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

Course Examinations 2012

Course Code & Title : PhD Qualifying Exam – Information Systems

Time allowed : ..... **hours** ..... **minutes**

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

IMPORTANT: Students are required to solve **ALL** the questions if their primary area is Information Systems; or solve any **FOUR** questions if their secondary area is Information Systems.

Question 1 [10 marks] Database – SQL

Consider the following relational database related to a university:

STUDENT (sid, sname)

PROFESSOR (pid, pname, dept)

CLASS (cid, cname)

OFFERING (offid, *cid*, semester, year, *pid*)

ENROLLED (*offid, sid*, grade)

In the above schemas, primary keys are underlined and foreign keys in each relational schema are in *italics*. All foreign keys are NOT NULL. The meaning of the attributes is self-explanatory. We assume all domains are strings (characters). Express the following queries in SQL expressions.

(a) Find the distinct names of all students taught by Michael in the Spring semester of 2012. **[5 marks]**

(b) Find the names of all classes that have at least two professors in the Spring semester of 2012 (note: we only need the names of these classes and not the number of offerings). **[5 marks]**

Question 2 [20 marks] Database – Query Plan

Consider a database consisting of the following three relation schemas:

CUSTOMERS (CID, Cname, job, age, street, city): 10,000 tuples

BRANCHES (BID, Bname, city): 2,000 tuples

ACCOUNTS (*CID, BID*, balance): 12,000 tuples

The meaning of the attributes in the above schemas is self-explanatory. For example, Bname is the name of the branch. Keys are underlined and foreign keys are in *italics*.

Consider the following SQL query:

select Cname, Bname, balance

from CUSTOMERS, BRANCHES, ACCOUNTS

where CUSTOMERS.cid = ACCOUNTS.cid

and BRANCHES.bid = ACCOUNTS.bid

and age > 55

and CUSTOMERS.city = "BOSTON"

and BRANCHES.city = "BOSTON"

and balance > 100,000

Assume that 30% of customers are older than 55, 20% of customers live in BOSTON, 10% of branches are located in BOSTON and 30% of balance are greater than 100,000.

Using the above information for optimizing a query, draw the most efficient query tree for the SQL query. You should clearly indicate all the selection and join operations on nodes in the tree. Justify your answer by analyzing the size of intermediate results.

Question 3 [20 marks] Data Mining – Clustering

The following lists the hierarchical clustering algorithm.

Hierarchical Clustering Algorithm

Input: n data points
Output: A hierarchical clustering

1. Compute the distance matrix containing pairwise distance between all data points
2. **repeat**
3. Merge the closest two clusters
4. Update the distance matrix to reflect the distance between the new cluster and the original clusters
5. **until** Only one cluster remains.

(a) Explain why the time complexity of hierarchical clustering is $O(n^3)$. **[10 marks]**

(b) Improve the algorithm to reduce the time complexity to $O(n^2 \log n)$. **[10 marks]**

Question 4 [20 marks] Data Mining – Location-based Social Networks

Location-based social networks (LBSNs), such as *Foursquare*, *Jiepang*, etc., have been increasingly popular recently. Users are sharing their locations or geo-tagged information with friends through check-ins. Table 1 shows three users and their check-in sequences. A check-in record contains check-in time, latitude, longitude, location name and check-in category. Assume you have collected a large set of historical user check-in activities as shown in Table 1. In the historical dataset, the number of distinct locations can be very large, but the number of check-in categories is usually small, e.g., no more than 10.

Given a new user t with a sequence of his check-in activities within a day:

User t, 5/15/2011, ⟨12:30, 31.88°, −47.12°, *Pizza Hut*, Food⟩,

⟨14:30, 31.28°, −47.42°, *IMAX*, Entertainment⟩,

design a method which can accurately predict t's next check-in location and category, based on what you have learnt from the historical data. Describe which model and which information in the historical data you will use, and how.

Table 1: Example of User Check-in Sequences

User	**Date**	**Check-in Records**
John	1/13/2010	⟨08:17, 41.89°, −87.65°, *Starbucks*, Food⟩, ⟨09:30, 41.88°, −87.63°, *City Hall*, Community⟩, ⟨12:35, 41.88°, −87.62°, *Subway*, Food⟩, ⟨17:22, 41.99°, −87.73°, *Macy's*, Shopping⟩, ⟨19:45, 41.99°, −87.72°, *Cinema*, Entertainment⟩
Andy	1/27/2010	⟨11:37, 37.42°, −122.17°, *McDonald's*, Food⟩, ⟨18:01, 37.53°, −122.07°, *Park*, Outdoors⟩, ⟨19:30, 37.54°, −122.06°, *Italian*, Food⟩
Eva	2/21/2010	⟨20:37, 40.77°, −73.96°, *Ocean Bay*, Food⟩, ⟨22:39, 40.87°, −73.92°, *Hi-Fi*, Night life⟩, ⟨23:51, 40.86°, −73.90°, *Live City*, Night life⟩

Question 5 [15 marks] Operating Systems – Storage

(a) Suppose a disk has the following properties:

Average seek time: 10ms

Time for one rotation: 11ms (5600RPM)

Transfer rate: 4 Mbytes/s

Sector Size: 2 Kbyte

What is the expected total access time to read 1000 consecutive sectors (within the same track) from a random place on the disk? **[5 marks]**

(b) Suppose a disk drive has 100 cylinders numbered 0 to 99. The disk is currently serving a request at the cylinder 30 and the previous request was at the cylinder 27. The queue of the pending request (in FIFO order) are:

15, 38, 8, 52, 96, 70, 24, 44

(i) Please specify the sequences of disk head movement and calculate the total number of movements (in term of number of cylinders) in (1) FCFS, (2) SSTF, (3) C-SCAN and (4) C-LOOK. **[8 marks]**

(ii) In a system that places a heavy load on the disk, which of the following two schemes is better, SSTF or C-SCAN? Briefly justify your answer. **[2 marks]**

Question 6 [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**	**A**	**B**	**C**	**D**
P0	0	0	3	2	0	0	4	4
P1	1	0	0	0	2	6	8	0
P2	1	3	5	4	3	6	10	10
P3	0	0	3	2	0	6	8	4
P4	0	0	1	4	0	6	6	10

Table 2: Resource Allocation

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

(b) Assume the available resources are <A:1, B:6, C:2, D:2>. Is the system in a safe state? If so, list ALL the possible safe sequences. **[10 marks]**

-End-

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

Course Examinations 2013

Course Code & Title : PhD Qualifying Exam – Information Systems

Time allowed : ..... hours ..... minutes

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

IMPORTANT: Students are required to solve **ALL** the questions if their primary area is Information Systems; or solve any **FOUR** questions if their secondary area is Information Systems.

Question 1 [20 marks] Database – Relational Database

Consider the following relations about NBA:

PLAYER (pid, pname, nation). Each tuple describes a player with *pid* being the player's id, *pname* his name, and *nation* his nationality. The table has a candidate key pid.

TEAM (tid, tname). Each tuple describes a team with *tid* being the team's id, and *tname* its name. The table has a candidate key tid.

REGISTER (pid, tid, salary, year). Each tuple records the fact that a certain player (indicated by *pid*) played for a certain team (indicated by *tid*) in a specific *year* with an annual income given in *salary*. The table has a candidate key (pid, year). Note that a player may belong to various teams in different years.

(a) Write relational algebra queries for the following tasks. **[10 marks]**

(1) Find the names of the teams and the corresponding year that "Michael Jordan" ever played for.

(2) Find the names of all players of the team "Heat" in 2013.

(b) Write SQL statements for the following tasks. **[10 marks]**

(1) For each player, display his pid, and the first and last years in which he played.

(2) Find the pids of all players that played from 1996 through 2005 (i.e., such a player played in 1996, 1997, ..., and 2005).

Question 2 [20 marks] Database – Hash Join

Consider the following two relations R and S. The number of records in R and S is also given.

$R(A, B, C, D)$: 20,000 records

$S(A, E)$: 36,000 records

Assume the size of one memory page is 4K bytes (for simplicity, assume 4K bytes=4,000 bytes) and the memory buffer has 30 pages. Assume the size of every attribute value is 10 bytes.

Consider the following SQL query:

select B, E

from R, S

where $R.A = S.A$

(a) What are the sizes of R and S in terms of the number of pages? **[5 marks]**

(b) Let h be a hash function that assigns a given record into one of ten partitions (or ten buckets). Assume using hash join to process the above query as follows:

Step 1: Partition R using h on $R.A$ and write the partitions to the disk.

Step 2: Partition S using h on $S.A$ and write the partitions to the disk.

Sept 3: For each partition of R, we transfer it from disk to memory buffer and match it against the corresponding partition of S (page by page). The matching can be done by an in-memory hash built on $R.A$. Assume the ten partitions of R have an equal number of pages, and the ten partitions of S have an equal number of pages.

Estimate the I/O (page read/write) cost of each step and then compute the total I/O cost of hash join. **[10 marks]**

(c) One further optimization of Part (b) is to keep the first partition of R in the memory buffer in Step 1 and then use the rest of memory to generate only the other partitions in Step 2. The records that belong to the first partition of S are matched directly with all records of the first partition of R on-the-fly. After matching the first partitions of R and S, the comparison in other partitions (the second, third and so on) follows Step 3.

Estimate the total I/O cost of hash join using this optimization. **[5 marks]**

Question 3 [20 marks] Data Mining – Association Rule Mining

(a) Table 1 shows a transaction database. Given minimum support=2, list all frequent itemsets. Among them, which are closed itemsets and which are maximal itemsets? **[12 marks]**

TID	A	B	C	D	E
t1	0	1	1	1	1
t2	1	0	1	1	1
t3	1	1	0	1	1
t4	1	1	1	0	1
t5	1	1	1	1	0

Table 1: A Transaction Database

(b) Table 2 shows the contingency table of coffee and milk sales.

	coffee	$\overline{coffee}$	$\sum_{row}$
milk	1000	1500	2500
$\overline{milk}$	2000	500	2500
$\sum_{col}$	3000	2000	5000

Table 2: Coffee and Milk Sales

(1) Given an association rule "coffee→milk", compute its support and confidence. **[3 marks]**

(2) Given another association rule "$\overline{coffee}$ →milk", compute its support and confidence. **[3 marks]**

(3) Are the sales of coffee and milk positively correlated or negatively correlated? Show your calculation. **[2 marks]**

Question 4 [10 marks] Data Mining – Link Prediction

The *link prediction* problem studies, given a snapshot of a social network at time t, how to accurately predict the edges that will be added to the network during the interval from time t to a given future time t'. An effective measure for link prediction is the number of common friends of two users. If two users have many common friends, it is very likely that they will become friends too.

Consider a new type of social networks, called *location-based social networks* (LBSNs), such as *Foursquare*. In LBSNs, people make check-ins to share their locations and activities. They can also write reviews and upload photos or videos.

Compared with the traditional link based method for link prediction, in LBSNs, how can you exploit the above rich features for link prediction more accurately? Describe your proposed method with technical details. For example, simply using the number of common locations visited by two users may not work well, as not all places have the same importance to foster new social ties among individuals who visit them, e.g., train stations, supermarkets which are visited by many people, but may not foster new friendships.

Question 5 [15 marks] Operating Systems – File System

Consider a magnetic disk system which rotates at 3000 rpm. Its average seek time is 15 milliseconds and the transfer rate is 2MByte/s. The sector size is 1KByte. (1K is defined as 2^{10} and 1M is defined as 2^{20}.)

(a) What is the average rotation delay? **[3 marks]**

(b) How long does it take to transfer one sector (ignore seek and rotation times)? **[2 marks]**

(c) Consider a file that consists of 8 sectors. Please compute the total time needed for reading this file under different allocation schemes. Show your calculation details.

(1) Contiguous allocation of disk space (assume that all 8 sectors are stored on the same track); **[5 marks]**

(2) Indexed allocation of disk space (the 8 sectors are randomly placed on different tracks of the disk. Ignore the time for accessing the index block). **[5 marks]**

Question 6 [15 marks] Operating Systems – Virtual Memory

Consider the following sequence of page references in term of page numbers:

1, 2, 1, 3, 4, 4, 5, 2, 1

Suppose there are 3 frames allocated for this process. Illustrate the contents of the frames under 3 different page replacement algorithms and compute the number of page faults for each algorithm.

	1	2	1	3	4	4	5	2	1
Frame 1									
Frame 2									
Frame 3									

(a) FIFO **[5 marks]**

(b) LRU: Least recently used replacement algorithm **[5 marks]**

(c) OPT: the optimal replacement algorithm **[5 marks]**

-End-

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.

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?

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]**

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

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

(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]**

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

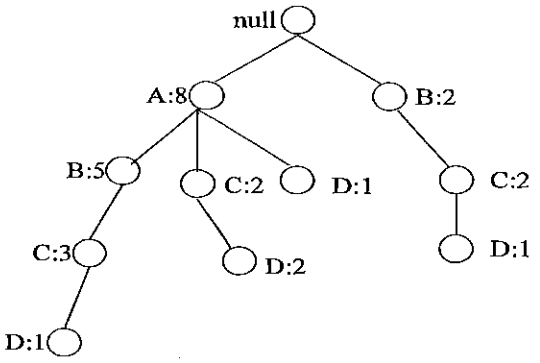


Figure 1: An FP-Tree for Question 3

(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]**

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

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 −.

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]**

(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]**

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) 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]**

(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]**

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]**

(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]**

(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]**

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

-End-

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

Course Examinations 2015

Course Code & Title : PhD Qualifying Exam – Information Systems

Time allowed : 3 hours 0 minutes

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

Question 1 [10 marks] Data Structures and Algorithms

Consider a balanced binary search tree T with n nodes carrying real values. For any node v in T, the node values in the left subtree of v are no greater than the value of v, and the node values in the right subtree of v are no less than the value of v. See Figure 1 for an example.

(a) Given a number x, write a function that finds whether x is in T. If yes, return true, otherwise, return false. What is the time complexity of your function? **[4 marks]**

(b) Given a number z, design an algorithm that finds whether there exist two values x, y in T such that $x + y = z$. For example, $z = 20$, there exist 5 and 15 in T such that $5 + 15 = 20$. What is the time complexity and space complexity of your algorithm? **[6 marks]**

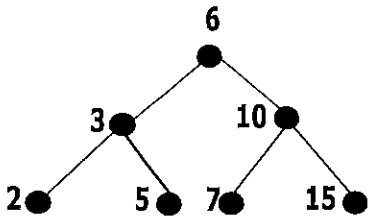


Figure 1: A balanced binary search tree T for Question 1

Question 2 [10 marks] Data Structures and Algorithms

Let $G(V, E)$ be an undirected and connected graph with n vertices and m edges. A positive weight is assigned on each edge. Given two different vertices $s \in V$ and $t \in V$, design an $O(m \cdot \log m)$ time algorithm to find a path from s to t in G, such that the maximum edge weight in the path is minimized. Explain your algorithm steps to justify the time complexity. (**Note**: If you design an algorithm that works, but is more expensive than the required time complexity, you can get partial marks.)

For example, for the graph G in Figure 2, for nodes a and g, the path (a, f, e, b, c, g) is the answer, as the maximum edge weight as 4 is minimized among all paths between a and g.

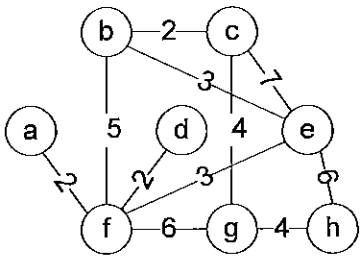


Figure 2: An undirected weighted graph G for Question 2

Question 3 [20 marks] Operating Systems

(a) Consider the following sequence of page references in terms of page numbers:

1, 2, 3, 1, 4, 4, 5, 1, 3

Suppose there are 3 frames allocated for this process. Illustrate the contents of the frames under 3 different page replacement algorithms and compute the number of page faults for each algorithm.

(1) FIFO: first in first out **[4 marks]**

(2) LRU: least recently used replacement algorithm **[4 marks]**

(3) OPT: the optimal replacement algorithm **[4 marks]**

(b) Consider a single-level paging scheme with a translation look-aside buffer (TLB) used to store part of the page table. Assume that each memory access time is 200 nanoseconds, TLB access time is 10 nanoseconds and TLB hit ratio is 0.9. What is the effective memory access time? **[4 marks]**

(c) Consider a segmentation with paging scheme, i.e., each segment is further divided into fixed-size pages (frames), can you list two main advantages of using such a scheme over a pure segmentation scheme? **[4 marks]**

Question 4 [20 marks] Database

Consider the following relations:

Book(bid, title, category): Each tuple represents a book. The attribute *category* describes the genre of the book (e.g., novel, sci-fi, science, music, ...). The underlined is the candidate key.

Student(sid, sname, dept): Each tuple represents a student. The attributes' meanings should be self-explanatory. The underlined is the candidate key.

Borrow(sid, bid, checkout-time, return-time): A tuple means that student *sid* checked out book *bid* at *checkout-time*, and returned it at *return-time*.

All attributes are strings, except checkout-time and return-time, which are integers. A smaller checkout-time represents an earlier timestamp (same for return-time).

The relation Student has 6,000 tuples and 10 tuples of Student fit into one block. The relation Book has 20,000 tuples and 20 tuples of Book fit into one block. The relation Borrow has 30,000 tuples and 30 tuples of Borrow fit on one block. There is no index on all the relations.

(a) Write SQL queries for the tasks below:

(1) Find the category with at least 2,000 books. **[3 marks]**

(2) Define the popularity of a book as the number of distinct students that have ever borrowed it. Find the titles of the books with the highest popularity (note that multiple books may have the same popularity). **[5 marks]**

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

```
SQL: select sname, bid, checkout-time
     from Student, Borrow
     where Student.sid=Borrow.sid
```

(1) The memory buffer has no restriction in size. **[3 marks]**

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

(c) Consider the following two strategies for computing the join:

Strategy 1: (Student $\bowtie$ Book) $\bowtie$ Borrow

Strategy 2: (Student $\bowtie$ Borrow) $\bowtie$ Book

Which strategy is better? Explain the reason(s) of your choice by first estimating the size (in terms of tuples) for the first join product in the bracket in each strategy. **[6 marks]**

Question 5 [20 marks] Data Mining

Table 1 shows a data set with three attributes A, B, C and two class labels + and −. Build a two-level decision tree.

A	B	C	Number of Instances	
			+ class	− class
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

Table 1: A data set for Question 5

(a) According to the classification error rate criterion, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate. **[8 marks]**

(b) Repeat step (a) for the two children of the root node to find the best split in the second level of the decision tree. For each attribute tested, show the contingency table and the gains in classification error rate. **[3 marks]**

(c) How many instances are misclassified by the resulting decision tree? **[2 marks]**

(d) If we choose attribute C as the first splitting attribute in the decision tree, for the two children of the root node, find the best split in the second level of the decision tree. For each attribute tested, show the contingency table and the gains in classification error rate. **[3 marks]**

(e) How many instances are misclassified by the resulting decision tree in step (d)? **[2 marks]**

(f) Compare the two decision trees you build to conclude about the greedy nature of the decision tree induction algorithm. **[2 marks]**

Question 6 [20 marks] Information Retrieval

(a) F measure combines precision and recall.

(1) Describe and explain some advantages of using balanced F measure rather than just taking the average of precision and recall. **[3 marks]**

(2) Mention an application where precision is more appropriate than F measure, and provide explanation. **[3 marks]**

(b) Consider a text classification problem involving two classes. Suppose we employ the standard Rocchio text classification method.

(1) Write the pseudo-codes for the training process and the testing process when cosine similarity is used. **[4 marks]**

(2) Derive the formulation for calculating the separating hyperplane. **[4 marks]**

(c) In multi-modal text classification problems, some or all classes exhibit internal multi-modal distribution. For such problems, the k-nearest-neighbor (kNN) method generally handles better than the standard Rocchio method. Use an example of a data set to illustrate this observation. **[6 marks]**

-End-

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

Course Examinations 2016

Course Code & Title : PhD Qualifying Exam – Information Systems

Time allowed : 3 hours 0 minutes

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

Question 1 [10 marks] Data Structures and Algorithms

Consider a balanced binary search tree T with n nodes carrying real values. For any node v in T, the node values in the left subtree of v are no greater than the value of v, and the node values in the right subtree of v are no less than the value of v. See Figure 1 for an example. The tree height is denoted as h. Suppose we use the following data structure to describe a tree node:

```
TreeNode {
    int value;
    TreeNode *left, *right;
};
```

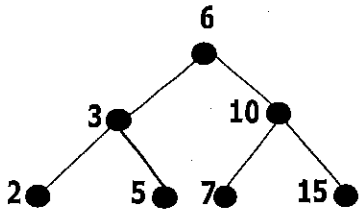


Figure 1: A balanced binary search tree T for Question 1

(a) Design an algorithm to find the k-th minimum value in T, where k is a query parameter and $1 \leq k \leq n$. For example, for $k = 3$, the answer is 5 in Figure 1. What is the time complexity of your algorithm? **[4 marks]**

(b) Show how to modify the TreeNode structure, so that you can find the k-th minimum element in $O(h)$ time. Describe the modified data structure and your algorithm. **[6 marks]**

Question 2 [10 marks] Algorithms

Given an $m \times n$ matrix filled with non-negative numbers, find a path from the top left element to the bottom right one, which minimizes the sum of all numbers along this path. For example, for the 4×4 matrix in Figure 2, the answer path is "1–2–11–5–12–4–10", and the sum is 45.

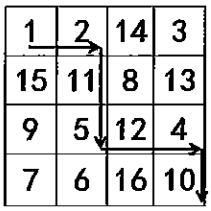


Figure 2: A 4×4 matrix

Question 3 [20 marks] Operating Systems

(a) Consider the following snapshot of a system:

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

Table 1: Resource Allocation

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

(2) Assume the available resources are <A:1, B:2, C:0, D:0, E:3>. Is the system in a safe state? If so, list ALL the possible safe sequences. **[5 marks]**

(b) Consider the following system:

(1) Byte-addressable memory

(2) 24-bit logical address

(3) Maximum segment length of 2M bytes

(4) Page size of 128 bytes

Using segmentation with one-level paging scheme, the logical address can be partitioned into three parts. List the name of each part and their size in terms of bits. Show your calculation details. **[6 marks]**

(c) According to the following page table, given the logical addresses (in the form of <page number, offset>; all numbers are decimal numbers), compute the corresponding physical addresses. Assume that one-level pure paging scheme is used, and the sizes of page and frame are both 1024 bytes.

(1) logical address <0, 857> **[2 marks]**

(2) logical address <5, 989> **[2 marks]**

Page number	Frame number
0	8
1	25
3	44
5	22
6	18
8	12

Table 2: Page Table

Question 4 [20 marks] Database

Consider the following relational database related to a university:

STUDENT (sid, sname, dept)

PROFESSOR (pid, pname, dept)

COURSE (cid, cname, credit)

TEACH (*pid, cid*, year)

ENROLL (*sid, cid*, year, grade)

In the above schemas, primary keys are underlined and foreign keys in each relational schema are in *italics*. All foreign keys are NOT NULL. The meaning of the attributes is self-explanatory.

(a) Give relational algebra queries for the tasks below:

(1) Find the names of all the courses that have not been taken by any student; **[3 marks]**

(2) Find the names of all students that have taken all the courses taught by the professor with pid = "p123"; **[3 marks]**

(b) Write SQL queries for the tasks below:

(1) Find the names of all the courses that have been taken by students from more than one department; **[3 marks]**

(2) Find the name of the professor that teaches a course with the largest enrollment number; if there is more than one professor with the same largest enrollment number, report all of them; **[3 marks]**

(c) Consider the following SQL query:

select sname, cname, grade
from STUDENT, COURSE, ENROLL
where STUDENT.sid = ENROLL.sid
and COURSE.cid = ENROLL.cid
and dept = "SEEM"
and credit = 2
and grade = "A"

Assume that there are 10,000 students, and 20% of them are from SEEM department; there are 200 courses, and 50% of them have a credit of 2; and there are 100,000 tuples in ENROLL, and 5% of them have grade A.

Using the above information for optimizing a query, draw the most efficient query tree for the SQL query. You should clearly indicate all the selection and join operations on nodes in the tree. Justify your answer by calculating the size of intermediate results. **[8 marks]**

Question 5 [20 marks] Data Mining

(a) Table 3 shows a simple database with six transactions where `TID` is the transaction identifier, and `Items` are the products the customer bought. Suppose the minimum support $min\_sup = 3$. Find all frequent itemsets and list their support counts. Find all closed itemsets and list their support counts. **[9 marks]**

TID	Items
T1	{M, O, N, K, E, Y}
T2	{D, O, N, K, E, Y}
T3	{M, A, K, E}
T4	{M, U, C, K, Y}
T5	{C, O, K, I, E}
T6	{C, A, K, Y}

Table 3: A Transaction Database

(b) For an association rule $E \rightarrow C$, compute its confidence and lift. How are E and C correlated? **[4 marks]**

(c) An interesting measure for an itemset $\{a_1, a_2, \ldots, a_k\}$ is defined as

$$h(\{a_1, a_2, \ldots, a_k\}) = \frac{sup(\{a_1, a_2, \ldots, a_k\})}{\max[sup(a_1), sup(a_2), \ldots, sup(a_k)]} \quad (1)$$

Given a minimum threshold min_h, is the constraint $h(\{a_1, a_2, \ldots, a_k\}) \geq min\_h$ **monotone** or **anti-monotone**? Prove your answer. **[3 marks]**

(d) An interestingness measure for an itemset $\{a_1, a_2, \ldots, a_k\}$ is defined as

$$\zeta(\{a_1, a_2, \ldots, a_k\}) = \min \; [\; c(\{a_1\} \rightarrow \{a_2, a_3, \ldots, a_k\}), c(\{a_2\} \rightarrow \{a_1, a_3, \ldots, a_k\}), \ldots, c(\{a_k\} \rightarrow \{a_1, a_2, \ldots, a_{k-1}\}) \;],$$

where $c(\{a_1\} \rightarrow \{a_2, a_3, \ldots, a_k\}) = sup(\{a_1, a_2, \ldots, a_k\})/sup(\{a_1\})$ is the confidence of the rule $\{a_1\} \rightarrow \{a_2, a_3, \ldots, a_k\}$.

Given a minimum threshold min_ζ, is the constraint $\zeta(\{a_1, a_2, \ldots, a_k\}) \geq min\_\zeta$ **monotone** or **anti-monotone**? Prove your answer. **[4 marks]**

Question 6 [20 marks] Information Retrieval

(a) The following questions are concerned with Boolean retrieval model.

(1) The method INTERSECT presented in the recommended book "Introduction to Information Retrieval" can return the intersection operation of the elements in two posting lists. Suppose there are M_1 and M_2 elements in the two posting lists respectively. What is the minimum number of elements, in terms of M_1 and M_2, in total the method may visit? What is the maximum number of elements, in terms of M_1 and M_2, in total the method may visit? Explain your answer. **[4 marks]**

(2) Write the pseudo-code for the operation UNION(t, s) where t and s refer to two terms and the aim of this operation is to obtain the documents for the query (t OR s). **[4 marks]**

(b) Both k-nearest-neighbor (kNN) method and Rocchio method are under the family of vector space text classification models. Describe an example of a document data set for which kNN will perform better than the standard Rocchio method. **[6 marks]**

(c) The following questions are related to the standard tf-idf weighting.

(1) What is the basic idf value of a term that occurs in every document? How does the idf value of such kind of term compared with a term in a stop word list? Can we completely rely on the idf value to identify stop words? **[3 marks]**

(2) The basic idf formulation involves a logarithm function. Assume that the term weight is the standard tf-idf weighting. Consider that the retrieval score of a document d given a query is the sum, over all query terms, of the number of times each of the query terms occurs in d. Explain clearly how the base of the logarithm affects the calculation of the retrieval score. Also, explain clearly how the base of the logarithm affects the relative retrieval scores of two documents given the same query. **[3 marks]**

-End-

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

Course Examinations 2017

Course Code & Title : **PhD Qualifying Exam – Information Systems**

Time allowed : **3** **hours** **0** **minutes**

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

Question 1 [20 marks] Data Structures and Algorithms

(a) Given a string S with n lower-case letters, partition S such that every substring of the partition is a **palindrome** (a palindrome is a string that reads the same forward and backward, for example "madam" and "dad"). Return the minimum number of cuts needed for a palindrome partitioning of S. Analyze the time complexity and space complexity of your algorithm.

For example, given S = "acaddgeg", the return value is 2 since the palindrome partitioning ["aca","dd","geg"] could be produced using 2 cuts. **[10 marks]**

(b) In the same string S as defined above, find the length of the longest alphabetically increasing substring (which is NOT required to be consecutive). Analyze the time complexity and space complexity of your algorithm.

For example, given S = "acaddgeg", the longest increasing substring is "acdeg", therefore the length is 5. **[10 marks]**

Question 2 [20 marks] Operating Systems

(a) Consider the following two processes for the Producer-Consumer problem. Semaphore "full" is initialized to 0; semaphore "empty" is initialized to N which is the buffer size; and semaphore "mutex" is initialized to 1.

Producer Process:

```
do {
        ...
        produce an item nextp;
        wait(empty);
        wait(mutex);
        add nextp to buffer;
        signal(mutex);
        signal(empty);
        ...
} while(true);
```

Consumer Process:

```
do {
        ...
        wait(full);
        wait(mutex);
        remove an item from buffer to nextc;
        signal(mutex);
        signal(empty);
        consume the item nextc;
        ...
} while(true);
```

(1) Explain the purpose of the three semaphores. **[6 marks]**

(2) There is one error in the Producer/Consumer process. Point out the error and correct it. **[4 marks]**

(b) What is the cause of thrashing? How does the system detect thrashing? **[4 marks]**

(c) Consider the following sequence of page references:

1, 2, 3, 4, 5, 3, 4, 1, 6, 7, 8, 7, 8, 9, 7, 8, 9, 5, 4, 5, 4, 2.

Suppose there are 3 frames allocated for this process. Illustrate the contents of the frames under LRU page replacement algorithm step by step, and compute the number of page faults. **[6 marks]**

Question 3 [20 marks] Database

Consider the following relational database related to a car rental company:

CUSTOMER (cid, cname, address)

VEHICLE (vid, make, model, color)

RENTAL (*cid, vid*, rental-date, return-date)

In the above schemas, primary keys are underlined and foreign keys are in *italics*. All foreign keys are NOT NULL. The meaning of the attributes is self-explanatory.

The relation CUSTOMER has 5,000 tuples and 100 tuples fit into one block. The relation VEHICLE has 1,000 tuples and 100 tuples fit into one block. The relation RENTAL has 20,000 tuples and 50 tuples fit into one block.

(a) What is the size of the join operations CUSTOMER⋈RENTAL, and VEHICLE⋈CUSTOMER in terms of the number of tuples, respectively? **[6 marks]**

(b) Assume that we use block nested-loop join to join CUSTOMER and RENTAL, and the memory buffer has 4 blocks. Which relation should be the outer relation, so that the number of block transfers of the join operation is smaller? Show your calculation details. **[8 marks]**

(c) Assume that there is a B^+-tree index on the foreign key *cid* in RENTAL, and the height of the tree is 3. If we use indexed nested-loop join to compute CUSTOMER⋈RENTAL using CUSTOMER as the outer relation, what is the number of block transfers? Show your calculation details. **[6 marks]**

Question 4 [20 marks] Data Mining

Consider a set of one dimensional points {0, 200, 300, 900, 1100, 1600}.

(a) For two initial centroids $c_1 = 1100, c_2 = 1600$, create two clusters by K-means. Show the steps in K-means clustering. **[6 marks]**

(b) Compute the SSE and BSS measures of the clustering created in (a). **[4 marks]**

(c) Compute the silhouette coefficient for points 200 and 1100 respectively in the clustering created in (a). **[4 marks]**

(d) Perform hierarchical clustering using "MAX" (complete link), list the merge operations and updated distance matrices step by step, and show your results by drawing a dendrogram. The dendrogram must be clearly shown the order in which the points are merged. **[6 marks]**

Question 5 [20 marks] Information Retrieval

(a) Consider the Rocchio classification for text documents for two classes, namely, C_1 and C_2. The class C_1 contains documents d_1 and d_2, and the class C_2 contains documents d_3 and d_4. The document vectors are given as follows:

$d_1 = (0, 0.447, 0.548, 0.707, 0)$

$d_2 = (0, 0.548, 0.447, 0.707, 0)$

$d_3 = (0.710, 0, 0.710, 0, 0)$

$d_4 = (0.447, 0, 0.707, 0.548, 0)$

(1) Suppose that we wish to classify a document with vector $d_t = (0.548, 0, 0.707, 0, 0.447)$. Illustrate clearly the process of classification. Show all intermediate steps and assumptions clearly. **[5 marks]**

(2) For a general two-class Rocchio classification model, state the main property of the decision boundary to the class centroids. **[2 marks]**

(3) Describe one drawback of a general Rocchio classification. Explain your answer by drawing a diagram showing a sample set of training documents. **[3 marks]**

(b) The following questions are concerned with the evaluation of unranked retrieval sets.

(1) Explain the meaning of "true positive", "false positive", "false negative", and "true negative". **[2 marks]**

(2) Define the accuracy metric using the concepts mentioned in (1) above. **[1 mark]**

(3) Give a sample value of each of "true positive", "false positive", "false negative", and "true negative" to illustrate why accuracy is not a good metric for IR. **[3 marks]**

(4) Define precision and recall from the concepts mentioned in (1) above. **[2 marks]**

(5) Give a sample information retrieval scenario where precision is a preferred evaluation metric than recall. **[2 marks]**

-End-

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

Course Examinations 2018

Course Code & Title : **PhD Qualifying Exam – Information Systems**

Time allowed : **3** **hours** **0** **minutes**

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

Question 1 [20 marks] Data Structures and Algorithms

(a) Given a binary tree, we can traverse the tree in Pre-order, In-order or Post-order, and the corresponding visiting sequence will be generated. Assume that you are given two visiting sequences only without the binary tree, determine whether you can reconstruct the tree uniquely according to the sequences, for each case below. If yes, briefly explain why. If no, create a counter-example. **[9 marks]**

(1) Pre-order sequence + In-order sequence

(2) Post-order sequence + In-order sequence

(3) Pre-order sequence + Post-order sequence

(b) Given a binary tree, the distance between two nodes is the minimum number of edges to be traversed from one node to reach the other. For example, for the binary tree in Figure 1, the distance between f and e is 3. The longest distance in the tree is 6, which is between h and i.

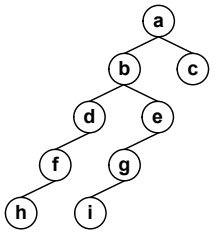


Figure 1: A binary tree example

Design an algorithm to find the longest distance in a binary tree with $O(n)$ time complexity, where n is the number of nodes in the tree. **[11 marks]**

Question 2 [20 marks] Operating Systems

(a) Consider a memory system with paging. The page table has 256 entries. The physical address has 32 bits; there are a total of 2^{16} frames in the physical memory. The following shows a part of the page table. Answer the following questions.

Page number	Frame number
0	0xA3C5
...	...
17	0xB235
...	...
32	0x4E86
...	...
98	0xC2EA
...	...
107	0x92AE
...	...
111	0xA89C
...	...
132	0x4455
...	...
173	0x22E5
...	...
211	0xEAC3
...	...
255	0x4A6C

Table 1: Page Table

(a.1) What is the size of the physical memory? **[2 marks]**

(a.2) How many bits are there in the logical address? **[2 marks]**

(a.3) What is the physical address (in hex) that corresponds to virtual address 0xD3E2F7? **[2 marks]**

(a.4) What is the virtual address (in hex) that corresponds to physical address 0xA89C78C2? **[2 marks]**

(b) Consider the following C program. How many lines of output does the function print_hellos print? **[6 marks]**

```
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>

void print_hellos(int n){
        int i;
        for(i = 0; i < n; i++){
                fork();
        }
        printf("hello\n");
        exit(0);
}

int main(){
        int n = 10;
        print_hellos(n);
        return 0;
}
```

(c) Consider the following sequence of page references in terms of page numbers:

1, 3, 2, 3, 4, 1, 5, 2, 5, 6

Suppose there are 3 frames allocated for this process, illustrate the contents of the frames step by step under FIFO and LRU page replacement algorithms, respectively, and compute the number of page faults for each algorithm. **[6 marks]**

Question 3 [20 marks] Database

Consider the following two relations R and S. The number of records in R and S is also given.

$R(A, B, C)$: 18,000 records

$S(A, D, E)$: 48,000 records

Assuming the size of each memory page is 3,000 bytes and the memory buffer has 36 pages. For simplicity we assume all attribute values are 10 bytes.

(a) What is the size of R and S in terms of pages? **[4 marks]**

(b) Consider the following SQL query:

select A

from R

where $B = 20$;

Assume that there is a B^+-tree index on attribute B, and a B^+-tree node contains 20 pointers to its children nodes. If we use the B^+-tree to process the above query, and there are 10 records as the answer. What is the largest possible number of block access to process the above query? **[6 marks]**

(c) Consider the following SQL query:

select B, E

from R, S

where $R.A = S.A$;

Let h be a hash function that assigns a given record in one of 6 partitions. Assume using hash join to process the query as follows:

Step 1: Partition R using h on $R.A$ and write the partitions to the disk.

Step 2: Partition S using h on $S.A$ and write the partitions to the disk.

Sept 3: For each partition of R, we transfer it from disk to memory buffer and match it against the corresponding partition of S (page by page). The matching can be done by an in-memory hash built on $R.A$.

Estimate the page read/write cost of each step and then compute the total cost of hash join. **[5 marks]**

(d) One further optimization of Part(c) is to keep the first partition of R in the memory buffer in Step 1 and then use the rest of memory to generate only the other partitions in Step 2. The records that belong to the first partition of S are matched directly with all records of the first partition of R on-the-fly. The comparison in other partitions (the second, third and so on) follows Step 3.

Estimate the total cost of the hash join using this optimization. **[5 marks]**

Question 4 [20 marks] Data Mining

Consider Table 2 for a binary classification problem. The first column shows the instance id, the second and third columns show the values of attributes A and B respectively, and the fourth column shows the true class label of each instance.

ID	A	B	Class Label
x_1	T	T	+
x_2	T	T	+
x_3	F	T	-
x_4	T	F	+
x_5	T	F	+
x_6	F	T	-
x_7	F	T	-
x_8	F	F	+
x_9	F	F	-
x_{10}	F	F	-

Table 2: A Training Data Set for Decision Tree Induction

(a) According to the Gini index criterion, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gain in Gini index. **[6 marks]**

(b) Use the remaining attribute to split in the second level of the tree for nodes that have both positive and negative instances. Draw the two-level decision tree. Mark the class label in each leaf node. **[3 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.) **[5 marks]**

(d) Assume there is a classifier that predicts the probability of belonging to the positive class for each instance, i.e., $P(+|x_i)$, in Table 3. According to the predicted probability, plot the ROC curve and calculate the Area under the Curve (AUC). **[6 marks]**

ID	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
$P(+\|x_i)$	0.90	0.85	0.82	0.80	0.75	0.70	0.50	0.45	0.40	0.35

Table 3: Predicted Probability of Belonging to the Positive Class

Question 5 [20 marks] Information Retrieval

(a) Consider the following fragment of a positional index with the format:

word – document: <position, position, ... >; document: <position, position, ... >; ...

Gates – 1: <3>; 2: <6>; 3: <2, 17>; 4: <1>;

IBM – 4: <3>; 7:<14>;

Microsoft – 1: <1>; 2: <1, 21>; 3: <3>; 5: <16, 22, 51>;

For the '$/k$' operator, "word1 $/k$ word2" finds occurrences of word1 within k words of word2 (on either side), where k is a positive integer argument. Thus $k = 1$ demands that word1 be adjacent to word2. Describe the set of documents and term position information that satisfy the query "Gates /2 Microsoft". **[5 marks]**

(b) Suppose that we employ B-trees for storing existing terms in a dictionary.

(1) Give an outline how a standard B-tree can get all terms satisfying a trailing wildcard query. An example of a trailing wildcard query is "bird$*$". **[2 marks]**

(2) Suppose that we wish to handle leading wildcard queries. An example of a leading wildcard query is "$*$bird". Use this example to show how we should build a B-tree that is useful for handling leading wildcard queries. **[3 marks]**

(c) Consider a text classification problem involving two classes. Suppose we employ the standard Rocchio text classification method. Write the pseudo-code for the training process and the testing process. You can assume that there is a cosine similarity function available for use. **[10 marks]**

-End-

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

二0一八至一九年度

科目編號及名稱
Course Code & Title : **PhD Candidacy Examination – Information Systems**

時間 小時 分鐘
Time allowed : 3 hours 0 minutes

學號
Student I.D. No. :

Question 1 [20 marks] Data Structures and Algorithms

(a) Consider a binary tree T with n nodes. Suppose we use the following data structure to store a tree node:

```
struct TreeNode{
        int value;
        TreeNode *left, *right;
};
```

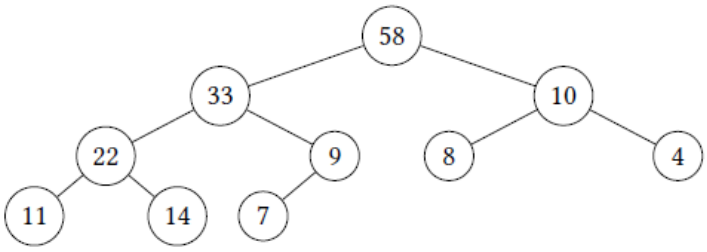
Design an algorithm to verify if a binary tree T is a binary search tree or not. What is the time complexity of your algorithm? (8 marks)

```
int isBST(struct TreeNode* node){



}
```

(b) Given the following max-heap T_1,



(i) Draw the max-heap after inserting 37 into T_1. What is the time complexity of max-heap insertion if the number of nodes in the max-heap is n? (3 marks)

(ii) Draw the max-heap after deleting the root from T_1. (3 marks)

- Given a hash table T of size 7 and a hash function $h(x) = x\%7$, insert 10, 2, 12, 19, 9, 47 into T (use linear probing to resolve collisions if any). Show table T after the insertion of all the elements. After all insertions are done, show the locations examined in order when searching for 16 and 47 separately in T. (6 marks)

Question 2 [20 marks] Operating Systems

(a) Consider the following sequence of page references:

1, 2, 3, 4, 5, 1, 5, 1, 6, 7, 8, 5, 8, 9, 2, 4, 5, 4, 2, 9.

Suppose there are 3 frames allocated for this process. Illustrate the contents of the frames under LRU page replacement algorithm step by step, and compute the number of page faults. (6 marks)

(b) Consider the following snapshot of a system:

	Allocation						Max Request				
	A	**B**	**C**	**D**	**E**		**A**	**B**	**C**	**D**	**E**
P1	2	3	0	0	2		5	3	2	1	6
P2	1	0	3	2	0		1	4	9	2	0
P3	0	2	1	3	2		0	2	8	6	3
P4	4	3	0	3	3		6	10	0	3	3
P5	1	1	3	0	1		4	3	5	0	1

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

(ii) Assume the available resources are <A:3, B:4, C:3, D:1, E:1>. Is the system in a safe state? If so, list ALL the possible safe sequences. (5 marks)

(c) Consider a single-level paging scheme with a translation look-aside buffer (TLB) used to store part of the page table. Assume that each memory access time is 200 nanoseconds, TLB access time is 20 nanoseconds and TLB hit ratio is 0.8. What is the effective memory access time? (4 marks)

Question 3 [20 marks] Database Systems

Consider a movie database with the following schema:

- ACTOR (AID, AName, Gender, DOB)
- MOVIE (MID, MName, Year, Profit)
- ROLE (AID, MID, RoleName, Pay)

ACTOR and MOVIE record information about actors and movies, respectively. Whenever an actor is cast in a movie, the ROLE table records the actor's role and pay in the movie. The relation ACTOR has 5,000 tuples and 100 tuples fit into one block; relation MOVIE has 1,000 tuples and 100 tuples fit into one block; the relation ROLE has 100,000 tuples and 100 tuples fit into one block. Answer the following questions.

(a) Represent the following queries with relational algebra. You may use the following operators:
σ (selection), Π (projection), ∪ (set union), ∩ (set intersection), − (set difference),
← (assignment), ρ (rename), ⋈ (natural join), G (grouping and aggregation)

(i) Find the names of the female actors who have appeared in at least one movie with a profit over 1,000,000. (2 marks)

(ii) For every male actor, list his AID, the movies that he appeared in from 2009 to 2019, as well as the total pay he received during this period. (3 marks)

(b) Write the SQL query for a.i and a.ii, respectively. (5 marks)

(c) Assume that we use block nested-loop join to join ACTOR and ROLE. Further assume that the block nested loop join includes the following optimization: If the memory buffer has M blocks, we read in $M - 2$ blocks of the outer relation at a time, and when we read each block of the inner relation we join it with all the $M - 2$ blocks of the outer relation.

(i) If the memory buffer has 4 blocks, which relation should be used as the outer relation so that the IO cost (in terms of the number of block transfers and disk seeks) of the join operation is smaller? Show your calculation details. (6 marks)

(ii) If the memory buffer has 60 blocks, what is the IO cost (in terms of the number of block transfers and disk seeks) if ACTOR is used as the outer relation? Show your calculation details. (4 marks)

Question 4 [20 marks] Data Mining

(a) The following table shows a transaction database where TID is the transaction identifier, and Items are the products a customer bought.

TID	Items
T1	{a, c}
T2	{b, c, d}
T3	{a, b, d, e}
T4	{a, d, e}
T5	{b, c, d, e}
T6	{c, d, f}
T7	{b, d, e}
T8	{c, d, e}

(i) Suppose the minimum support count min_sup=3. Among all frequent itemsets, list the closed itemsets and maximal itemsets and their support counts. (5 marks)

(ii) For an association rule b→de, compute its support, confidence and lift. (3 marks)

(b) Consider a set of one dimensional points {10, 20, 30, 40, 50, 60}.

(i) For three initial centroids c1=10, c2=20, c3=30, create three clusters by K-means. Show the steps in K-means clustering. (6 marks)

(ii) Compute the SSE and BSS measures of the clustering created in b.i. (6 marks)

Question 5 [20 marks] Information Retrieval

(a) Consider an information need for which there are 4 relevant documents in total in the collection. The top 10 retrieved documents are judged for relevance as shown below. The leftmost item is the top ranked search document. R and N denote relevant and non-relevant document respectively.
Result: R N R N N N N N R R

(i) Calculate the Precision and Recall if we only consider the top 5 retrieved documents. (4 marks)

(ii) Calculate the Mean Average Precision (MAP) of this single query, i.e. considering all 10 retrieved documents. (2 marks)

(b) (i) What is isolated-term spelling correction? (3 marks)

(ii) Given a set V of strings corresponding to terms in the vocabulary and a query string q, describe an outline of a method for conducting spelling correction by using edit distance. (Assume that the edit distance function is available.) (3 marks)

(c) Consider the following documents in the training set. For simplicity, each document is represented as a 2-dimensional vector. There are two classes denoted as "A" and "B".

class A: (1,1); (1,3); (3,3)

class B: (3,1); (5,1)

Draw the Voronoi tessellation (using single lines) and decision boundaries (using double lines) for 1-Nearest-Neighbor (1NN) classification model. (8 marks)

End

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

Course Examination 2020

科目編號及名稱
Course Code & Title : **PhD Qualify Exam - Information Systems**

時間 Time allowed : 3 小時 hours 0 分鐘 minutes

學號 Student I.D. No. : _____ 座號 Seat No. : _____

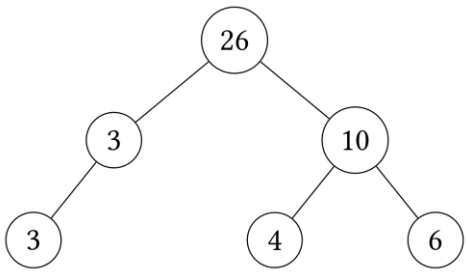
*Candidates whose primary/secondary area is IS should answer all the below questions.

Question 1 [13 marks] Data Structures and Algorithms

I. [7 makrs] Given a binary tree T, assume that we use the below structure to store a tree node:

```
struct TreeNode{
        int value;
        TreeNode *left, *right;
};
```

Design an algorithm to check if T is a *sum tree*, i.e., a binary tree such that the *value* field of each non-leaf node is equal to the sum of the *values* of all nodes in both its left subtree and right subtree. Below shows an example of a sum tree. The root node has a value of 26 since the sum of the values of all nodes on the left subtree and right subtree is 26.




```
int isSumTree(struct TreeNode* node){
//write your code here
}
```

II. [6 makrs] Given a hash table T of size 11. Assume that we use double hashing to resolve collisions. The double hashing function is $h(k,i) = \big(h(k) + i \cdot h'(k)\big)\%11$, where $h(k) = k\%11$ and $h'(k) = 1 + k\%5$. Show the table T after inserting 11, 13, 5, 12, 27, 24, 33, 9 in order. After all insertions, show the locations and keys examined in order when searching for 44.

III. [7 marks] There are two sorted arrays A_1 and A_2 of size n_1 and n_2 respectively. Design an algorithm to find the medium of the two sorted arrays. You may just describe the main idea of your algorithm and show your analysis of the time complexity. Notice that we do not accept a straightforward solution with $O(n_1 + n_2)$ (or even higher) time complexity.

Example 1: $A_1 = [1,2]$, $A\_2 = [3,4,5]$: the medium is 3.

Example 2: $A_1 = [1,2]$, $A_2 = [3,4]$, the medium is $(2 + 3)/2 = 2.5$.

Example 3: $A_1 = [1,2]$, $A_2 = [2,4]$, the medium is $(2 + 2)/2 = 2$.

Question 2 [20 marks] Operating Systems

I. Briefly explain when deadlock and starvation occurs. [3 marks]

II. Explain the difference between a virtrual memory address and a physical memory address [3 marks]

III. Given an integer shared variable x initialized as 1, assume that two threads are updating x as shown in the sampled code and do not update x elsewhere.

a. What are the possible values of x when Line 1.1 finishes without incurring any data race? [2 marks]

b. Given the compare&swap atomic instruction, design an **atomic_add** function so that we can replace Line 1.1 and Line 2.1 with atomic_add(&x, 2) and atomic_add(&x, 5) without bringing data races. You may use the following interface:

bool compare_and_swap(int word, int testval, int newval):* It compares the contents of a memory location (*word) with testval, and only if they are the same, modifies the content of that memory location to newval and returns true. Otherwise, it does not update the value and returns false [5 marks]

```
Thread 1:
…
x = x + 2;  (Line 1.1)
…
```

```
Thread 2:
…
x = x + 5;   (Line 2.1)
…
```

```
bool atomic_add(&word, value){
//write your code here
}
```

IV. Consider the following sequence of page references: 0, 2, 3, 7, 1, 3, 2, 3, 7, 6, 5, 3, 2, 6, 5, 6. Suppose that there are 3 frames allocated for this process. Illustrate the contents of the frames under the LRU and FIFO page replacement algorithm step by step, respectively. Compute the number of page faults for each page replacement algorithm. [7 marks]

Question 3 [20 marks] Database Systems

Consider a database that stores information about tennis players, tennis courts, and tennis matches, with the following schema:

PLAYER(PID, Name, Gender, DOB)
COURT(CID, Location, Type)
MATCH(MID, P1, P2, CID, Date, P1Wins)

PLAYER records information about tennis players. COURT records the ID and location of each tennis court, as well as its types (i.e., whether it is grass, hard, or clay). MATCH records the result of each match. In particular, MATCH.P1 and MATCH.P2 are foreign keys referencing PLAYER.PID, while MATCH.CID is a foreign key referencing COURT.CID. PLAYER.P1Wins is a Boolean attribute that takes value TRUE if player P1 wins the match.

Answer each of the following questions.

(a) Represent the following queries with relational algebra. You may use the following operators: σ (selection), Π (projection), ∪ (set union), ∩ (set intersection), − (set difference), ← (assignment), ⋈(natural join), G (grouping and aggregation). You may further use "X as Y" to rename attribute from X to Y for simplicity, e.g., for aggregations.

I. Find the CIDs of tennis courts on which there have been at least 100 matches. [3 marks]

II. For each tennis player, list the number of matches that he/she has won. [4 marks]

(b) Write the SQL query for (a).I and (a).II, respectively. [6 marks]

(c) Assume that relation COURT includes 1000 records and 50 records fit into a block; relation MATCH includes 50000 records and 40 records fit into one block. Further assume that MATCH has a B^+-Tree on the foreign key CID, which contains 20 entries per index node. If we use the indexed nested-loop join to compute COURT⋈MATCH, what is the number of *block transfers*? Show your calculation details. [7 marks]

Question 4 [20 marks] Data Mining

(i) Consider the table below for a binary classification problem. The first column shows the instance id, the second and third columns show the values of attributes A and B respectively, and the fourth column shows the true class label of each instance.

ID	A	B	Class label
x_1	1	0	+
x_2	0	1	+
x_3	0	0	+
x_4	1	1	+
x_5	0	1	-
x_6	0	1	-
x_7	1	0	+
x_8	0	0	-

Table 1: A Training Dataset

(a) If we use the decision tree for the classification task and Gini index is selected as the splitting criteria. Which attribute, A or B, will be chosen as the first splitting criteria? Please justify your answer by showing the gain of the Gini index when choosing each attribute. [5 marks]

(b) If the Naïve Bayes classifier is trained based on the above training dataset, what will be the predicted class label for a new record x_9 with $A = 1, B = 0$. Please show your calculated probability for each class. [5 marks]

(ii) Consider a set of one dimensional points $\{7, 13, 20, 25, 30, 42, 50, 60\}$.

(a) If we choose 7, 50, 60 as the initial centroid, show the steps in k-means clustering. [5 marks]

(b) Perform the single-linkage clustering for the above data points. Show the merge operations step by step and draw the dendrogram. [5 marks]

Question 5 [20 marks] Information Retrieval

(i) For measuring ad hoc information retrieval performance, a typical way is to prepare a test collection comprised of three core components. State these three core components. [6 marks]

(ii) When we increase the number of documents retrieved, typically how precision varies and how recall varies? [3 marks]

(iii) F-measure is the weighted harmonic mean of precision and recall. Why does a harmonic mean is more preferred than the arithmetic mean? [3 marks]

(iv) We consider Bayesian text classification from the perspective of generative process. Suppose that C denotes the set of all classes. We decide class membership of a document d by assigning it to the class, denoted as g, with the maximum a posterior probability. [8 marks]

a. Write the equation that can determine g using d and C as input. [3 marks]
b. Describe your equation in (i) above as a generative process. Also describe the two different document representations for two variants, namely, multinomial model and Bernoulli model. [3 marks]
c. Among multinomial model and Bernoulli model, which one requires feature selection more? State a reason for that. [2 marks]

End

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

Course Examination 2021

科目編號及名稱
Course Code & Title : **PhD Qualify Exam - Information Systems**

時間 小時 分鐘
Time allowed : 3 hours 0 minutes

學號 座號
Student I.D. No. : Seat No. :

---

*Candidates whose primary/secondary area is IS should answer all the below questions.

Question 1 [20 marks] Data Structures and Algorithms

I. [7 marks] Given the output of the inorder traversal and postorder traversal of the same binary tree T as shown below, draw the binary tree T.
   a. Inorder traversal output of T: 48, 22, 9, 15, 28, 3, 33, 37, 17
   b. Postorder traversal output of T: 48, 9, 22, 28, 33, 3, 17, 37, 15

II. [7 marks] You are given a sorted array A containing n distinct integers in the range from 1 to $n+1$. Please design an algorithm to find in $O(\log n)$ time the only number $x \in [1, n+1]$ that is missing in A. For example, if $n = 5$, and A stores 1,3,4,5,6, then $x = 2$.

III. [6 marks] Consider a directed graph where the weights of its edges can be only 1, 2, or 3. Design an algorithm to compute the shortest path from a source to the destination with $O(n+m)$ running cost. No pseudo-code is required and you may only describe your main idea.

Question 2 [20 marks] Operating Systems ch 8

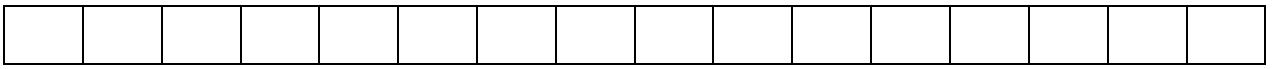
I. [7 marks] Assuming a page size of 2K bytes (K=1024) and that a page table entry takes 4 bytes, how many levels of page tables would be required to map a 32-bit address space, if the top-level page table fits into a single page? Explain why.

II. [7 marks] Table 1 shows a set of processes and their associated arrival time and running time. ch9

Process ID	Arrival Time	Running Time
1	0	2
2	1	6
3	4	1
4	7	4
5	8	3

Table 1: Info of processes

Show the scheduling order for these processes under the Round-Robin schedule with timeslice quantum = 1, by filling in the Gantt chart (as shown below) with the ID of the process currently running in each time quantum. Assume that the context switch overhead is 0 and that new Round-Robin processes are added to the head of the queue. For example, if we have two processes A and B, where A has an arrival time of 0 and B has an arrival time of 1. Both A and B have a running time of 4 timeslices. Then, the Gatta chart will be A B A B A B A B for time 0-8 under the Round-Robin schedule.



0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

III. [6 marks] Consider the following incomplete implementation of a spin-lock.

```
typedef struct __lock_t {
        int flag;
} lock_t;

void init(lock_t *lock) {
        lock->flag = abc;
}
void acquire(lock_t *lock){
        while(TestAndSet(&lock->flag, xyz)==xyz)
                ; // spin-wait (do nothing)
}
void release(lock_t *lock) {
        lock->flag = 1;
}
```

TestAndSet(int*old_pointer, int new_value) is an atomic instruction that returns the old value stored at old_pointer and simultaneously updates its stored value to new_value. For example, if B is initially 2, then TestAndSet(&B, 3) will set B to 3 and simultaneously return 2, which is the old value of B. Answer the following questions about the spin-lock.

- To what value should the lock->flag be initialized (shown as abc in the above code)? [2 marks]
- To what value should the lock->flag be tested and set (shown as xyz in the above code)? (There might exist multiple valid answers and you are only required to provide one valid solution) [2 marks]
- List one disadvantage of spin-lock and briefly explain how to avoid this disadvantage? [2 marks]

Question 3 [20 marks] Database Systems

Consider a sales database with the following schema:

- PRODUCT (PID, Name, Category)
- EMPLOYEE (EID, Name, DOB)
- CUSTOMER (CID, Name, Address)
- TRANSACTION (EID, PID, CID, Price, Quantity, Date)

The PRODUCT, EMPLOYEE, and CUSTOMER tables record information about products, employees, and customers, respectively. The TRANSACTIONS table records which employee sells which customer which product on which date, as well as the unit price and quantity of the product in the transaction.

Answer each of the following questions.

I. Represent the following queries with relational algebra. You may use the following operators: σ (selection), Π (projection), ∪ (set union), ∩ (set intersection), − (set difference), ← (assignment), ⋈(natural join), G (grouping and aggregation). You may further use "X as Y" to rename attributes from X to Y for simplicity, e.g., for aggregations.

   a. Find the EIDs and Names of the employees who have sold products in the "Toy" category to customers whose addresses are "Shatin". [2 marks]
   b. List the EID and Name of each employee, as well as the total amount of money involved in the transactions that he/she has made.

II. Write the SQL query for I.a and I.b, respectively. [6 marks]

III. Assume that relation EMPLOYEE includes 2000 records and 100 records fit into a block; relation TRANSACTION includes 100,0000 records and 50 records fit into one block. Assume that the memory capacity is 50 blocks and that we do a join operation between EMPLOYEE and TRANSACTION. [7 marks]

   a. What relation should be selected as the outer relation for the nested-loop join and block nested-loop join to reduce the I/O cost. [2 marks]
   b. Given your choice of outer relation, what is the number of block transfers and block seeks for EMPLOYEE ⋈ TRANSACTION by the nested-loop join and block nested-loop join, respectively? Show your calculation details. [5 marks]

Question 4 [20 marks] Data Mining

I. Consider a transactional database where a, b, c, d, e, f, g are items.

ID	Items
t_1	a, b, c, e
t_2	a, b, c, d, e
t_3	a, b, c, g
t_4	a, c, f
t_5	a, b, d, e, f

Table 2: the set $\mathcal{S}$ of transactions

a. Suppose the minimum support is 3, find all frequent itemsets according to set $\mathcal{S}$. [3 marks]

b. If the minimum support is 3 and the minimum confidence is 75%, list all association rules according to set $\mathcal{S}$. [4 marks]

II. What is model overfitting? Show two examples that try to avoid model overfitting. [4 marks]

III. Consider ten 1-dimension records $p_1 = 1, p_2 = 3, p_3 = 4, p_4 = 6, p_5 = 8, p_6 = 13, p_7 = 15, p_8 = 17, p_9 = 20, p_{10} = 25$.

a. If we choose $p_1 = 1, p_3 = 4, p_6 = 13$ as the initial centroid ($k = 3$), show the steps in k-means clustering. [4 marks]

b. Finish the clustering task using DBSCAN with $\epsilon = 2$, $MinPts = 3$. Please show the ϵ-neighbor of each point, identify the point type of each point, construct the core-point graph, and show the final clusters. [5 marks]

Question 5 [20 marks] Information Retrieval

I. [7 marks] The following is the INTERSECT algorithm for the Boolean retrieval model as described in the reference book.

```
INTERSECT(p1, p2)
1   answer ← emptylist
2   while p1 is not NIL and p2 is not NIL
3   do if docID(p1) = docID(p2)
4         then ADD(answer, docID(p1))
5              p1 ← next(p1)
6              p2 ← next(p2)
7         else if docID(p1) < docID(p2)
8              then p1 ← next(p1)
9         else p2 ← next(p2)
10  return answer
```

This algorithm returns the intersection operation of the elements in two posting lists. It is quite an efficient method. Suppose there are M1 and M2 elements in the two posting lists respectively.

a. What is the minimum number of elements, in terms of M1 and M2, in total the method may visit? Furthermore, suppose that M1 is 3 and M2 is 5. Draw sample posting lists and explain your answer.

b. What is the maximum number of elements, in terms of M1 and M2, in total the method may visit? Furthermore, suppose that M1 is 3 and M2 is 5. Draw sample posting lists and explain your answer.

II. [6 marks] In multi-modal text classification problems, some or all classes exhibit internal multi-modal distribution. For such problems, the k-nearest-neighbor (kNN) method generally handles better than the standard Rocchio method. Use an example of a data set to illustrate that 3-NN is better than Rocchio.

III. [7 marks] Consider multinomial Naïve Bayes text classification. Describe clearly how to estimate $P(x|c)$ where x denotes a word and c denotes a class. Assume that "Add 1" smoothing is employed.

End