\emptyset stands for the empty set)
- (e) What are the main differences between clustering and classification?
- (f) Give an example of a distance function that satisfies the triangle inequality.

SECTION B: Data Warehousing

Question 2

(10 marks)

Consider the following base cuboid *Sales* with *four* tuples and the aggregate function SUM:

<i>Location</i>	<i>Time</i>	<i>Item</i>	<i>Quantity</i>
Sydney	2005	PS2	1400
Sydney	2006	PS2	1500
Sydney	2006	Wii	500
Melbourne	2005	XBox 360	1700

Location, *Time*, and *Item* are dimensions and *Quantity* is the measure. Suppose the system has built-in support for the value **ALL**.

- (a) How many tuples are there in the complete data cube of *Sales*?
- (b) Write down an equivalent SQL statement that computes the same result (i.e., the cube). You can *only* use standard SQL constructs, i.e., no **CUBE BY** clause.
- (c) Consider the following *ice-berg cube* query:

```
SELECT Location, Time, Item, SUM(Quantity)
FROM Sales
CUBE BY Location, Time, Item
HAVING COUNT(*) > 1
```

Draw the result of the query in a tabular form.

- (d) Assume that we adopt a MOLAP architecture to store the full data cube of *R*, with the following mapping functions:

$$f_{Location}(x) = \begin{cases} 1 & \text{if } x = \text{'Sydney'}, \\ 2 & \text{if } x = \text{'Melbourne'}, \\ 0 & \text{if } x = \mathbf{ALL}. \end{cases}$$

$$f_{Time}(x) = \begin{cases} 1 & \text{if } x = 2005, \\ 2 & \text{if } x = 2006, \\ 0 & \text{if } x = \mathbf{ALL}. \end{cases}$$

$$f_{Item}(x) = \begin{cases} 1 & \text{if } x = \text{'PS2'}, \\ 2 & \text{if } x = \text{'XBox 360'}, \\ 3 & \text{if } x = \text{'Wii'}, \\ 0 & \text{if } x = \mathbf{ALL}. \end{cases}$$

Draw the MOLAP cube (i.e., sparse multi-dimensional array) in a tabular form of $(ArrayIndex, Value)$. You also need to write down the function you chose to map a multi-dimensional point to a one-dimensional point.

SECTION C: Data Cleaning

Question 3

(15 marks)

Consider running the prefix-based SSJoin algorithm with an overlap similarity threshold of $t = 3$ on the following dataset.

<i>objectID</i>	<i>elements</i>
1	$\{a, a, d, c, b\}$
2	$\{a, a, a, c, b\}$
3	$\{b, c, d\}$
4	$\{c, c, c\}$

- (a) What is the **overlap** similarity between the first and the second object?
- (b) Write down the prefixes of the objects.
- (c) Show the steps of performing the prefix-based SSJoin on the above dataset. The global ordering based on the document frequencies (DF) must be used.
- (d) If the similarity threshold t is relative, i.e., two objects, x and y , will be written to the output only if their overlap is no less than $\max(t \cdot |x|, t \cdot |y|)$ (where $|x|$ is the size of the multiset x), it is possible to compute the SSJoin results by extracting a prefix of length $|x| - \lceil t \cdot |x| \rceil + 1$ for every object x . Prove that this algorithm is correct (i.e., it won't miss any result).

SECTION D: Association Rule Mining

Question 4

(15 marks)

Consider the market basket transactions shown in the table below.

Transaction ID	Items
101	$\{M, B, D\}$
102	$\{B, Y, M\}$
103	$\{M, D, C\}$
104	$\{B, Y, C\}$
105	$\{B, Y, D, C\}$

- (a) What is the maximum number of size-3 frequent itemsets that can be derived from this data set (assuming $minsup > 0$)
- (b) Show the steps of running the FP-growth algorithm on the above dataset with the minimum support threshold of 50%. Ties should be broken according to the alphabetic order, i.e., if X and Y have the same support, X is deemed as less frequent as Y .
- (c) Suppose we know the price of each item (see the table below) and we want to mine frequent itemsets whose total price is no larger than 5 ($minsup = 50\%$). Show the steps of running the Apriori algorithm with such constraint pushed inside.

Item	Price
M	2
B	3
D	1
Y	3
C	2

Question 5

(10 marks)

Let $\text{conf}(\text{rule})$ be the confidence of an association rule rule .

Prove that

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

AB ABC

$$\text{conf}(U \rightarrow V) \leq \text{conf}(U \cup X \rightarrow V - X) \quad , \text{ where } X \subset V \text{ and } X \not\subseteq U$$

$$\text{sup}(U \cup X) \geq \text{sup}(U)$$

$$\text{sup}(V) \geq \text{sup}(V - X)$$

$$\text{conf}(U \rightarrow V) = \text{sup}(U \rightarrow V) / \text{sup}(U)$$

$$\text{conf}(U \cup X \rightarrow V - X) = \text{sup}(U \cup X) / \text{sup}(V - X)$$

SECTION E: Classification and Prediction

Question 6

(15 marks)

Consider the following training dataset.

<i>id</i>	<i>a</i> ₁	<i>a</i> ₂	<i>a</i> ₃	<i>class</i>
1	T	T	1.0	Y
2	T	T	6.0	Y
3	T	F	5.0	N
4	F	F	4.0	Y
5	F	T	7.0	N
6	F	T	3.0	N
7	F	F	8.0	N
8	T	F	7.0	Y
9	F	T	5.0	N

- (a) Assume a_3 values have to be discretised to a'_3 as follows:

a_3	a'_3
$0.0 \leq a_3 < 3.0$	L
$3.0 \leq a_3 < 6.0$	M
$6.0 \leq a_3 < 9.0$	H

Show the dataset after applying the above transformation. For the rest of the questions, we will use the transformed dataset (i.e., consisting of attributes a_1 , a_2 and a_3).

- (b) Show the decision tree obtained by the ID3 decision tree induction algorithm.
- (c) Build a Naive Bayes classifier for the training dataset and use it to classify a new tuple (10, T, F, M).

SECTION F: Cluster Analysis

Question 7

(15 marks)

Use the **similarity** matrix in the table below to perform hierarchical clustering using several different algorithms.

Show your final results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

	p_1	p_2	p_3	p_4	p_5
p_1	1.00	0.10	0.41	0.55	0.35
p_2	0.10	1.00	0.64	0.47	0.98
p_3	0.41	0.64	1.00	0.44	0.85
p_4	0.55	0.47	0.44	1.00	0.76
p_5	0.35	0.98	0.85	0.76	1.00

- (a) Show the steps and final result of running the **single-link hierarchical clustering** algorithm.
- (b) Show the steps and final result of running the **complete-link hierarchical clustering** algorithm.

single-link 选取最近的作为距离
complete-link 选取最远的作为距离

SECTION G: Bonus

Question 8

(5 marks)

Summarise several commonly used techniques that can make data mining algorithms scale to large datasets (i.e., datasets that cannot be accommodated in the main memory). You need to *briefly* describe how each of the techniques you listed works.

END OF EXAM PAPER