Course Code : CST 406 GISU/RS – 17 / 3310

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 answer properly.
(3) Assume suitable data and illustrate answers with neat sketches wherever necessary.

1. (a) Illustrate schema integration and instance integration with the help of an example. 3(CO1)

(b) Describe data warehouse development life cycle. 3(CO1)

(c) Explain the following terms and give suitable examples :

(i) data cleansing (ii) data transformation. 4(CO1)

2. (a) Explain data warehouse architecture. Comment on the importance of metadata repository. 5(CO1)

(b) Suppose a market shopping data warehouse consists of four dimensions: customer,date, product, and store, and two measures: count, and avg_sales, where avg_sales stores the real sales in pounds at the lowest level but the corresponding average sales at other levels.

(a) Construct a star schema for this data warehouse.

(b) Starting with the base cuboid [customer, date, product, store], what specific OLAP operations should be performed in order to list the average sales of each cosmetic product since January 2010 ? Explain your answer.

(c) Comment on de-normalization in multi-dimensional models. 5(CO2)

OR

(c) Suppose that the data warehouse for Big University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg_grade. When at the lowest conceptual level (e.g. for a given student, course, semester, and instructor combination), the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg_grade stores the average grade for the given combination.

(a) Construct a snowflake schema diagram for the data warehouse.

(b) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e. g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big University student.

(c) If each dimension has five levels (including all), such as "student<major<status<university<all", how many cuboids will this cube contain (including the base and apex cuboids) ?

5(CO2)

3. (a) Consider a table DEPT (dept_id,dept_name,mgr,loc_id)
Suppose block 1 of this heap-organized DEPT table contains rows as follows :

50,Shipping, 121,1500
20,Marketing, 201,1800

Block 2 contains rows for the same table as follows :

30,Purchasing, 114,1700
60,IT,103,1400

(i) Predict in what sequence will the blocks be scanned for a heap organized table.

(ii) Predict in what sequence will the blocks be scanned for a heap organized table with B-tree index on dept_id. List the contents of the B-tree index.

(iii) Predict in what sequence will the blocks be scanned for index-organized table with primary key dept_id. List the contents of IOT.

3(CO2)

(b) State what is a Bitmap Join Index ? List advantages of creating a bitmap join index over normal joins. Give the command for creating a bitmap join index. 3(CO2)

(c) Explain Range Partitioning and Hash Partitioning with the help of an example. 4(CO2)

4. (a) Describe the architecture of Online Analytical Mining (OLAM) system. 4(CO3)

(b) Assume an integer-valued attribute "A" whose values are distributed as follows is given :

0,0,0,1,1,1,2,2,5,6,7,17,18,19,20,25,28,29,33,39,43,44,44,46,51,58,59,60,61,65,77,78,81,99,120.

"A" has to be summarized in a histogram with 5 buckets using the following 3 methods :

(1) Equidepth (2) V-Optimal (3) MaxDiff

Give the 3 histograms that would be obtained using the 3 methods. Also explain how your histograms were derived. Compare the 3 histograms, which histogram(s) do you prefer ? Justify your answer. 6(CO3)

OR

(c) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13,15,16,16,19,20,20,21,22,22,25, 25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70.

(i) Show a boxplot of the data.

(ii) Use min-max normalization to transform the value 35 for age onto the range [0.0,1.0].

(iii) Plot an equal-width histogram of width 10

(iv) Sketch examples of each of the following sampling techniques : SRSWOR,SRSWR, cluster sampling stratified sampling. Use samples of size 5 and the strata "youth", "middle-aged", and "senior". 6(CO3)

Solve Q. 5 or Q. 6 :

5. (a) Apply the Apriori algorithm on the grocery store example with support threshold s = 33.34% and confidence threshold c = 60%. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association, rules that are generated and highlight the strong ones, sort them by confidence.

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

7(CO4)

(b) "Strong rules are not necessarily interesting". Clarify this statement. 3(CO3)

6. (a) General Motors (GM) is planning their production strategy for their next model. Three alternatives are being considered for their model Malibu: 30,000,20,000, and 12,000.GM decides to categorize the demand for Malibu for the next year as either High (H) or Low (L). The payoffs measured in millions of dollars and probabilities of states of nature are presented in the table below

	States of nature	
Decision Alternatives	High (H)	Low (L)
Produce 30K	29	–12
Produce 20K	18	8
Produce 12K	3	11
Probabilities	0.62	0.38

For this problem, if we want to do a decision tree,

(i) How many decision nodes are required ?

(ii) How many branches come out of each decision node ?

(iii) How many chance nodes are required ?

(iv) How many branches come out of each chance node ?

(v) Construct the decision tree. Label each branch completely including probabilities and payoffs.

(vi) Solve the decision tree and find the best production strategy. 10(CO4)

Solve Q. 7 or Q. 8 :

7. (a) Apply the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters :

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

Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 epoch only. At the end of this epoch show:

(a) The new clusters (i. e. the examples belonging to each cluster)

(b) The centers of the new clusters

(c) Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids. 10(CO4)

8. (a) Use single, complete, and average link agglomerative clustering to group the data described by the following distance metric. Produce the dendrograms.

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

10(CO4)

Course Code : CST 407

GISU/RS – 17 / 3311

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

INFORMATION SECURITY

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) All questions carry equal marks.
(2) Solve any **Two** from Q. **One**, Q. **Two**, Q. **Three**, Q. **Four** and Q. **Five**.
(3) Solve Q. **Six** is Compulsory.
(4) Assume suitable data wherever necessary.

1. Solve any **Two** :—

(a) Classify the types of crypt-analytic attacks based on what is known to the attacker. Also Differentiate active and passive attack. 5 (CO 1)

(b) Compare and contrast the operation of Worms and Virus Activity. Give suitable example of each. 5 (CO 1, 5)

(c) Generalize why network need security, state important parameter associated with network security Model. 5 (CO 1)

2. Solve any **Two** :—

(a) Explain about single Round of DES algorithm. Describe the key discarding process of DES.

Eve secretly gets access to one person's computer and using her cipher types "abcdefghij". The screen shows "CABDEHFGIJ" If eve knows that person is using a keyed transposition cipher.

(i) State what type of attack is eve launching.

(ii) What is the size of permutation key ? 5 (CO 2)

(b) The problem explores the use of one time pad version of the substitution cipher. In this scheme of random number 0-26 e.g. if key is 3, 9, 5, ... then 1st letter is encrypted with shift of 3 and 2nd by 9 and so on. In this manner design the ciphertext if the Plaintext. "SENDMOREMONEY" with the key stream is 9, 0, 1, 7, 23, 15, 21, 14, 11, 2, 8, 9, 4 ? 5 (CO 2)

(c) Design plaintext for the given ciphertext "oprlssewaercaifsaroeodik" using double columnar Transposition technique and guess the plaintext if key 1 "BANK" and key 2. "lock". (hint : plaintext ends with 5 digit number) 5 (CO 2)

3. Solve any **Two** :—

(a) Implement and demonstrate the idea of extended Euclidean algorithm, to compute inverse of a number in RSA. 5 (CO 1, 2, 4)

(b) Fill in the blanks.

The relation between the RSA encryption and decryption key is _____. Let the RSA modulus, n be 77. if the encryption key is 7. The decryption key is _____. now suppose if the ciphertext = 5 the correspinding plaintext is _____. analyze these information and tell what is the time complexities of RSA encryption _____ and decryption (as a function of key size, k)_____. 5 (CO 1, 2, 4)

(c) Show the use of discrete logarithm in key managements. In a Diffie-Hellman scheme let a common prime be $q = 11$ primitive root $\alpha = 2$. Show that 2 is a primitive root of 11. If a user A has public key $Y_A = 9$, what is A's private key X_A ? If user B has public key $Y_b = 3$, what is the shared secret key K ? 5 (CO 1, 2, 4)

4. Solve any **Two** :—

(a) Would the hash function defined below be appropriate fro security application ?

$h(x) = (x \bmod n) \bmod p$

Where n = 1024 bit prime and p is a 160 bit prime. Similarly.

Derive an expression for the number of messages that need to be created (variation of the malicious message and the innocuous message). So that Ravi's Birthday attack is successful with a probability 0.75. 5 (CO 4)

(b) Suppose that a bank wants to communicate with all its customers using digital certificates. What infrastructure would the bank need and the customers need ? Draw the Diagram showing the flow of event. 5 (CO 4)

(c) If we can create a certificate of our own, so can the attacker. Where is the security, then ? Discuss the important ingredients used in DSS signing and verifying with the help of neat sketch. 5 (CO 4)

(d) The authentication Protocol are used for both one-way and Mutual Authentication. In the real world can you think of application where –

(i) Client–to–server authentication is mandatory but server–to–client authentication is not.

(ii) Server–to–client authentication is mandatory but Client–to–server authentication is not.

(iii) Both Server–to–client and Client–to–server authentication are mandatory.

Illustrate your answer in brief. 5 (CO 4)

5. Solve any **Two** :—

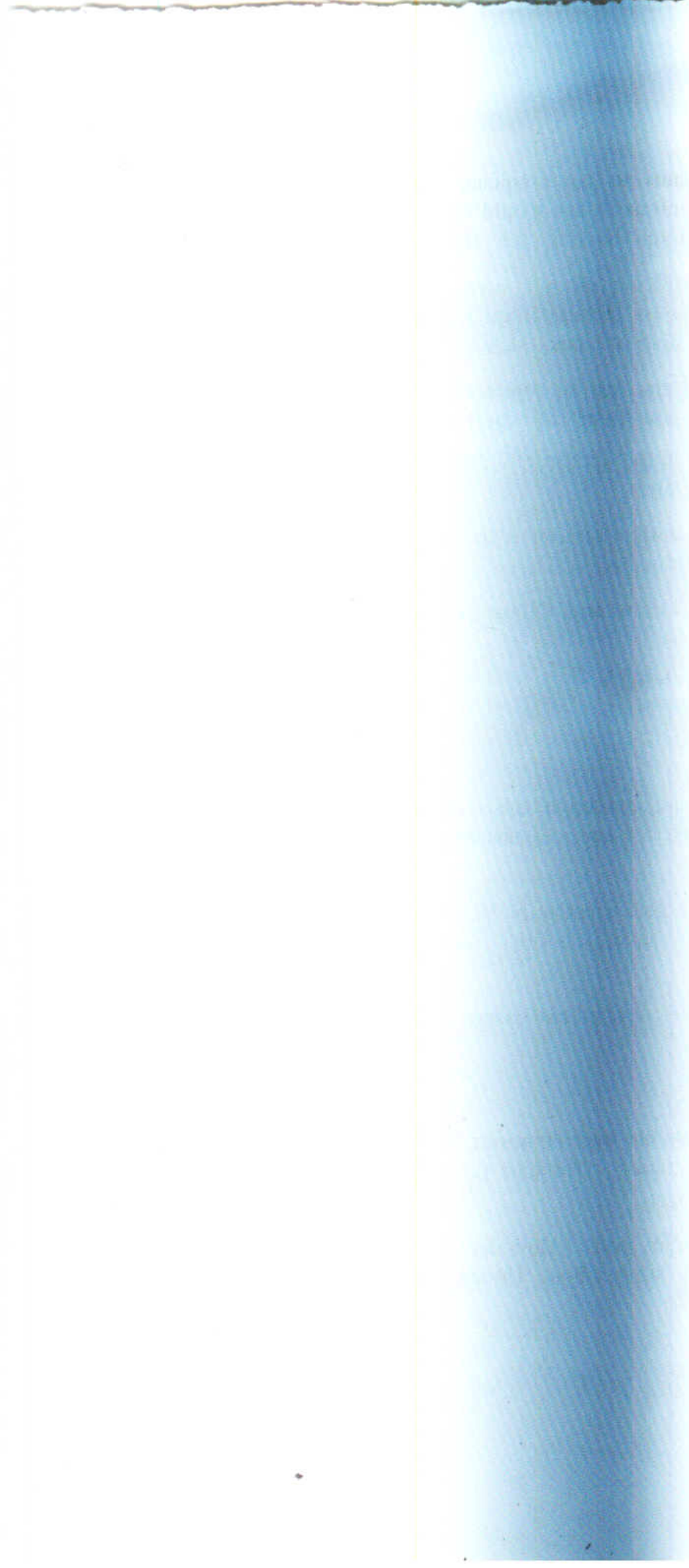
(a) How does inter realms authentication takes place in Kerberos ? Give a neat sketch of how request for services are made for one realm to another remote realm. 5 (CO 5)

(b) How does certificate trust gets created in PGP implementation ? Draw the flow of events related to several entities used in PGP certificate. 5 (CO 5)

(c) Compare and contrast the operational issues of SET and SSL. 5 (CO 5)

6. (a) List and describe three possible configurations of firewalls to protect against gaining access to private network. Also state the main limitation of a firewall. 5 (CO 5)

(b) How does audit record structure looks ? How does intrusion detection mechanism are classified ? Write in short about Honey Pots. 5 (CO 5)



Course Code : CST 408–1 GISU/RS – 17 / 3312

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

Elective - III

NATURAL LANGUAGE PROCESSING

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

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

1. (a) Compute maximum likelihood estimation for following sentences :

<S> I like Indian Food </S>

<S> I like India </S>

<S> Indian is my country </S>

Compute any three bi–gram probabilities. 4(CO1)

(b) The distribution of Eight horses participating in race is as given below:

H1 = 1/2, H2 = 1/4, H3 = 1/8, H4 = 1/16, H5 – H8 = 1/64

Compute the entropy of random variable X and find out total number of bits required for representation. 4(CO1)

(c) Define cross entropy with suitable example. 2(CO2)

2. (a) Define Language Model

Assuming size of corpus is 7322 sentences and V = 1535.

1	Want	To	Eat	Chinese	Food	Lunch
2322	455	4565	1244	565	665	1212

Contd.

Confusion Matrix :

	I	Wnat	To	Eat	Chinese	Food	Lunch
I	22	477	34	0	0	0	1123
Want	4	0	983	44	322	32	0
To	3	0	4	1455	0	44	0
Eat	55	0	0	4	199	8	0
Chinese	0	2	0	34	0	593	0
Food	0	0	34	0	324	0	578
Lunch	12	1	22	1	12	0	1

Find the probabiltiy matrix for MLE Estimation. [Use Bigram model]

Perform Smoothing using Add–2 method.

Comment on new values generated after smoothing.

Give sample example for witten–Bell Smoothing method.

Demonstrate the method to compute Perplexity and Entropy.

Coment upon improving performance of Language model. 10(CO2)

3. (a) Design Viterbi algorithm for following set of given probabilities :

Prob(the | ART)=0.54, Prob(a | ART) = 0.36, Prob(files | N) = 0.25,

Prob(a | N)=0.04 Prob(flies | V) = 0.076, Prob(Flower | N) = 0.07

Prob(like | V)=0.1, Prob(Flower | V) = 0.05, Prob(like | P) = 0.68,

Prob(like | N)=0.12

Sentence : Files like flowers

Assume suitable data if required 5(CO3)

(b) Implement CKY algorithm on following sentence "The flight includes a meal".

S→NPVP[0.7]

NP→DTN[0.4]

VP→VNP [0.25]

V→Inclues [0.15]

DT→the [0.45]

DT→a [0.30]

N→meal [0.16]

N→flight [0.2]

Find the best parsing sequence and value of parsing sequence. Also design the parsing rules. 5(CO3)

4. (a) Define the following terms.

- **Feature vector / Collocation vector**
- **Co–occurence vector.**

Generate the Feature and Co–occurrence vector for the following sentence considering two context words to left and right of target word "bass"

Consider context words

Fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, ban.

An electric guitar and bass player stand off to one side. 6(CO4)

(b) What is contrastive knowledge, apply and design transfer model for suitable example. 4(CO4)

OR

(c) Discuss how Naive Baye's classifier can be used to solve word sense disambiguation problem. Also discuss its associated drawbacks. 4(CO4)

5. (a) Consider the following hypothetical information retrieval scenario. Suppose it has been found at Edinburgh Royal Infirmary that due to equipement malfunction, the results of blood tests taken on 2013–12–04 are unreliable for diabetic patients. The hospital would like to contact all diabetic patients who had any kind of blood test on that day, to repeat the test. The hospital uses an information retrieval system to identify these patients. Suppose the collection of patients' medical records contains 10000 documents, 150 of which are relevant to the above query. The system returns 250 documents, 125 of which are relevant to the query.

(a) Calculate the precision and recall for this system, showing the details of your calculations.

(b) Based on your results from (a), explain what the two measures mean for this scenario. How well would you say that the hospital's information IR system works ?

(c) According to the precision–recall tradeoff, what will likely happen if an IR system is tuned to aim for 100% recall ?

6(CO4)

(b) Discuss the problems associated with identifying Named Entities from text. 4(CO4)

OR

(c) Identify the stages involved in automatic Text Summarization. 4(CO4)

6. (a) Explain basic steps of Phrase based machine translation. Convert the English sentence into Hindi using P(FE) model.

Sentence: Ram and Shyam are playing in garden during afternoon. 5(CO5)

OR

Write an algorithm for "direct translation", apply algorithm and convert sentence into Hindi.

Sentence:Students study hard during exams for good results.

Write the output of each step : Morphology, Lexical transfer, local reordering. 5(CO5)

(b) What is EM Training for alignment model. Explain any two examples for sentences of Q.6(a) to perform EM training. 5(CO5)

75

Course Code : CST 408–2

GISU/RS – 17 / 3313

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) Each questions carry marks as indicated.
(2) Due credit will be given to neatness.
(3) Assume suitable data whenever necessary.
(4) Illustrate your answers wherever necessary with the help of neat sketches.

1. (a) Identify the form of parallelism (inter query, inter operation, or intra operation) which is most important for each of the following task. Give reason.

(i) Increasing the throughput of a system with many small queries.

(ii) Increasing the throughput of a system with a few large queries, when the number of disks and processors is large. 2(CO1)

(b) Attempt any **Two** questions :—

Compare various data partitioning techniques of parallel databases. Apply these methods to the relation R(x, y) given below and partition it onto the 3 disks d1, d2 and d3.

For hash partitioning technique use hash function h(x) = (x mod 3) + 1

For Range partitioning technique use range vector on x attribute V[5, 10]

R	x	y
T1	1	1
T2	2	4
T3	15	6

Contd..

R	x	y
T4	6	6
T5	7	2
T6	9	3
T7	12	4
T8	5	1
T9	8	3

4(CO1)

(c) Discuss the following parallel database architecture with diagram. List and explain the strength and weakness of it.

(i) Shared Memory architecture.

(ii) NUMA architecture. 4(CO1)

(d) Discriminate Homogeneous and Heterogeneous database management system. Also discuss the components of distributed database management system. 4(CO1)

2. (a) Consider the Global Schema - > STUDENT (SNUM, NAME, DEPT).

Consider the cases in which an application which is repeated for many possible values of student number and display its details. Write the application for :

(i) Accessing the database for each student number given at the terminal.

(ii) Accessing the database after having collected several inputs from the terminal.

(iii) Accessing the database before collecting inputs from the terminal.

Also analyze the effect of these alternatives on the efficiency of application execution. 6(CO1)

(b) Consider the Global Schema, Fragmentation Schema and Allocation Schema:

Global Schema - >

SUPPLIER (SNUM, NAME, LOCATION)

SUPPLY (SNUM, PNUM, DEPT)

Fragmentation Schema - >

SUPPLIER1 = SL $_{\text{LOCATION = "Mumbai"}}$ SUPPLIER

SUPPLIER2 = SL $_{\text{LOCATION = "Pune"}}$ SUPPLIER

SUPPLY1 = SUPPLY SJ $_{\text{SNUM = SNUM}}$ SUPPLIER1

SUPPLIY2 = SUPPLY SJ $_{\text{SNUM = SNUM}}$ SUPPLIER2

Allocation Schema - >

SUPPLIER1 = Site 1 & 2, SUPPLEIR2 = site 3 & 4

SUPPLY1 = site 5, SUPPLY2 = site 6

Write an application that requires the part number from the terminal & retrieves the supplier name who supplies the given part at level 1, 2 and 3 of transparency. 4(CO1)

OR

(c) Explain the types of allocation of fragments to sites. Illustrate the equation for the measure of cost and benefit in case of horizontal fragmentation. 4(CO1)

3. Attempt any **Two** questions :—

(a) Discuss the role of time and timestamp in a distributed database to implement concurrency control technique.

Consider a data item x. Let RTM(x) = 25 and WTM(x) = 20. Let the pair (Ri(x), TS) (Wi(x), TS) denote the read and write request of transaction Ti on the item x with timestamp TS. Analyze 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) Using proper scenario explain distributed wait for graph showing a distributed deadlock. Also explain distributed deadlock detection algorithm. 5(CO2)

(c) Draw and explain the reference model for concurrency control. 5(CO2)

4. (a) Consider an Engineering database that maintains three tables :

EMP (ENO, NAME, TIRLE, SAL)

PROJ (PNO, PNAME, BUDGET, LOCATION)

ASG (PNO, ENO, TASK, DURATION)

Translate the following global queries into fragment queries and apply Criteria 1 to 6 to simplify them. Use the algebra of qualified relation for the elimination of fragments which are not required by the query.

(i) Assume that relation PROJ is horizontally fragmented in PROJ1 = $SL_{PNO < = "P2"}$ PROJ and PROJ2 = $SL_{PNO > "P2"}$ PEOJ

Assume that ASG is horizontally fragmented as

ASG1 = $SL_{PNO < = "P2"}$ ASG, ASG2 = $SL_{"P2" < PNO < = "P4"}$ ASG & ASG3 = $SL_{PNO > "P4"}$ ASG

Reduce the following query to the fragments :

$PJ_{RESP, BUDGET}$ (ASG JN $_{PNO = PNO}$ $SL_{PNAME = "CAD / CAM"}$ PROJ)

(ii) Assume that ASG is not indirectly fragmented as

ASG1 = ASG $SJ_{PNO = PNO}$ PROJ1, ASG2 = ASG $SJ_{PNO = PNO}$ PROJ2

And EMP is vertically fragmented as

EMP1 = $PJ_{ENO, NAME}$ EMP EMP2 = $PJ_{ENO, TITLE}$ EMP

Reduce the following query to the fragments :

PJ_{NAME} (EMP JN $_{ENO = ENO}$ (ASG JN $_{PNO = PNO}$ $SL_{PNAME = "CAD / CAM"}$ PROJ)

6(CO3)

(b) Explain the motivation behind the semi-join reduction in distributed database. Justify the use of semi-join programs for the join queries. 4(CO3)

OR

(c) Explain the database profile. Also Evaluate the estimation profiles of result of union and difference operation in terms of size, cardinality and distinct value. 4(CO3)

5. Attempt any **Two** questions :—

(a) Consider various failures that occur during 2PC for a transaction. For each possible failure, explain how 2PC ensures transaction atomicity despite the failure. 5(CO2)

(b) How does quorum based protocol behaves in case of network partitioning? 5(CO2)

(c) Discuss the role of replication management in context of distributed database. 5(CO2)

6. Attempt any **Two** questions :—

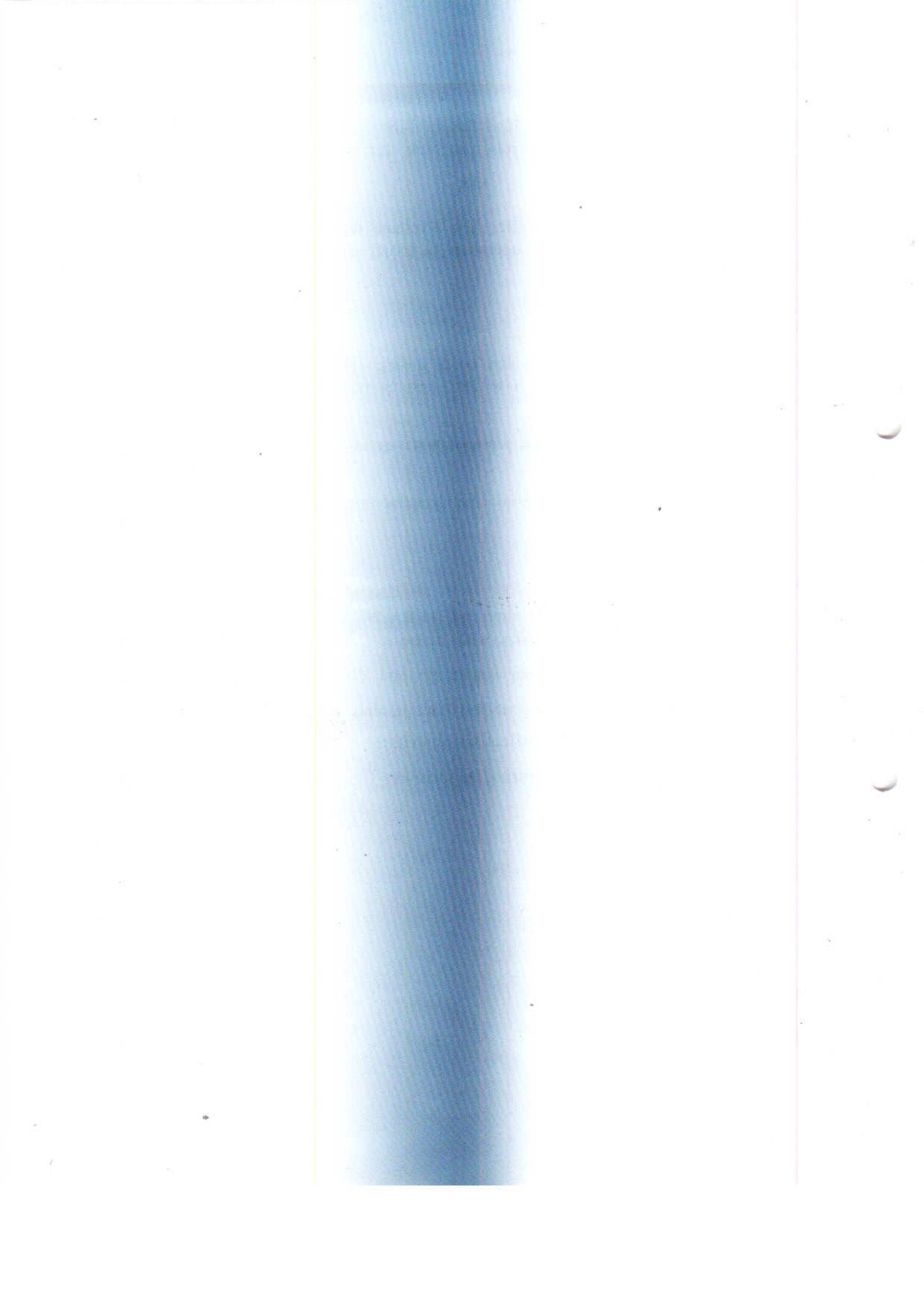
(a) List and explain the features of TERADATA database. Also discuss the components of TERADATA with diagram. 5(CO4)

(b) Discriminate between data warehouse and distributed database. 5(CO4)

(c) Discuss the following distributed data mining algorithm in detail.

(i) Distributed Classification algorithm.

(ii) Distributed Clustering Algorithm. 5(CO4)



Course Code : CST 409 – 1 GISU/RS – 17 / 3314

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 carry marks as indicated against them.

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

1. (a) Fill in the blanks with most appropriate words :—

(i) Web Intelligence exploits the fundamental and practical impact that ——— and ——— will have on the Web.

(ii) The 7 Vs (characteristics) of Big Data are velocity, ———, ———, ———, ———, ———, ——— .

(iii) Bit data management cycle : Look → ——— → ——— → ——— → ——— Correct. 3 (CO 1)

(b) Consider that the minhash signatures for 5 documents are given per column. Apply the banding theory to divide the signatures into 4 bands with each band having 3 rows. Use the given buckets to store the documents and show which documents are similar and which documents are dissimilar.

Bucket – 1	Bucket – 2	Bucket – 3	Bucket – 4	Bucket – 5			
D1	D2	D3	D4	D5	D6	D7	D8
0	0	0	1	1	2	0	0
2	0	1	3	3	2	0	2
1	0	3	0	0	1	1	1
2	3	0	2	1	0	1	0
2	2	1	3	2	0	1	0
1	1	0	1	2	1	1	1
2	3	0	2	2	1	1	2
2	2	0	1	1	3	1	2
3	2	0	2	2	3	1	3
0	1	1	3	1	1	1	1
0	0	1	3	0	3	1	1
1	0	1	3	1	1	0	3

3 (CO 1)

(c) Illustrate the use of Sparse Distributed Memory (SDM). Consider the following data
Reference Address as 0101101110
and 5 Address of the memory : 0001001010, 0001101100, 000000100, 0111101110, 0000011100
How would you design and organize the SDM for the input data 1001101110. Show the complete sparse memory with the output from memory. (Radius = 3). 4 (CO 1)

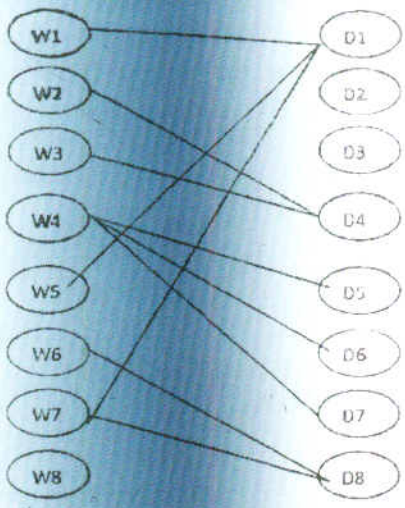
OR

(d) What is the significance of text indexing on the web ? Discuss the algorithm for index creation along with the complexity of index creation, searching word in the index and query search. Can you apply the concpet of page rank along with indexes ? Explain with example. 4 (CO 1)

2. (a) Consider the set of words on L. H. S. of the bipartite graph and set of documents on R. H. S. Discover :—

(i) documents that can be grouped together based on similarity

(ii) keywords that can be grouped together based on similarity.



2 (CO 1)

OR

(b) Justify how TF – IDF score can be good measure for characterization of documents. 2 (CO 1)

(c) What is the significance of Naïve Bayes. classifier in Web Intelligence ? How would you apply what you learned about Naïve Bayes for new document classification ? For the new document, find whether it belongs to the class India or not :—

	Doc id	Words in doc	In c = India ?
Training set	1	Chandrapur Ballarpur Chandrapur	yes
	2	Chandrapur Chandrapur Shimla	yes
	3	Chandrapur Manipur	yes
	4	Pune Nagpur Chandrapur	no
New Document	5	Chandrapur Chandrapur Chandrapur Pune Nagpur	?

5 (CO 4)

(d) Can the concept of Ad – sense be applied on the websites ?
Also solve to find the TF – IDF values considering the given information : Given a single document with terms A, B, C with the following frequencies : A : 3 , B : 3 , C : 1.
The documents belong to a collection of 10,000 documents. The document frequencies (Nw) are : A – 50 , B – 1300 and C – 250. 3 (CO 1)

3. (a) Devise a Map – Reduce based solution for the following query and set of tables, showning the dry – run for the given data.
Select sum (sale), City from sales, cities where sales. AddrId = cities. AddrId GROUP BY City

SALES	
AddrId	Sale
1	1000
2	20000
3	300
4	4000
5	3000
6	8000

CITIES	
AddrId	City
1	Nagpur
2	Delhi
3	Nagpur
4	Nagpur
5	Nagpur
6	Delhi

5 (CO 2)

OR

(b) Devise a Map – Reduce based solution for index creation using 3 Mappers and 2 reducers. Consider the following set of documents for index creation :

D1 – w1 , w1 , w2 , w4
D2 – w1 , w2 , w3 , w4
D3 – w2 , w3 , w4
D4 – w1 , w2 , w3
D5 – w1 , w3 , w4
D6 – w1 , w2 , w2 , w4
D7 – w4 , w2 , w1
D8 – w2 , w3 , w2
D9 – w1 , w3 , w2
D10 – w2 , w1 , w4 , w3

5(CO 2)

(c) MongoDB uses eventual consistency to write the data into the replicas. Demonstrate how eventual consistency can be achieved for three physical nodes X, Y and Z. 2 (CO 2)

(d) "Column oriented databases are gaining popularity over row oriented databases" — Justify. 3 (CO 2)

OR

(e) "Hadoop uses Hadoop Distributed File system (HDFS) rather than the conventional file systems found on single computers". — Justify, stating the use of namenode and datanode. 3 (CO 2)

4. (a) How do classes emerge in unsupervised learning ? Can you apply clustering on graph of web pages to determine the websites of similar domain by deleting the inconsistent edges ? Give example to justify your answer. 5 (CO 3)

(b) How can association rule mining be applied using Aprori algorithm. State all the steps and apply the steps to the given below set of transactions.

Transaction ID	Items
100	A , C , D
200	B , C , E
300	A , B , C , E
400	B , E

Find the rules along with support and confidence. 5 (CO 3)

OR

5. (a) Discuss the main idea of association rule mining ? State its need in a scenario where the output classes are missing. What challenges can you infer from association rule mining ? 4 (CO 3)

(b) Describe the long tail pheomena. Which techniques can deal with this problem in the online recommender system ? Elaborate the techniques.
Also suggest and apply the type of technique that can be used in the given scenario :
"We have set of transaction for each user giving the rating to a particular item and its rating value. There are some items to which the users do not give any rating".
How can you find the rating for such missing items ? 6 (CO 3)

6. (a) Solve using Naïve Bayes classifier to find the probabilities by considering class C = Clear the exam which can have value yes or no. S = Studied for the exam and P = went to party. Prior probabilities for C being yes is 0.3 and C being no is 0.7.

Table 1		
S	C	Probability (P)
Y	Y	0.9
N	Y	0.1
Y	N	0.2
N	N	0.8

Table 2		
P	C	Probability (P)
Y	Y	0.8
N	Y	0.2
Y	N	0.1
N	N	0.9

(1) Given the evidence S = yes, find the probabilities of clearing the exam. (Consider table 1 only)

(2) Consider another feature called party (P). But there is no evidence about party. Find the probabilities of clearing the exam. Will there be any change in the answer ?

(3) Given two evidences Party = yes and Study = yes. Find the probability of clearing the exam.

What can you conclude after each probability calculation ? 6 (CO 4)

(b) How would you combine different networks into Bayesian network ? State example and interpret the need of Bayesian networks. 2 (CO 4)

(c) Predicate logic statements are given. Solve to find if RPI (Joe) is true :—

~ Smart (x) v ~ LikesHockey (x) v RPI (x)

~ Canadian (y) v LikesHockey (y)

~ Skates (z) v LikesHockey (z)

Smart (Joe)

Skates (Joe)

2 (CO 4)

7. (a) Discuss the main idea of Blackboard architecture.
Design an efficient the Blackboard architecture for any **One** of the following scenario :

(1) "A mobile robot for corridor navigation"

(2) Blackboard system for web application

Explain in detail the working of all components of the blackboard.
6 (CO 4)

(b) How would you compare linear and non – linear predication models ?
4 (CO 4)