Lib - 05

Course Code : CST 406 NMPM/RS - 15 /5480

# 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.
(2) Figures to the right indicate marks. Carefully see internal choices.
(3) Solve Q. 6 or Q. 7.

1. (a) What are the differences between the three main types of data warehouse usage : information processing, analytical processing, and data mining ? Discuss the motivation behind OLAP mining (OLAM). 5

**OR**

(b) What is cube ? If we create a CUBE for retail application with three dimensions for time, product, and store, illustrate with an example how the subcubes in the lattice can be created. 5

(c) Write a short note on data warehouse development life cycle. 5

2. (a) With relevant examples discuss multidimensional online analytical processing (MOLAP) and multirelational online analytical processing (ROLAP). 5

(b) What is a STAR schema ? Draw a star schema for your college and specify the hierarchies in it. 5

**OR**

(c) Explain the distinction between :

(i) Measures and Dimensions.

(ii) Fact tables and Dimension tables.

(iii) Star and Snowflake data warehousing Schemes. 5

3. (a) Compare B–tree index and Bitmap index. 5

**OR**

(b) Discuss in detail about hash and functional indexing. 5

(c) Write a short on query optimization. 5

4. (a) 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.

Using this, answer the following :

(i) Use min–max normalization to transform the value 45 for age onto the range [0:0; 1:0].

(ii) Use z–score normalization to transform the value 45 for age, where the standard deviation of age is 12.94 years.

(iii) Use normalization by decimal scaling to transform the value 45 for age.

(iv) Comment on which method you would prefer to use for the given data, giving reasons as to why.

(v) Why is normalization required ? 5

(b) How data mining system are classified ? Discuss any two classification methods with an example. 5

**OR**

(c) How data mining system can be integrated with a data warehouse ? Discuss with an example. 5

5. (a) Discuss in detail the various steps involved in KDD with suitable illustration. 4

(b) Explain the uses and steps involved in Apriori algorithm. Discuss it with suitable example. 6

**OR**

(c) What is association–based classification ? Why does association–based classification achieve higher classification accuracy than a classical decision–tree method ? Explain how association–based classification can be used for text document classification with an example. 6

Solve Q. **6** or Q. **7** :—

6. (a) With a suitable example, explain K–Means Clustering algorithm. 5

(b) Describe the working of PAM (Partitioning Around Medoids) algorithm. 5

**OR**

7. (a) What is an outlier ? Give example in clustering. 3

(b) Distinguish between classification and clustering. 2

(c) Give categorization of major clustering methods. Explain any one method. 5

**Course Code : CST 407** **NMPM/RS – 15 / 5481**

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

## INFORMATION SECURITY

Time : 3 Hours] [Max. Marks : 60

**Instructions to Candidates :—**

(1) All questions carry marks as indicated.
(2) Assume suitable data wherever necessary.
(3) Illustrate your answers wherever necessary with the help of neat sketches.
(4) Mobile phones are prohibited in examination hall.

1. (a) What is the difference between passive and active security threats ? List and briefly define categories of passive and active security attacks. 6

(b) What is message integrity ? Does integrity differ from secrecy or confidentiality or is it implied ? 4

**OR**

2. (a) What are the ways through which protection of information is done in cryptography and how these implementations are supported by different mechanism ? Justify your answer. 6

(b) What is the role of inverses in cryptography ? Justify how it is used to decide the key domain. 4

3. (a) What do you mean by Cipher text only attack ? How this cryptanalysis generally used by the attacker to break different types of substitution Cipher. Justify your answer with at least one example. 4

(b) Encipher the following statement using Playfair Cipher :

"Advance right flank to Bunker Hill, then take up positions ready for attack."

Given Keyword : PLAYFAIR.

(Note : Consider the secret-key matrix filled Left-to-right, in alphabetic order.) 4

**OR**

4. (a) Which parameters and design choices determine the actual algorithm of a Feistel cipher ? 4

(b) What is the purpose of the S – boxes in DES ? 2

(c) What is the purpose of Modes of Operations in symmetric–key encipherment ? Name its modes of operation. 4

5. (a) Define the following terms with respect to public – key cryptography :

(i) Session key.

(ii) Master key.

(iii) Nonce. 3

(b) Describe the security of RSA with its proof, list the possible attacks on RSA. 4

(c) What types of information can be derived from a traffic analysis attack ? 3

**OR**

6. (a) With respect to placement of encryption function, explain the characteristics of :

(i) Link Encryption.

(ii) End – to – End Encryption. 4

(b) Which public key cryptography algorithm is used for selection of key pairs ? Brief about its mathematical proof to recover plain text. 4

(c) Give the significance of a trap – door one – way function. 2

7. (a) What is message authentication ? What types of functions can be used to produce an authenticator ? 4

(b) State whether the following statement is True or False. Justify your answer :

"It is necessary to recover the secret key in order to attack a MAC algorithm." 3

(c) What is the difference between a message authentication code and a one-way hash function ? 3

**OR**

8. (a) What types of attacks are addressed by message authentication ? 4

(b) What are the threats associated with a direct digital signature scheme ? 2

(c) What ways can a hash value be secured so as to provide message authentication ? 4

9. (a) What requirements were defined for Kerberos ? 4

(b) How does signing and verification function proceeds with digital signature algorithm ? 3

(c) Describe X·509 Authentication service. How does user obtain a certificate ? 3

**OR**

10. (a) Explain IPSec architecture. 6

(b) What are the applications of IPSec ? Explain. 4

11. (a) Write a short note on :

(i) SSL Session State.

(ii) SSL Session Connection. 6

(b) What is the role of encryption in the operation of Virus ? 2

(c) What properties are required of a reference monitor ? 2

**OR**

12. (a) What are the identified classes of intruders ? 3

(b) Write a short note on "Distributed Intrusion Detection". 4

(c) What are the design goals for a firewall ? 3

NMPM/RS – 15 / 5482

Course Code : CST 408-2

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

## Elective – III

## DISTRIBUTED AND PARALLEL DATABASES

[Max. Marks : 60

Time : 3 Hours]

**Instructions to Candidates :—**

(1) All questions carry marks as indicated.
(2) **Solve any two subquestions from each question.**
(3) Assume suitable data wherever necessary.
(4) Diagrams should be given wherever necessary.

1. (a) Give a brief account of architectural models for distributed DBMS. 5

(b) Explain the role of the following in parallel execution of a query :

(1) Dependence Analysis.

(2) Optimistic data transmission.

(3) Load Balancing. 5

(c) Explain the following distributed DBMS architecture :

(1) Peer to peer architecture.

(2) Collaborative systems. 5

2. (a) What do you mean by optimal benefits in allocation of fragments ? Give an benefit expression for each of the followings :

(1) Horizontal fragmentation.

(2) Vertical fragmentation.

(3) Vertical clustering. 5

**Contd.**

(b) Consider the following global, fragmentation and allocation schemata :

Global Schema : STUDENT (NUMBER, NAME, DEPT)

Fragmentation Schema : STUDENT1 = $SL_{DEPT = "EE"}$ STUDENT

STUDENT2 = $SL_{DEPT = "CS"}$ STUDENT

Allocation Schema : STUDENT1 at sites 1, 4

STUDENT2 at sites 3, 2

Assume that "EE" and "CS" are the only possible values for DEPT.

(a) Write an application that moves the student having number 232 from department "EE" to department "CS", at levels 1, 2 and 3 of transparency.

(b) Consider the case in which application (a) is repeated for many possible values of student number. Write the application accessing the database after having collected several inputs from the terminal. 5

(c) What is horizontal fragmentation ? How will you ensure the correctness of horizontal fragmentation ? 5

3. (a) What do you mean by distributed concurrency control ? Explain the reference model for distributed concurrency control. 5

(b) What is distributed deadlock ? State and explain any one algorithm for distributed deadlock detection ? 5

(c) Consider a data item x stored at site 1 in a distributed database with 2 sites. Let the triple <Ri, j, TS> (<Wi, j, TS>) denote a read (write) request of transaction Ti on the item x generated at site j with time stamp TS. Assume that initially the following operations are buffered :

<R1, 2, 12>, <W2, 1, 8>, <W3, 2, 9>

Indicate the behavior of the conservative timestamp method with the following sequence of requests :

<R4, 1, 10>, <R5, 1, 11>, <W6, 1, 13>

<W7, 2, 14>, <R8, 1, 14>, <R9, 2, 17> 5

4. (a) Describe simplification techniques for each of the followings with proper example :

(1) Horizontally fragmented relation.

(2) Vertically fragmented relation. 5

(b) Consider the following two scenarios :

(1) Performing join operation without semi join program between two relations.

(2) Implementation of join operation with the help of semi join programs.

Which one will be better over the other one and in which circumstances ? Derive the relationship between the transmission costs of these two scenarios, assuming relations are located on different sites. 5

(c) Comment on the following things when join, semi join and union is performed between two relations :

(1) Cardinality of the result.

(2) Size of the result.

(3) Distinct values for any attribute in result. 5

5. (a) What do you mean by replication management in the context of database reliability ? Explain strict Replica Control Protocol in brief. 5

(b) Draw and explain the reference model for distributed transaction recovery. 5

(c) Explain the behavior of the 2 – phase – commitment protocol in the presence of following :—

(1) Site failures.

(2) Lost messages.

(3) Network Partitions. 5

6. (a) Write short note on the followings :—

(1) Teradata.

(2) Gamma. 5

(b) Explain the mapping and implementation of data warehousing on parallel system. Also explain the advantages and disadvantages of it. 5

(c) Explain the distributed classifier learning algorithm in the context of distributed data mining. 5

**Course Code : CST 409-1** **NMPM/RS – 15 / 5483**

# 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 question carry marks as indicated.
(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) Mobile phones are prohibited in examination hall.
(6) Carefully check internal choices.

1. (a) Explain types of searches: Private and Enterprise. 4

(b) Illustrate the basic text indexing done on the web. Give the algorithm for index creation with its complexity.

Use the following documents and words for creating the index and explanation

D1- "Book of web intelligence is in the library"
D2- "Web intelligence is a course in final year"
D3- "All final year students are placed"

Search Words: web, web intelligence. 3

**OR**

(c) Consider that 1st fingerprint has probability of 0.2 that each of 3 grid squares have minutiae. If another fingerprint from same person has probability of 0.8. Both have grid squares as 3. Assume b = 1024 (no. of functions used). 3

Construct a new family F1 by performing 1024-way OR on F.

(i) Find the probability that fingerprint from same person will be together in at least one bucket.

(ii) Probability that fingerprints from different person will be in same bucket.

(iii) Find the percentage of false negative and false positive.

Comment on what will happen if we increase the number of functions.

(d) What is sparse distributed memory ? Explain how it can be used for approximate recall. 3

2. (a) Explain the concept of mutual information with example and formula.

Find the mutual information :

(1) Between word-"hate" and sentiment(s)

(2) Between word-"course" and sentiment(s)

| Count | | Sentiment |
|---|---|---|
| 2000 | I really **like** this course and am learning a **lot** | Positive |
| 800 | I really **hate** this course and think it is a **waste** of time | Negative |
| 200 | The course is **simple** and quite **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 |

5

(b) Explain Naive Bayes Classifier with an example. 5

**OR**

3. (a) There is a collection of three documents as shown below

D1: "new York times"
D2 : "new York post"

D3 : "los angeles times"
Given the following query:
Q. "new new times"
Find the following :

(1) Term Frequency (TF)

(2) Inverse Document Frequency (IDF)

TF-IDF for documents and query 5

(b) How can machine learning be used to find whether a person intends to shop or surf ? Also explain the use of conditional probabilities. 5

4. (a) Derive the equation of parallel efficiency of map reduce. Is the equation independent of P (processors) ? Show how P affects the efficiency. 4

(b) Consider the following input files and show the map-reduce process to find the word count of words in the set of 8 documents. Use 3 mappers and 2 reducers.

D1- The game of life is a game of everlasting learning.

D2- The unexamined life is not worth living

D3- Never stop learning'

D4- Some papers are unexamined

D5- life is a game

D6- life is full of surprises

D7- life is worth living

D8- Surprises are part of life. 6

**OR**

5. (a) Why id Dremel more consistent than other NO SQL systems ? 3

(b) Consider the following input documents and show the map-reduce process to find the word count of words in the set of documents. Use 5 mappers and 3 reducers.

D1-w1, w1, w2, w3, w5, w6
D2- w1, w3, w4, w5
D3- w2, w3, w4, w5, w5
D4- w2, w2, w3, w3
D5- w1, w3, w4, w6
D6- w1, w2, w3, w4, w5
D7- w1, w2, w2, w3, w6
D8- w1, w3, w2, w5, w6
D9- w2, w2, w2, w2
D10- w3, w4, w4, w5, w6
D11- w1, w2, w3
D12- w4, w5, w6
D13- w2, w5, w6, w3
D14- w2, w3, w4, w4
D15- w5, w6, w2, w3, w3

7

6. (a) Consider the following baskets :

(1) {cat, and, dog, bites}

(2) {yahoo, news, claims, a, cat, and dog, produced, offspring}

(3) {cat, killer, likely, is, a big, dog}

(4) {professional, free, advice, on, dog, training, puppy, training}

(5) {cat, and, kitten, training, and behavior}

(6) {dog, and cat, in, Oregon, provides, training}

(7) {dog, and, cat, is, a, slang, term, used, by, the, police, officers, for, relationship}

(8) {shop, for, your, show, dog, grooming, and, pet, supplies}

Find the singletons, doubles, and triplets having high support, Assume s = 3. Also find the confidence for the following rules :

(1) {cat, dog} à and

(2) {cat} à kitten

5

(b) Write short note on learning. 5

**OR**

7. (a) How are facts learned from text using Hidden Markov Model (HMM) ? Also explain the <subject, verb, object> type of learning. 4

(b) Explain the Searle's Chinese room concept. 2

(c) Explain the long tail phenomenon. 4

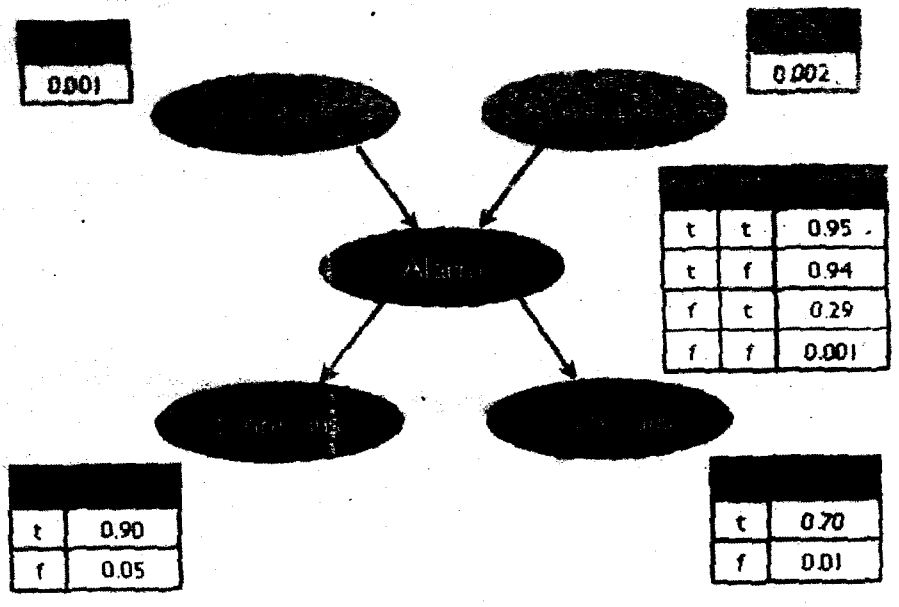
8. (a) Why is going beyond learning to reasoning needed ? What is the motivation behind connecting the learned concepts ? 4

(b) Explain logical propositions and predicates with examples. State its importance in web intelligence and big data. 4

(c) Explain the concept of causality and classification. 2

**OR**

9.




(a) For the above network find the P(burglary|Johncalls) and P(Johncalls|Burglary) 6

(b) Explain the use of probability tables and marginalization in connection of facts. 4

10. (a) Illustrate sparse distributed representations and sequence learning. 5

(b) State the applications of predication where the relationships between prediction, making decisions based on the predictions, and then controlling when actions become abundantly clear. 5

**OR**

(c) Explain hierarchical memory with advantages. 5

28/11/15