Course Title: Data Warehousing and Data Mining **Credit: 3**
Course No: CSIT.416.1 **Number of period per week: 3+3**
Nature of the Course: Theory + Lab **Total hours: 45+45**
Year: Fourth, Semester: Seventh
Level: B. Sc. CSIT

1. Course Introduction

Data warehousing and data mining are two major areas of exploration for knowledge discovery in databases. As more data is collected by businesses and scientific institutions alike, knowledge exploration techniques are needed to gain useful business intelligence. Data mining is for relatively unstructured data for which more sophisticated techniques are needed. The course aims to cover powerful data mining techniques including clustering, association rules, and classification.

2. Objectives

Upon completion of the course, the student should:

- → Be able to define and critically analyze data warehouse and mining approaches
- → Understand the technology of data warehousing.
- → Understand data mining concepts and techniques.
- → Be able to develop applications of higher order database systems.

3. Specific Objectives and Contents

Specific Objectives	**Contents**
• Discuss data mining and KDD and their relationships • Describe data warehouse concepts and needs • Explain functionalities and applications of data mining • Demonstrate data pre-processing steps	**Unit I: Introduction (6 hr)** 1.1. Data Mining Definition, KDD vs.Data Mining, KDD Process, Architecture of Data Mining Systems 1.2. Data Warehouse, Framework of Data Warehouse, Data Mining Functionalities, Classification of Data Mining Systems, Interestingness of Patterns 1.3. Integrating Data Mining with Data Warehouses and Databases, Data Mining Task Primitives, Data Mining Issues and Applications 1.4. Importance of Data Pre-processing, Data Summarization, Data Cleaning, Data Integration and Transformation, Data Reduction, Data Discretization and Concept Hierarchy Generation
• Understand differences between OLAP and OLTP • Describe multidimensional data	**Unit II: Data Warehouse and OLAP (10 hr)** 2.1. Overview of Data Warehouse, Features of Data Warehouse, Operational Database Systems vs Data Warehouse, Need of Separate Data Warehouse 2.2. Multidimensional Data Model and Data Cube,

and their representation using cube • Demonstrate the different schema used for data warehouse representation • Apply DMQL to create data warehouse schema • Demonstrate different OLAP operations • Understand data cube computation and materialization	Schema for Multidimensional Data-Star Schema, Snowflake Schema, Fact Constellation Schema 2.3. DMQL introduction and Syntax, Defining Multidimensional schema by using DMQL, Measures and Its Categories, Using DMQL for finding Measures 2.4. Concept Hierarchies, OLAP Operations- Roll-up, Drill-down, Slicing, Dicing, Pivoting 2.5. Data Warehouse Architecture, Data Warehouse Models, Data Warehouse Backend Tools and Utilities, Metadata, Types of OLAP Servers 2.6. Data Cube Computation, Data Cube Computation , Finding number of Cuboids, Data Cube Materialization, OLAP Query Processing, Data Warehouse Usage 2.7. Cube Materialization- Full Cube, Iceberg Cube, Closed Cube, Shell Cube, Optimization of Cube Computation
• Understand need and importance of association mining • Demonstrate the use of Apriori and FP-Growth algorithms in finding frequent item sets • Use above mentioned algorithms to generate association rules	**Unit III: Association Mining (8 Hrs)** 3.1. Frequent Item Sets, Closed Item Sets, Association Rules, Support & Confidence 3.2. Finding Frequent Item Sets by using Apriori Algorithm, Mining Association Rules from Frequent Items, Improving Efficiency of Apriori Algorithm 3.3. Finding Frequent Item Sets by using FP-Growth Algorithm, Generating Association Rules
• Understand need and importance of classification and prediction • Apply classification algorithms to find class labels • Apply prediction algorithms to make predictions	**Unit IV: Classification and Prediction (8 Hrs)** 4.1. Defining Classification and Prediction, Comparison of Classification and Prediction 4.2. Classification by Decision Trees, Naive Bays Classification, Rule Based Classification, Support Vector Machines 4.3. Prediction-Linear and Non-linear Regression, Accuracy and Error Measures, Evaluating Accuracy of Classifiers and Predictors, Ensemble Methods
• Explain different measures of distances • Understand difference between classification and clustering • Categorize different clustering	**Unit V: Cluster Analysis (8 Hrs)** 5.1. Defining Cluster Analysis, Distance Measures, Types of Data in Cluster Analysis, Categorization of Clustering 5.2. Partition Based Clustering: K-Means Algorithm, K-Medoid Algorithm 5.3. Hierarchical Clustering: Agglomerative Clustering, Divisive Clustering

algorithms • Apply clustering algorithms to divide data into number of groups	5.4. Density Based Methods: DBSCAN Clustering, OPTICS Clustering 5.5. Clustering High Dimensional Data (CLIQUE), Outlier Analysis (Statistical Distribution-Based Outlier Detection
• Explain use of data mining techniques in different areas	**Unit VI: Advanced Data Mining Concepts (5 Hrs)** 6.1. Mining Data Streams, Graph Mining, Social Network Analysis, Multi-relational Data Mining 6.2. Text Mining, Web Mining, Object Mining, Spatial Data Mining, Multimedia Data Mining

Evaluation System

Undergraduate Programs							
External Evaluation	**Marks**	**Internal Evaluation**	**Weight age**	**Marks**	**Practical**	**Weight age**	**Mark**
End semester examination	60	Assignments	20%	20	Practical Report copy	25%	20
(Details are given in the separate table at the end)		Quizzes	10%		Viva	25%	
		Attendance	20%		Practical Exam	50%	
		Internal Exams	50%				
Total External	60	Total Internal	100%	20		100%	20
Full Marks 60+20+20 = 100							

External evaluation

1. **End semester examination:**
   It is a written examination at the end of the semester. The questions will be asked covering all the units of the course. The question model, full marks, time and others will be as per the following grid.

2. **External Practical Evaluation:**
   After completing the end semester theoretical examination, practical examination will be held. External examiner will conduct the practical examination according to the above mentioned evaluation. There will be an internal examiner to assist the external examiner. Three hours time will be given for the practical examination. In this examination Students must demonstrate the knowledge of the subject matter.

Full Marks: 100, Pass Marks: 45, Time: 3 Hrs

Nature of question	Total questions to be asked	Total questions to be answered	Total marks	Weightage
Group A: multiple choice*	20	20	20×1 = 20	60%
Group B: Short answer type questions	7	6	6×8 = 48	60%
Group C: Long answer type questions	3	2	2×16 =32	60%
			100	100%

Each student must secure at least 50% marks in internal evaluation in order to appear in the end semester examination. Failed student will not be eligible to appear in the end semester examinations.

Internal evaluation

Assignment: Each student must submit the assignment individually. The stipulated time for submission of the assignment will be seriously taken.

Quizzes: Unannounced and announced quizzes/tests will be taken by the respective subject teachers. Such quizzes/tests will be conducted twice per semester. The students will be evaluated accordingly.

Attendance in class: Students should regularly attend and participate in class discussion. Eighty percent class attendance is mandatory for the students to enable them to appear in the end semester examination. Below 80% attendance in the class will signify NOT QUALIFIED (NQ) to attend the end semester examination.

Presentation: Students will be divided into groups and each group will be provided with a topic for presentation. It will be evaluated individually as well as group-wise. Individual students have to make presentations on the given topics.

Mid-term examination: It is a written examination and the questions will be asked covering all the topics in the session of the course.

Discussion and participation: Students will be evaluated on the basis of their active participation in the classroom discussions.

Instructional Techniques: All topics are discussed with emphasis on real-world application. List of instructional techniques is as follows:

- Lecture and Discussion
- Group work and Individual work
- Assignments
- Presentation by Students
- Quizzes
- Guest Lecture

Students are advised to attend all the classes and complete all the assignments within the specified time period. If a student does not attend the class(es), it is his/her sole responsibility to cover the topic(s) taught during that period. If a student fails to attend a formal exam/quiz/test, there won't be any provision for re-exam. Unless and until the student clears one semester he/she will not be allowed to study in the following semesters.

Laboratory Work

Student should design data warehouse by using SQL Server or any other tool and then practice different OLAP operations and DMQL queries on it. Besides this students need to implement different association mining, classification and clustering algorithms.

Prescribed Text

- Data Mining Concepts and Techniques, Morgan Kaufmann J. Han, M. Kamber Second Edition

References

- Data Warehousing in the Real Worlds, Sam Anahory and Dennis Murray, Pearson Edition Asia.
- Data Mining Techniques – Arun K. Pajari, University Press.