請勿攜去
Not to be taken away

香港中文大學
The Chinese University of Hong Kong

二〇一八至一九年度

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

時間
Time allowed : 3 小時 hours 0 分鐘 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:

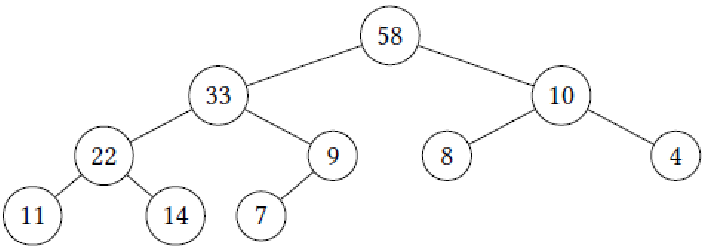
```
struct TreeNode{
        int value;
        TreeNode *left, *right;
};
```

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), $\cup$ (set union), $\cap$ (set intersection), $-$ (set difference), $\leftarrow$ (assignment), ρ (rename), $\bowtie$ (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