Course Code : CST 406 DJPV/MS – 16 / 4486

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

DATA WAREHOUSING AND MINING

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) Solve Q. 6 or Q. 7.
(2) Assume suitable data wherever necessary.
(3) Illustrate your answers wherever necessary with the help of examples.
(4) Mobile phones are prohibited in examination hall.

1. (a) What is data warehousing ? Enlist and explain key features of data warehouse.

OR

Compare and contrast OLTP and OLAP system. 4

(b) Explain the following data warehouse model :

(i) Enterprise warehouse.

(ii) Data Mart.

(iii) Virtual warehouse. 6

2. (a) Explain the process of data transformation. 4

(b) 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 docor charges a patient for a visit :

(i) Draw a star schema diagram for the above data warehouse. 3

(ii) Starting with the base cuboid [day, doctor, patient, give a list of OLAP operations to be performed in order to list the total fee collected by each doctor in 2004 ? 2

(iii) To obtain the ... list, write an SQL query assuming the data are ... a relational database with the schema fee (day, mo... r, doctor, hospital, patient, count, charge). 1

3. (a) Explain list partitioni... range partitioning with the help of examples. 5

(b) Write detailed notes ...

(i) Domain Index.

(ii) B–tree Index. 5

OR

(c) Explain how query ... tion can be performed in data warehouse system. 5

4. (a) Describe the steps in ... in data mining when viewed as a process of knowledge discov... 4

(b) Suppose that a goverm... ted age and population data for 18 randomly selected region with ... llowing results :

age	23	...	27	39	41	47	49	50
% Population	9.5	...	17.8	31.4	25.9	27.4	27.2	31.2

3

age	52	...	56	57	58	58	60	61
% Population	34.6	...	33.4	30.2	34.1	32.9	41.2	35.7

(i) Draw boxplot ... ta.

(ii) Draw Scatter pl... q–q plot based on these two variables. 3

5. (a) The following table shows the midterm and final exam grades obtained for students in a database course.

Midterm exam (x)	Final exam (y)
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

(i) Plot the data. Do x and y seem to have a linear relationship? 4

(ii) Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course. 4

(iii) Predict the final exam grade of a student who received an 86 on the midterm exam. 2

OR

(b) A database has five transact ions Let min_sup=60% and min_conf=80%

TID	items bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}

Contd.

Contd.

TID	Items bought
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, K, I, E}

(i) Find all frequent item sets using FP–growth. 7

(ii) List all of the strong association rules (with support s and confidence c) matching the following metarule, where X is a veriable representing customer, and item, denotes variables representing items (e.g. "A", "B", etc) :

$\forall_x \in$ transaction, buys(X, $item_1$) $\wedge$ buys(X, $item_2$) $\Rightarrow$ buys(X, $items_3$) [s,c] 3

6. (a) Both K–means and K–medoids algorithm can perform effective clustering. Illustrate the strength and weakness of k–means in comparsion with k–medoids. 4

(b) Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are :

A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 0)

The distance function is Euclidean distance. Suppose initially we assign A1, B1 and C1 as the center of cluster, respectively. Use the K–means algorithm to find only :—

(i) The three cluster after the first round of execution.

(ii) The final three clusters. 6

7. Explain the working of PAM (partitioning around medoids) algorithm with an example. 10

Course Code : CST 407 **DJPV/MS – 16 / 4487**

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) Due credit will be given to neatness
(3) Assume suitable data wherever necessary.
(4) Solve any two from all qauestions.
(5) Illustrate your answers wherever necessary with the help of neat sketches.
(6) Mobile phones are prohibited in examination hall.

1. (a) What are goals (aspects) of information Security ? Comment on each one of its benefits, in brief. 5

(b) Differentiate between cryptanalytic attack and non-cryptanalytic attacks. Explain the non-cryptanalytic attacks atfer categorizing them into groups related to the security golas. 5

(c) What are the security services, state the relation between serivices and security mechanism. Classify these services according to layers functionality. 5

2. (a) A mode of operation is a technique for enhancing the effect of a cryptographic algorithm or adapting the algorithm for an application. What are the five modes of operation have been standardized by NIST for use with symmetric block ciphers such as DES and AES ? 5

(b) Compare DES and AES. Which one is bit oriented ? Which one is byte oriented ? Why only one substitution table (S box) is needed in AES but several in DES ? Why are expansion and compression permutations required in DES, but not in AES ? 5

(c) Euclidean algorithm is a simple procedure for determining the greatest common division of two positive integers. Explain Extended Euclidean algorithm which is used in RSA to verify private key. Write a program for same. 5

3. (a) What are the requirements of public key cryptography system ? Explain the characteristics of public key cryptography. 5

(b) Differentiate conventional and public key cryptography. 5

(c) Define linear congruence. Which algorithm can be used to solve an equatuion of type ac = b (mod n). How can we solve a set of linear equations ? 5

4. (a) Explain working of HMAC. Though it's efficient technique which can be applied for applications where it is infeasible forgery still it's not a choice for application needing non-repudiation. Indentify that problem by stating example. 5

(b) What are the limitations of message authentication? Explain properties and requirements of digital signature. Illustrate with a neat sketch with two approachers for digital signature. 5

(c) User A and B exchange the key using Diffie-Hellman algorthim.

q=11, XA=2, XB=3, Find the value of YA,YB and K? Explain man in middle attack on Diffie-Hellman. 5

5. (a) Explain the authentication application Kerberos versions 5. How does the request for service in another realm is made ? 5

(b) State the format of Certificate Revocation List.Also state the authentication procedure with respect to X.509 certificate. 5

(c) What are the services provided by PGP services? Explain the reasons for using PGP. 5

6. (a) Explain the technical details of firewall and describe any one type of firewall with neat diagram. 5

(b) What are the common techniques used to protect a password file? Explain any one in detail. 5

(c) Explain the types of Host based intrusion detection. List any two IDS software available. 5

Course Code : CST 408-2 **DJPV/MS –16 / 4488**

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

Elective - II

DISTRIBUTED AND PARALLEL DATABASES

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) All questions carry equal marks.
(2) Due credit will be given to neatness.
(3) Assume suitable data wherever necessary.
(4) Illustrate your answers wherever necessary with the help of neat sketches.
(5) Mobile phones are prohibited in examination hall.

1. Solve any **two** :—

(a) Draw and explain reference architecture of distributed database management system. 5

(b) Explain the following with a neat sketch along with the advantages and disadvantages of each.

(1) Shared Memory Architecture

(2) Hierarchical Architecture. 5

(c) Explain parallel nested loop algorithm in the context of parallel database. Also explain the shortcomings of the algorithms. 5

2. Solve any **two** :—

(a) Write benefit equations used in designing horizontal fragmentation for :

(1) Best fit approach

(2) All beneficial sites approach

(3) Additional replication approach. 5

(b) Discuss the correctness condition for horizontal and vertical fragmentation. 5

(c) What do you mean by transparency in the database ? Explain the following in brief :

(1) Local mapping transparency (2) No transparency. 5

3. Solve any **two** :—

(a) How concurrency control is achieved in distributed database ? What should be the characteristics of concurrency control mechanism ? 5

(b) Explain how time stamps are used in distributed database to implement the concurrency control techniques. Give proper example for it. 5

(c) Describe two phase commit protocol and discuss the behaviour of the protocol in presence of :

(1) Lost message

(2) Site failure. 5

4. Solve any **two** :—

(a) Write short note on parallel query optimizations. 5

(b) Explain the optimization plan for the canonical query tree in the processing of distributed grouping and aggregate functions. Give example. 5

(c) Explain how SDD-1 algorithm is used to decide the execution strategy of join queries. Give the algorithm. Also state the shortcomings of this algorithm. 5

5. (a) Write short notes on :

(1) Reliability in distributed database.

(2) Distributed transaction recovery. 8

(b) What do you mean by network partitioning in the context of distributed database ? Explain any approach to overcome this. 2

6. (a) Enlist and explain distributed data mining challenges. 5

(b) Explain any two access methods for data warehousing ? Also explain how parallelization can be applied to improve the performance. 5

Course Code : CST 409-1 **DJPV/MS –16 / 4489**

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 equal marks, carefully see internal choices.
(2) Due credit will be given to neatness and adequate dimensions.
(3) Assume suitable data wherever necessary.
(4) Illustrate your answers wherever necessary with the help of neat sketches.
(5) Retain the construction lines.
(6) Mobile phones are prohibited in examination hall.

1. (a) Google, Facebook, LinkedIn, eBay, Amazon did not use 'traditional' databases for 'big data', why ? 4

OR

(b) What is a page rank and how can it be used in search engines ? 4

(c) There was a murder and an investigating team found a finger prints from a crime spot. There are 100000 FPs available in the database of the agency against which they have to match the FP. Suppose the probability of finding minutia in random grid square of a finger print (FP) is 25%. If a grid is having minutia in all squares of a grid, then the corresponding grid of other FP will also have the minutia with a probability of 75% if the FP is taken from the same finger. Consider each function f in a family of **F** is defined by a 4 grid squares. f says 'yes' if both FPs have minutia in all 4 grid squares otherwise it says 'no'. If we choose 1000 such functions randomly chosen from **F**, find:

(1) What is the probability that F1 will put fingerprints from the same finger together in at least one bucket ?

(2) What is the probability that two fingerprints from different fingers will be placed in the same bucket ?

(3) Calculate % false negatives and % false positives. 6

2. (a) Suppose there is a repo[illegible] of ten million documents, and word w appears in 320 of them. In a [illegible]ular document d, the maximum number of occurrences of a word [illegible]. Approximately what is the TFIDF score for w if that word [illegible] (a) once (b) five times ? 4

(b) Define mutual informat[illegible]ulate in the given table the mutual information for the features *hate* [illegible]*e* towards behaviour *sentiment*. Also justify that mutual informatio[illegible]e a measure for selection of a feature.

Count		Sentiment
2000	I really li[illegible] course and am learning a lot	positive
800	I really h[illegible] course and think it is waste of [illegible]	negative
200	The cour[illegible]lly too simple and quite a bore	negative
3000	The cours[illegible]ple, fun and very easy to follow	positive
1000	I'm enjoyi[illegible] course a lot and learning something [illegible]	positive
400	I would [illegible]yself a lot if I did not have to t[illegible]is course	negative
600	I did not [illegible]his course enough.	negative

6

[illegible]R

3. What is Bayes theorem ? How c[illegible]l probability is used for classification ? Why we require to use Naïve Bayes [illegible]r ? Considering the table given in Question 2 b, find the sentiment of the [illegible] *"I have hated this course since beginning as it is boring and waste of mone[illegible]* Naïve Bayes classification. Consider suitable features to be included. 10

4. (a) Considering an example o[illegible]ord count explain the approach for how map reduce can be used to [illegible]ate a TF.IDF scores of different keywords/ features. 5

(b) Justify parallel efficiency of map-reduce paradigm. Will it be scalable in situations when the input size grows to a considerable extent ? Prove. 5

OR

(c) Explain the process model of map-reduce. Why do you think that this framework will be robust and not fail where sequential systems will fail if used to solve the same problem ? 5

5. (a) What is association rule mining ? Giving an example justify the role of ARM in data mining. 5

(b) Explain Distributed File System and its components. 5

OR

(c) Write a note on Mongo DB. Comment why it is popular over other No SQL databases. 5

6. (a) Write a short note on how reasoning is important when web intelligent applications are built. 4

(b) What is proposition and predicate logic ? Will these systems be capable of handling uncertainty ? Why ? 6

OR

(c) What are Baysian Networks ? With an example describe their use. 6