香 港 中 文 大 學
The Chinese University of Hong Kong

Course Examinations 2014

Course Code & Title : PhD Qualifying Exam – Information Systems

Time allowed : 2 hours 0 minutes

Student I.D. No. : ..... Seat No. : .....

Question 1 [15 marks] Database – Indexing

Consider the following relation:

STUDENT (sid, sname, did, gpa). Each tuple describes a student with *sid* being the student's id, *sname* his name, *did* his department id, and *gpa* his GPA. The table has a candidate key sid.

The relation STUDENT has 30,000 tuples and 30 tuples of STUDENT fit into one page. The four attributes in the relation have the same byte size. Assume that the gpa of students are uniformly distributed in the range from 1.0 to 4.0. $30000 \times \frac{0.3}{3.1} = 3000$

To process the query: *Display the names of students whose gpa is in the range from 3.0 to 3.3*, we can consider two indexing approaches:

- Approach 1: Build a clustered index on gpa;
- Approach 2: Build a multi-attribute index on ⟨gpa, sname⟩ and do index-only scan.

For each indexing approach, a B-tree with height 3 is built. Assume all index pages are stored in disk. Calculate the number of disk page access for the above two approaches. Which approach is better?

$b = \frac{30000}{30} = 1000$

Approach 2:

Question 2 [20 marks] Database – Query Processing

Consider the following relations:

STUDENT (sid, sname, did, gpa). Each tuple describes a student with *sid* being the student's id, *sname* his name, *did* his department id, and *gpa* his GPA. The table has a candidate key sid.

DEPT (did, dname). Each tuple describes a department with *did* being the department's id, and *dname* its name. The table has a candidate key did.

SQL: select sname, dname

from STUDENT, DEPT

where STUDENT.did=DEPT.did

The relation STUDENT has 30,000 tuples and 30 tuples of STUDENT fit into one block. The relation DEPT has 1,000 tuples and 5 tuples of DEPT fit into one block. There is no index on both relations.

(a) Assume that we use block nested-loop join to perform the above SQL query using STUDENT as the outer relation. Estimate the cost in terms of block accesses of the join under the following three memory buffer settings:

(1) The memory buffer has no restriction in size. **[5 marks]**

$b_r = \frac{30000}{30} = 1000 \quad b_s = \frac{1000}{5} = 200 \quad \text{cost} : b_r + b_s = 1200$

(2) The memory buffer has 3 blocks. **[5 marks]**

$\text{cost} = \lceil \frac{b_r}{M-2} \rceil \times b_s + b_r = 201000$

(3) The memory buffer has 22 blocks. **[5 marks]**

$\text{cost} = \lceil \frac{b_r}{M-2} \rceil \times b_s + b_r = \lceil \frac{1000}{20} \rceil * 200 + 1000$

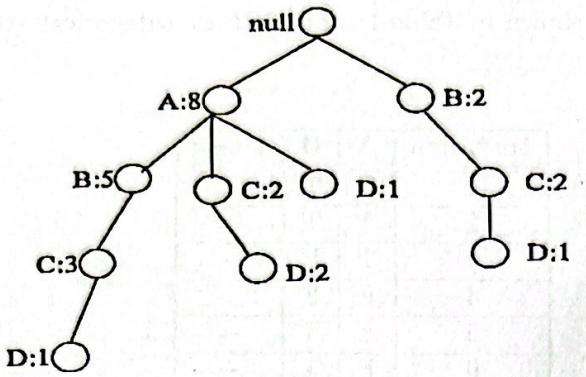
$= 11000$

(b) If an index is available on the join attribute *did*, what is the main benefit that we can obtain in processing the SQL query? **[5 marks]**

Less disk page access cost is needed.

Question 3 [15 marks] Data Mining – Association Rule Mining

A 8
B 7
C 7
D 5



A, B, C, D 1
A, C, D 2
A, D 1
B, C, D 1
A, B, C 2
B, C 1
A, B 2

Figure 1: An FP-Tree for Question 3

(a) A database with 10 transactions has its FP-tree shown in Figure 1. Suppose the minimum support $min\_sup = 4$. List all frequent itemsets and their support counts. [9 marks]

{A}: 8 {C}: 7 {A, B}: 5 {A, D} 4 {C, D}: 4

{B}: 7 {D}: 5 {A, C}: 5 {B, C}: 5

(b) For an association rule $A \rightarrow B$, compute its support, confidence and lift measure. [6 marks]

$$\text{support}(A \rightarrow B) = \frac{5}{10} = 0.5$$

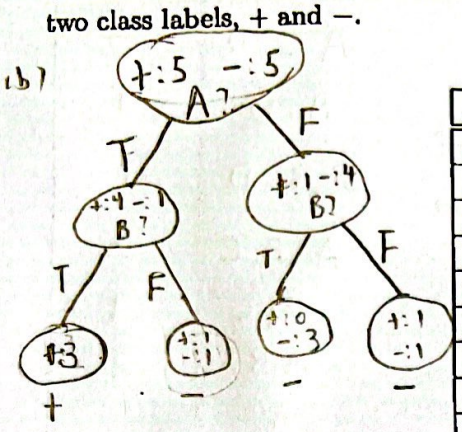
$$\text{confidence}(A \rightarrow B) = \frac{\text{support}(A\,B)}{\text{support}(A)} = \frac{0.5}{0.8} = 0.625$$

$$\text{lift}(A \rightarrow B) = \frac{\text{confidence}(A \rightarrow B)}{\text{support}(B)} = \frac{0.625}{0.7} = \frac{25}{28}$$

Question 4 [20 marks] Data Mining – Decision Tree Induction

Consider the training dataset shown in Table 1. A and B are categorical attributes. There are two class labels, + and −.

(b)



Instance	A	B	Class
1	T	T	+
2	T	T	+
3	T	F	−
4	F	F	+
5	F	T	−
6	F	T	−
7	F	F	−
8	T	F	+
9	F	T	−
10	T	T	+

Table 1: A Data Set for Question 4

(a) Calculate the gain in the Gini index when splitting on A and B, respectively. Show your calculation details. Which attribute would be chosen as the first splitting attribute? [8 marks]

$Gini(D) = 1-(\frac{1}{2})^2-(\frac{1}{2})^2 = \frac{1}{2}$

A	+	−
T	4	1
F	1	4

B	+	−
T	3	3
F	2	2

$Gini(A) = \frac{1}{2}\times(1-(\frac{1}{5})^2-(\frac{4}{5})^2)+\frac{1}{2}\times(1-(\frac{1}{5})^2-(\frac{4}{5})^2) = \frac{8}{25}$

$Gini(B) = \frac{3}{5}\times(1-(\frac{1}{2})^2-(\frac{1}{2})^2)+\frac{2}{5}(1-(\frac{1}{2})^2-(\frac{1}{2})^2) = \frac{1}{2}$

(b) Draw a two-level decision tree using both attributes for splitting. Mark the class label in each leaf node. In case of a tie on the positive and negative instances in a leaf node, mark the node as −. [6 marks]

(c) Show the confusion matrix based on the induced decision tree, and calculate the accuracy, precision, recall, and F_1-measure. (Note that precision, recall and F_1-measure are defined with respect to the + class.) [6 marks]

Confusion Matrix

		Predicted class +	Predicted class −
Actual class	+	3	2
	−	0	5

Accuracy: $\frac{3+5}{10} = 0.8$

Precision: $\frac{3}{3+0} = 1$

Recall: $\frac{3}{3+2} = 0.6$

$F = \frac{2\times 1\times 0.6}{1+0.6} = 0.75$

Question 5 [15 marks] Operating Systems – Deadlock

Answer the following questions using Banker's algorithm.

Consider the following snapshot of a system:

	Allocation					Max Request				
	A	B	C	D	E	A	B	C	D	E
P0	0	0	1	2	3	0	0	3	2	3
P1	2	0	0	0	1	2	7	5	0	2
P2	0	0	3	4	2	6	6	5	6	4
P3	2	3	5	4	0	4	3	5	6	3
P4	0	3	3	2	0	0	6	5	2	0
P5	3	2	4	0	0	4	2	5	2	6

Table 2: Resource Allocation

(a) Need

	A	B	C	D	E
P0	0	0	2	0	0
P1	0	7	5	0	1
P2	6	6	2	2	2
P3	2	0	0	2	3
P4	0	3	2	0	0
P5	1	0	1	2	6

(a) What is the content of the matrix *Need* denoting the number of additional resources needed by each process? **[5 marks]**

(b) Assume the available resources are <A:2, B:1, C:2, D:0, E:0>. Is the system in a safe state? If so, list ALL the possible safe sequences. **[5 marks]**

① $P_0 \rightarrow V = \langle 2\ 1\ 3\ 2\ 3 \rangle$

② $P_3 \rightarrow V = \langle 4\ 4\ 8\ 6\ 3 \rangle$

③ $P_4 \rightarrow V = \langle 4\ 7\ 11\ 8\ 3 \rangle$

④ $P_1 \rightarrow V = \langle 6\ 7\ 11\ 8\ 4 \rangle$

⑤ $P_2 \rightarrow V = \langle 12\ 13\ 13\ 10\ 6 \rangle \rightarrow P_5$

$\therefore P_0 \rightarrow P_3 \rightarrow P_4 \rightarrow P_1 \rightarrow P_2 \rightarrow P_5$

(c) In the current snapshot of the system, can request <A:0, B:1, C:2, D:0, E:0> by P2 be granted immediately? Why? **[5 marks]**

$\langle 2\ 1\ 2\ 0\ 0 \rangle \Rightarrow \langle 0\ 1\ 2\ 0\ 0 \rangle = \langle 2, 0, 0, 0, 0 \rangle$

Not safe.

P_2 cannot be granted.

Question 6 [15 marks] Operating Systems – Memory Management

(a) Consider a paging scheme, with the page table stored in the main memory. If each memory reference takes 250 nanoseconds, what is the effective memory-access time (average time to access any memory location)? **[4 marks]**

250 + 250 = 500 ns.

(b) Suppose a translation look-aside buffer (TLB) is added to store part of the page table in the cache. The hit ratio of the TLB is 90%, and the access time of the TLB is 20 nanoseconds. What is the effective memory-access time? **[4 marks]**

(250+20) × 90% + (500+20) × 10% = 295 ns

(c) Suppose a computer has an 8-bit address space, i.e., each logical address is 8-bits long. Each page has size 32 Bytes. How many entries does the page table contain? **[4 marks]**

$\log \frac{2^8}{32} = 3$ bit ∴ $2^3 = 8$ page number contain

(d) What is the benefit and drawback of hierarchical page tables compared to a flat page table? **[3 marks]**

- Avoid page table using too much memory.
- Lead to additional access cost.

-End-