Course Code : CST 406 HJRT/MS – 17 / 2296

Eighth Semester B. E. (Computer Science and Engineering) Examination

DATA WAREHOUSING AND MINING

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) All questions carry marks as indicated against them.
(2) Number your answers properly.
(3) Assume suitable data and illustrate answers with neat sketches wherever necessary.

1. Solve any **Two** :—

(a) List the problems faced in maintaining data quality at instance and schema level by giving two examples each of single source and multi source data. 5 (CO 1)

(b) Define Multidimensional Cube. How do they apply in OLAP system ? Why populating data into OLAP system from operational systems is inappropriate ? 5 (CO 1)

(c) Consider the following schema :

SALES (time_key, item_key, branch_key, location_key, units_sold, dollars_sold)

BRANCH (branch_key, branch_name, branch_type)

LOCATION (location_key, street, city, state, country)

ITEM (item_key, item_name, brand, type, supplier_type)

TIME (time_key, day, day_of_week, month, quarter, year)

Construct the following queries in SQL :

— roll – up on total sales by year and by quarter.
— roll – up on total sales by country, by state, and by city.
— roll – up on total sales by item brand by item type (digital or analog).
— drill down on total sales by month and by day.
— drill down on total sales by street address. 5 (CO 2)

2. (a) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit :

(a) Construct a star schema diagram for the above data warehouse. Also list the concept hierarchies.

(b) Starting with the base cuboid [day, doctor, patient], what specific OLAP operation should be performed in order to list the total fee collected by each doctor in 2010 ?

(c) Compute how many cuboids will be generated from the base cuboid [day, doctor, patient] if each dimension has 4 hierarchy levels including all (*) ? 5 (CO 2)

(b) Describe Relational OLAP(ROLAP) and Multidimensional OLAP(MOLAP) systems. 5 (CO 1)

OR

(c) Explain degenerate and junk dimensions with examples. Explain what a factless fact is. Explain additive, semi – additive, and non – additive facts with the help of an example. 5 (CO 1)

3. (a) Rebuilding of index can improve the clustering factor. Comment on this statement. What inferences can be drawn from your argument ? 2 (CO 1)

(b) Describe what will the following commands do ?

(i)
```
CREATE CLUSTER trial_cluster (trialno NUMBER(5, 0))
TABLESPACE users
STORAGE(INITIAL 250 K  NEXT 50K
MINEXTENTS 1  MAXEXTENTS 3
PCTINCREASE 0)
HASH IS trialno HASHKEYS 150 ;
```

(ii)
```
DROP CLUSTER emp_dept INCLUDING TABLES CASCADE
CONSTRAINTS ;
```
3 (CO 2)

(c) Find all instructors in finance dept with salary of 80,000 or more. Outline the steps in answering the query, and demonstrate the final and intermediate bitmaps constructed to answer the query.

ID	name	dept_name	salary
10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	EI Said	History	60000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	62000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000

5 (CO 2)

Solve Q. **Four** or Q. **Five** :—

4. (a) In real–world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. 4 (CO 3)

(b) Table 1 shows attributes of two classes of planes : civil and military. Solve the following :

Table 1

Plane Weight (*10^3kg)	Plane Velocity (km/h)	Plane Type
26	1460	War
33	1450	War
36	1550	War
37	1350	War
72	1564	War
505	595	Civil
477	825	Civil
590	600	Civil

(i) Construct Boxplot for plane weight.

(ii) Use min–max normalization to transform the value 400 (*10^3 kg) for plane weight onto range [0·0, 1·0].

(iii) Use z–score normalization to transform the value 400 (*10^3 kg) for plane weight. 6 (CO 3)

5. (a) Consider the following table showing continuous feature F with corresponding class K :

Sample	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

Apply ChiMerge algorithm for discretization to generate the final reduced intervals. (Given : degree of freedom $d = 1$, $\chi_2 = < 2{\cdot}706$ for threshold, $\alpha = 0{\cdot}1$) 10 (CO 3)

Solve Q. **Six** or Q. **Seven** :—

6. (a) Consider the grocery store example with support threshold $s = 33.34\%$ and confidence threshold $c = 60\%$. Build a frequent pattern tree (EP – Tree) and show for each transaction how the tree evolves. Use Fp–Growth to generate the frequent itemsets form this FP – tree.

Transaction ID	Items
T1	HotDogs, Buns, Ketchup
T2	HotDogs, Buns
T3	HotDogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	HotDogs, Coke, Chips

10 (CO 4)

7. (a) The following table consists of training data from an employee database. The data have been generalized. For example, "31....35" and "46K....50K" stands for the age range of 31 to 35 years old and the salary range of 46K to 50K pounds.

department	status	age	salary
sales	senior	31....35	46K....50K
sales	junior	26....30	26K....30K
sales	junior	31....35	31K....35K
systems	junior	21....25	46K....50K
systems	senior	31....35	66K....70K
systems	junior	26....30	46K....50K
systems	senior	41....45	66K....70K
marketing	senior	36....40	46K....50K
marketing	junior	31....35	41K....45K
secretary	senior	46....50	36K....40K
secretary	junior	26....30	26K....30K

Use the data in the above table to train a naïve Bayesian classifier using status attributes as class label and regarding all the remaining attributes as input. Test the following unseen instances :

(i) (Marketing, 31....35, 46K....50K)

(ii) (sale, 31....35, 66K....70K)

on your naïve Bayesian classifier and generate the results. 10 (CO 4)

Solve Q. **Eight** or Q. **Nine** :—

8. (a) Use single–link, complete–link agglomerative clustering to create the clusters of following 8 points :

A1 = (2, 10), A2 = (2, 5), A3 = (8, 4), A4 = (5, 8), A5 = (7, 5), A6 = (6, 4), A7 = (1, 2), A8 = (4, 9).

The distance matrix is

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

7 (CO 4)

(b) Explain about classifier accuracy. Explain the process of measuring the accuracy of a classifier ? 3 (CO 3)

9. (a) Discuss the characteristics of clusters K–Medoids and K–means are trying to find. What can be said about the optimality of the clusters they find ? Both algorithms are sensitive to initialization; explain why this is the case. 3 (CO 3)

(b) Apply PAM (Partition Around Medoids) algorithm to the following data points (show 2 iterations) :

X1(2, 6), X2(3, 4), X3(3, 8), X4(4, 7), X5(6, 2), X6(6, 4), X7(7, 3), X8(7, 4), X9(8, 5), X10(7, 6) 7 (CO 4)

Course Code : CST 407 **HJRT/MS – 17 / 2297**

Eighth Semester B. E. (Computer Science and Engineering) Examination

INFORMATION SECURITY

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) All questions carry marks as indicated against them.
(2) Due credit will be given to neatness and adequate dimensions.
(3) Assume suitable data and illustrate answers with neat sketches wherever necessary.

1. (a) Which security mechanism(s) are provided in each of the following cases ?

(i) A company demands employee identification and a password to let employee log into the company server.

(ii) A company server disconnects an employee, if he is logged into the system for more than two hours.

(iii) A teacher refuses to send students grades by email unless they provide identification assigned by the teacher.

(iv) A bank requires the customer's signature for a withdrawal. 4(CO1)

OR

(b) How are cryptographic system characterized ? Explain each characteristic in one line. 4(CO1)

(c) Classify the different types of attacks on information security. Clearly give the examples to illustrate the classification of above attacks. 6(CO1)

2. (a) Use the playfair cipher to encipher the message "The key is hidden under the door pad". The secret key can be created by filling the first and part of the second row the word "GUIDANCE" and filling the rest of the matrix with the rest of the alphabet. 5(CO1)

(b) What do you mean by differential and linear cryptanalysis ? How differential cryptanalysis can be used by an attacker to guess the key values by using the difference relation between plaintext and cipher text. Give the analogy. 5(CO1)

OR

(c) Show the design of product cipher (at least consisting of 2 rounds) which makes use of the invertible, non-invertible and self-invertible component ? How this design is extended for the design of modern block cipher DES ? 5(CO1)

3. (a) Differentiate between statistically random numbers and pseudo random number. Write and explain the ANSI X9.17 Pseudorandom Number Generator. 5(CO1,2,4)

OR

(b) Which public key cryptography algorithm is used for selection of key pairs? Provide its mathematical proof to recover plain text. 5(CO1,2,4)

(c) In the Diffie-Hellman protocol, what happens if x and y have same value, that is Rita and Shyam have accidentally chosen the same number ? Are R1 and R2 the same ? Do the session keys calculated by Rita and Shyam have the same value ? Give an example to prove your claims. 5(CO1,2,4)

4. (a) What are the security requirements for cryptographic Hash function ? Clearly describe how these requirements can be satisfied by a secure hash function. 6(CO3,4)

OR

(b) What are Message Detection Code (MDC) and Message Authentication Code (MAC) ? Predict how Hash-based message authentication code is useful. 6(CO3,4)

(c) In what order should be the signature function and the confidentiality function be applied to a message, and why ? 4(CO3,4)

5. (a) What is PGP ? Explain the format of private key ring table and public key ring table in PGP. List the inputs needed to extract information at the sender side in PGP. 5(CO3,4)

OR

(b) Classify Transport Mode and Tunnel Mode in IPSec. List and describe the fields of Authentication Header Protocol. 5(CO4)

(c) Define a session key and show how a KDC can create a session key between Alice and Bob. 5(CO4)

6. (a) Give the security threats associated with the web security. Analyze the solutions provided by different protocols provided in TCP/IP protocol stack along with their locations at different layer. 6(CO5)

OR

(b) Differentiate between a session and a connection in SSL protocol. List the four protocols in SSL with their purpose. 6(CO5)

(c) How firewall is used to control the access and enforce the security policy inside a network. Justify your answer at least with one example. 4(CO5)

Course Code : CST 408-1 HJRT/MS – 17 / 2298

Eighth Semester B. E. (Computer Science and Engineering) Examination

Elective - III

NATURAL LANGUAGE PROCESSING

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

Assume suitable data and illustrate answers with neat sketches wherever necessary.

1. (a) Consider the sentence. "I made her duck". Explain ambiguity at various levels of NLP. 4(CO1)

   (b) Justify "Natural Language Processing is complex" with suitable example. 4(CO1)

   (c) Differentiate between : Training corpus and Testing corpus. 2(CO1)

2. (a) **Design** Language model for the following set of uni-gram and confusion matrix.

   (i) Probability matrix.

   (ii) Add–1 smoothing and probability matrix.

   (iii) Witten Bell Smoothing sample cases.

	Please	Visit	Your	Registered	Mail	Box
Please	10	1237	59	0	815	0
Visit	122	0	4342	0	60	11
Your	90	343	0	3478	436	0
Registered	232	212	878	0	982	125
Mail	340	0	343	348	0	8765
Box	1232	0	987	565	0	0

Unigram values :

Please	Visit	Your	Registered	Mail	Box
1033	4155	899	1799	6523	4342

In which of unigram, maximum smoothing is required. 4+3+3=10(CO2)

3. (a) Implement CKY algorithm on following sentence. "The filght includes a meal".

S → NP VP[0.8]

NP → DTN[0.3]

VP → V NP[0.20]

V → Includes [0.05]

DT → the [0.40]

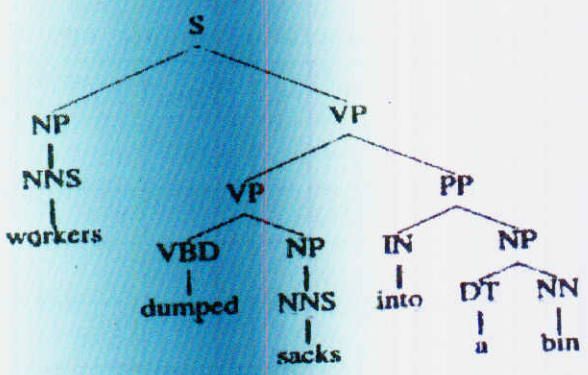
DT → a [0.40]

N → meal [0.01]

N → flight [0.02]

Find the best parsing sequence and value of parsing sequence. Also design the parsing rules. 7(CO3)

(b) Define Lexicalized PCFG. For the following tree, identify the "head words" using suitable rules.



Write the lexicalized probability rule for

P(S(dumped) → NP (Workers) VP(dumped)). 3(CO3)

4. (a) Identify Suitable Thematic roles for the verb ***serve*** in following sentence:–

(a) Well, there was the time they served, green-lipped muscles from New Zeland.

(b) Which airlines serve Denver ?

(c) Which airlines serve breakfast ?

Identify Selectional restrictions that can be used to augment thematic roles for the above example. 6(CO4)

(b) Compute the perplexity of following sentence, using bi-gram module.

I	Like	Chinese	Food	Lunch
2450	900	210	3434	9232

Confusion matrix :

	Chinese	Food	Like
Chinese	20	2211	1020
Food	412	10	2120
Like	1450	2001	12

Sentence : Chinese food like 4(CO4)

OR

(c) Differentiate between Dictionaries, Wordnet and Thesaurus with respect to their contents. 4(CO4)

5. (a) Analyze the different Evaluation parameters commonly used for Text Summarization systems. 5(CO4)

(b) Discuss how NER can be useful for Automatic Summarization and machine translation applications. 5(CO4)

OR

(c) Discuss the challenges associated with Sentimental Analysis task. 5(CO4)

6. (a) Create alignment matrix for following English sentence with destination language as Hindi: "Jaipur popularly known as pink city". 5(CO5)

(b) What are various steps in designing training alignment model. Explain EM step with suitable example. 5(CO5)

Course Code : CST 408-2 HJRT/MS – 17 / 2299

Eighth Semester B. E. (Computer Science and Engineering) Examination

Elective – III

DISTRIBUTED AND PARALLEL DATABASES

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) All questions carry marks as indicated.

(2) Assume suitable data and illustrate answers with neat sketches wherever necessary.

1. (a) How inter-operator parallelism is different from Intra-Operator parallelism ? Explain with example. Also discuss the various data placement techniques. 5(CO1)

(b) Compare Homogeneous and Heterogeneous database management system. Also discuss the reference architecture of distributed database management system. 5(CO1)

2. Solve any **Two** :—

(a) Consider the following Global relation schema

CUSTOMER(Cus_id,Name,Address,Phone)
PRODUCT(Id,Name,Quantity,Cost)
INVOICE(Invoice_no,Cus_id,Date,Total)

Fragmentation Schema
Product1=SL cost<=5000
Product2=SL cost>5000

Allocation schema
Product1 at site 1,2
Product2 at site 3,4

Write an application which takes product details and customer id from the terminal and updates the product table by the quantity purchases by the user and generates an invoice for the purchase at level 1, 2 and 3 of transparency. 5(CO2)

(b) Consider the following Global Relation :-

Emp(Eno,name,sal,dept)
Project(Pno,Pname,Budget,Location)
Asg(Eno,Pno,task,role)

Consider the following applications:

(1) Find the name and budget of projects given their location issued at three sites for ex. Mumbai, Nagpur, Chennai.

(2) Access project information according to budget

One site accesses <200000 and other accesses ≥200000.

Infer the simple and minterm predicates for horizontal fragmentation on Project. Is it possible to design derived horizontal fragmentation on Asg on the attribute which is not in Asg ? Also discuss the concept of complete and minimal set of predicates. 5(CO2)

(c) Derive the measure of cost and benefit in redundant and non redundant cases in horizontal fragmentation and vertical fragmentation. 5(CO2)

3. (a) Discuss the ACID properties of a transaction in detail. 5(CO2)

(b) Compare the concurrency based locking in centralized databases and distributed databases. 5(CO2)

4. (a) For estimating profiles of results of algebraic operations Selection and Semijoin, comment on the following :

(i) Cardinality

(ii) Size

(iii) Distinct values. 5(CO3)

(b) Consider the following fragmentation schema :

Global Schema : DOCTOR(DNUM,NAME,DEPT)

PATIENT(PNUM,NAME,DEPT, TREAT,DNUM)

CARE(PNUM,DRUG,QUAN)

Fragmentation Schema :

$$DOCTOR_1 = SL_{DEPT="SURGERY"} DOCTOR$$

$$DOCTOR_2 = SL_{DEPT="PEDIATRICS"} DOCTOR$$

$$DOCTOR_3 = SL_{DEPT\ !="SURGERY"\ AND\ DEPT=!"PEDIATRICS"} DOCTOR$$

$$PATIENT_1 = SL_{DEPT="SURGERY"\ AND\ TREAT\ =\ "INCENTIVE"} PATIENT$$

$$PATIENT_2 = SL_{DEPT="SURGERY"\ AND\ TREAT\ =\ !"INCENTIVE"} PATIENT$$

$$PATIENT_3 = SL_{DEPT!="SURGERY"}$$

$$CARE_1 = CARE\ SJ_{DNUM\ =\ DNUM}\ PATIENT_1$$
$$CARE_2 = CARE\ SJ_{DNUM\ =\ DNUM}\ PATIENT_2$$
$$CARE_3 = CARE\ SJ_{DNUM\ =\ DNUM}\ PATIENT_3$$

Assume that a patient is always assigned to the same department as his or her doctor.

Translate the following global queries into fragment queries and simplify them using criteria 1 to 5.

(1) $PJ_{NAME}\ SL_{DRUG\ =\ "ASPIRIN"\ AND\ TREAT\ =\ "INCENTIVE"}\ (DOCTOR\ JN_{DNUM\ =\ DNUM}\ PATIENT\ NJN\ CARE)$

(2) $GB\ DEPTNUM,\ AVG\ (SAL)\ SL_{DRUG\ =\ "ASPIRIN"}\ DOCTOR\ JN_{DNUM\ =\ DNUM}\ PATIENT\ NJN\ CARE)$ 5(CO3)

5. (a) Consider a data item x. Let RTM(x) = 25 and WTM(x) = 20. Let the pair {Ri(x), TS} ({Wi(x), TS}) denote a read (write) request of transaction Ti on the item x with timestamp TS. Determine the behavior of the basic timestamp method with the following sequence of requests :-

{R1(x),19}, {R2(x),22}, {W3(x),21}

{W4(x),23}, {R5(x),28}, {W6(x),27} 5(CO2)

(b) In case of a fully redundant database, how does a strict replica control protocol works ? 5(CO2)

6. (a) Explain importance of parallel and distributed processing in the context of data mining techniques. Explain its benefits for any of the technique. 5(CO4)

(b) Write short notes on following :

(i) Terradata

(ii) Gamma 5(CO4)

Course Code : CST 409-1 **HJRT/MS – 17 / 2300**

Eighth Semester B. E. (Computer Science and Engineering) Examination

Elective - IV

WEB INTELLIGENCE AND BIG DATA

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) All questions are compulsory and carry equal marks.

(2) Figures to the right indicate marks. Each question is mapped to a course object given against it.

(3) Carefully see internal choices.

1. (a) What is Moore's Law ? What were the predictions and clarify if is it relevant today ? 4(CO1)

OR

(b) List some of the points why Google is the most used search engine in the internet today ? 4(CO1)

(c) There are million Finger Prints (FPs) available in the database of the investigating agency. Suppose the probability of finding minutia in random grid square of a finger print (FP) if fifteen percent. If a grid has minutia in a square of a grid, then the corresponding grid of other FP will also have the minutia with a probability of eighty percent if the FP is taken from the same finger. Consider each function f in a family of **F** is defined by a 3 grid squares. f says 'yes' if both FPs have minutia in all three grid squares otherwise it says 'no'. If we choose 1000 such functions randomly chosen from f, find how many FPs will wrongly be investigated even if there is no similarity between them for each FP if LHS is used. Can you use some method to improve the situation ? 6(CO1)

2. (a) Explain with an example the significance of taking TF and IDF for the words within documents. 4(CO1)

(b) Define mutual information. Calculate in the given table the mutual information for the features *like* and *course* towards behavior *sentiment*. Also justify that mutual information can be a measure for selection of a feature.

Count		Sentiment
2000	I really like this course and am learning a lot	Positive
800	I really hate this course and think it is waste of time.	negative
200	The course is really too simple and quite a bore.	negative
3000	The course is simple, fun and very easy to follow.	positive
1000	I'm enjoying this course a lot and learning something too.	positive
400	I would enjoy myself a lot if I did not have to be in this course.	negative.
600	I did not enjoy this course enough.	negative.

6(CO3)

3. (a) Why we require using Naïve Bayes Classifier ? Considering the table given in Question 2 b, determine the sentiment of the statement *"I like the course and don't think it is a waste of money."* using Naïve Bayes classification. Consider suitable features to be included. 6(CO1)

(b) Explain the concept of Sparse-Distributed Memory. How is it used in actual applications ? 4(CO1)

4. (a) Considering an example of word count explain the approach for how map reduce can be used to calculate a TF.IDF scores of different keywords/ features. 5(CO2)

(b) Why map-reduce paradigm is scalable and efficient ? Justify mathematically. 5(CO2)

OR

(c) What is a distributed file system ? How GFS/HDFS implements its read/ rights on replicas without creating bottlenecks at the master ? 5(CO2)

5. (a) What is association rule mining ? Why is ARM important in data mining applications ? Give any one algorithm of ARM. 5(CO3)

(b) Explain how semantic web vision will be used to infer 'Who is the leader of USA ?' From the facts listed in different web pages. 5(CO4)

OR

(c) Why there is a need of parallel database ? What were the contributions of Mongo DB in this domain making it a popular choice ? 5(CO2)

6. (a) Design the Bayes Network for the given data. Consider P(W|S,R) not joint. P(R) and P(S) given in the following probability tables respectively.

(1) Find the probability of it rained yesterday when we find grass to be wet.

(2) Find the probability that the sprinkler was on.

W	S	R	P
y	y	y	0.95
y	y	n	0.75
y	n	y	0.85
y	n	n	0.05
n	n	n	0.9
n	n	y	0.15
n	y	n	0.25
n	y	y	0.1

R	P
y	0.25
n	0.75

S	P
y	0.4
n	0.6

4(CO3)

(b) What is proposition and predicate logic ? Clarify if these systems are capable of handling uncertainty ? 6(CO4)

OR

(c) What is linear regression ? Explain with an example how we can predict a value of an output variable using linear regression. 6(CO4)