Syllabus

Course Code	Course Name	Credits
CSC603	Data Warehousing and Mining	4

Course Objectives :

1. To identify the scope and essentiality of Data Warehousing and Mining.
2. To analyze data, choose relevant models and algorithms for respective applications.
3. To study spatial and web data mining.
4. To develop research interest towards advances in data mining.

Course Outcomes : On successful completion of course learner will be able to :

1. Understand Data Warehouse fundamentals, Data Mining Principles.
2. Design data warehouse with dimensional modelling and apply OLAP operations.
3. Identify appropriate data mining algorithms to solve real world problems.
4. Compare and evaluate different data mining techniques like classification, prediction, clustering and association rule mining.
5. Describe complex data types with respect to spatial and web mining.
6. Benefit the user experiences towards research and innovation.

Prerequisite : Basic database concepts, Concepts of algorithm design and analysis.

Module No.	Topics	Hrs.
1.0	**Introduction to Data Warehouse and Dimensional modeling :** Introduction to Strategic Information, Need for Strategic Information, Features of Data Warehouse, Data warehouses versus Data Marts, Top-down versus Bottom-up approach. Data warehouse architecture, metadata, E-R modelling versus Dimensional Modelling, Information Package Diagram, STAR schema, STAR schema keys, Snowflake Schema, Fact Constellation Schema, Factless Fact tables, Update to the dimension tables, Aggregate fact tables.**(Refer chapter 1)**	8
2.0	**ETL Process and OLAP :** Major steps in ETL process, Data extraction : Techniques, Data transformation : Basic tasks, Major transformation types, Data Loading : Applying Data, OLTP Vs OLAP, OLAP definition, Dimensional Analysis, Hypercubes, OLAP operations : Drill down, Roll up, Slice, Dice and Rotation, OLAP models : MOLAP, ROLAP. **(Refer chapter 2)**	8

Module No.	
3.0	Intro Mini Data Attr Pre His Di Co
4.0	C u p
5.0	
6.0	

Scanned by CamScanner

Module No.	Topics	Hrs.
3.0	**Introduction to Data Mining, Data Exploration and Preprocessing :** Data Mining Task Primitives, Architecture, Techniques, KDD process, Issues in Data Mining, Applications of Data Mining, Data Exploration : Types of Attributes, Statistical Description of Data, Data Visualization, Data Preprocessing : Cleaning, Integration, Reduction : Attribute subset selection, Histograms, Clustering and Sampling, Data Transformation and Data Discretization : Normalization, Binning, Concept hierarchy generation, Concept Description : Attribute oriented Induction for Data Characterization. **(Refer chapter 3)**	10
4.0	**Classification, Prediction and Clustering :** Basic Concepts, Decision Tree using Information Gain, Induction : Attribute Selection Measures, Tree pruning, Bayesian Classification : Naive Bayes, Classifier Rule - Based Classification : Using IF-THEN Rules for classification, Prediction : Simple linear regression, Multiple linear regression Model Evaluation and Selection : Accuracy and Error measures, Holdout, Random Sampling, Cross Validation, Bootstrap, Clustering : Distance Measures, Partitioning Methods (k-Means, k-Medoids), Hierarchical Methods (Agglomerative, Divisive). **(Refer chapter 4)**	12
5.0	**Mining Frequent Patterns and Association Rules :** Market Basket Analysis, Frequent Item sets, Closed Item sets and Association Rule, Frequent Pattern Mining, Efficient and Scalable Frequent Item set Mining Methods : Apriori Algorithm, Association Rule Generation, Improving the Efficiency of Apriori, FP growth, Mining frequent Itemsets using Vertical Data Format, Introduction to Mining Multilevel Association Rules and Multidimensional Association Rules. **(Refer chapter 5)**	8
6.0	**Spatial and Web Mining :** Spatial Data, Spatial Vs. Classical Data Mining, Spatial Data Structures, Mining Spatial Association and Co-location Patterns, Spatial Clustering Techniques : CLARANS Extension, Web Mining : Web Content Mining, Web Structure Mining, Web Usage mining, Applications of Web Mining. **(Refer chapter 6)**	6
	Total	52

Scanned by CamScanner

Module 1

Chapter 1 : Introduction to Data Warehouse and Dimensional Modelling 1-1 to 1-60

Syllabus : Introduction to Strategic Information, Need for Strategic Information, Features of Data Warehouse, Data warehouses versus Data Marts, Top-down versus Bottom-up approach, Data warehouse architecture, metadata, E-R modelling versus Dimensional Modelling, Information Package Diagram, STAR schema, STAR schema keys, Snowflake Schema, Fact Constellation Schema, Factless Fact tables, Update to the dimension tables, Aggregate fact tables.

✓ **Syllabus Topic :** Introduction to Strategic Information 1-1

1.1 Introduction to Strategic Information 1-1

✓ **Syllabus Topic :** Need for Strategic Information 1-2

1.2 Need for Strategic Information 1-2

1.2.1 Desired Characteristics of Strategic Information 1-2

1.2.2 Why are Operational Systems not Suitable for Providing Strategic Information? 1-2

1.2.3 Operational v/s Decisional Support System **(May 2016)** 1-3

✓ **Syllabus Topic :** Features of Data Warehouse 1-3

1.3 Intriduction of Data Warehouse **(May 2010, Dec. 2010, May 2011, May 2012, Dec. 2013)** 1-3

1.3.1 Definition Data Warehouse 1-3

1.3.2 Benefits of Data Warehousing 1-4

1.3.3 Features of a Data Warehouse **(Dec. 2010, Dec. 2014)** 1-5

✓ **Syllabus Topic :** Data Warehouses Versus Data Marts 1-6

1.4 Data Warehouses versus Data Marts **(May 2013, May 2016)** 1-6

✓ **Syllabus Topic :** Top-Down versus Bottom-Up Approach 1-7

1.5 Top-Down versus Bottom-Up Approach 1-7

1.5.1 The Top Down Approach : The Dependent Data Mart Structure **(May 2010, Dec. 2010, May 2011, May 2012)** 1-7

1.5.2 The Bottom-Up Approach : The Data Warehouse Bus Structure **(May 2010, Dec. 2010, May 2011, May 2012)** 1-8

1.5.3 Hybrid Approach 1-10

1.5.4 Federated Approach 1-11

1.5.5 A Practical Approach 1-11

✓ **Syllabus Topic :** Data Warehouse Architecture 1-12

1.6 Data Warehouse Architecture 1-12

1.6.1 The Information Flow Mechanism 1-12

Scanned by CamScanner

1.6.2 Data Warehouse Architecture **(May 2010, Dec. 2010, May 2011, Dec. 2011, May 2012, May 2013, May 2014, Dec. 2014, May 2016)** 1-15
1.6.3 Three Tier/ Multi-tier Data Warehouse Architecture **(Dec. 2013)** 1-18
✓ **Syllabus Topic :** Metadata 1-19
1.7 Metadata **(Dec. 2010, Dec. 2012, May 2013, May 2014, Dec. 2014)** 1-19
1.7.1 Definition **(May 2010, Dec. 2011, May 2012, Dec. 2013)** 1-19
1.7.2 Describe Metadata of a Book Store 1-20
1.7.3 Data Warehouse Metadata **(May 2010, May 2012, Dec. 2013)** 1-20
1.7.4 Classification of Metadata or Types of Metadata in Data Warehouse **(May 2010, Dec. 2011, May 2012, Dec. 2013)** 1-21
✓ **Syllabus Topic :** E-R Modelling versus Dimensional Modelling 1-22
1.8 E-R Modelling versus Dimensional Modelling 1-22
1.8.1 What is Dimensional Modeling ? **(Dec. 2011, May 2012)** 1-22
1.8.2 Difference between Data Warehouse Modeling and Operational Database Modeling 1-22
1.8.3 Comparison Database and Data Warehouse Database 1-23
1.8.4 Comparison between Dimensional Model and ER model 1-23
✓ **Syllabus Topic :** Information Package Diagram 1-24
1.9 Information Package Diagram 1-24
✓ **Syllabus Topic :** Star Schema 1-25
1.10 The Star Schema **(Dec. 2010, May 2012, Dec. 2014)** 1-25
✓ **Syllabus Topic :** STAR Schema Keys 1-27
1.11 STAR schema Keys 1-27
✓ **Syllabus Topic :** Snowflake Schema 1-28
1.12 The Snowflake Schema **(May 2011, Dec. 2011, May 2012, Dec. 2012, May 2013, May 2014, Dec. 2014)** 1-28
1.12.1 Differentiate between Star Schema and Snowflake Schema 1-29
1.12.2 Steps of Designing a Dimensional Model 1-29
✓ **Syllabus Topic :** Fact Constellation Schema 1-31
1.13 Fact Constellation Schema or Families of Star **(May 2012, May 2014)** 1-31
✓ **Syllabus Topic :** Factless Fact Tables 1-32
1.14 Factless Fact Tables 1-32
1.14.1 Fact Tables and Dimension Tables 1-32
1.14.2 Factless Fact Table **(Dec. 2012, May 2013)** 1-33
✓ **Syllabus Topic :** Update to the Dimension Tables 1-35
1.15 Update to the Dimension Tables **(May 2016)** 1-35
1.15.1 Slowly Changing Dimensions 1-35
1.15.2 Large Dimension Tables 1-38

Scanned by CamScanner

1.15.3 Rapidly Changing or Large Slowly Changing Dimensions 1-38
1.15.4 Junk Dimensions 1-40
✓ **Syllabus Topic :** Aggregate Fact Tables 1-40
1.16 Aggregate Fact Tables **(May 2014)** 1-40
1.17 Examples on Star Schema and Snowflake Schema 1-40
1.18 University Questions and Answers 1-55
• Chapter Ends 1-60

Module 2

Chapter 2 : ETL Process and OLAP ..... 2-1 to 2-35

Syllabus : Major steps in ETL process, Data extraction : Techniques, Data transformation : Basic tasks, Major transformation types, Data Loading : Applying Data, OLTP Vs OLAP, OLAP definition, Dimensional Analysis, Hypercubes, OLAP operations : Drill down, Roll up, Slice, Dice and Rotation, OLAP models : MOLAP, ROLAP.

2.1 An Introduction to ETL Process **(May 2012)** 2-1
2.1.1 What is ETL Tool? 2-1
2.1.2 Desired Features 2-1
✓ **Syllabus Topic :** Major Steps in ETL Process 2-2
2.2 Major Steps in ETL Process **(May 2012)** 2-2
✓ **Syllabus Topic :** Data Extraction Techniques 2-2
2.3 Data Extraction **(May 2016)** 2-2
2.4 Identification of Data Sources **(May 2016)** 2-3
2.5 Data in Operational Systems **(May 2010, Dec. 2010, May 2011, Dec. 2011, May 2012, Dec. 2012, May 2013, Dec. 2013, May 2014, Dec. 2014, May 2016)** 2-3
2.5.1 Immediate Data Extraction 2-5
2.5.2 Deferred Data Extraction 2-7
✓ **Syllabus Topic :** Data Transformation - Basic Tasks, Major Transformation Types 2-8
2.6 Data Transformation : Tasks Involved in Data Transformation **(May 2010, Dec. 2010, May 2011, Dec. 2011, May 2012, Dec. 2012, May 2013, Dec. 2013, May 2014, Dec. 2014, May 2016)** 2-8
2.6.1 The Set of Basic Tasks 2-10
2.7 Data Integration and Consolidation **(May 2010, Dec. 2010, May 2011, Dec. 2011, May 2012, Dec. 2012, May 2013, Dec. 2013, May 2014, Dec. 2014, May 2016)** 2-10
✓ **Syllabus Topic :** Data Loading - Applying Data 2-11
2.8 Data Loading : Techniques of Data Loading **(May 2010, Dec. 2010, May 2011, Dec. 2011, May 2012, Dec. 2012, May 2013, Dec. 2013, May 2014, Dec. 2014, May 2016)** 2-11
2.8.1 Loading the Dimension Tables 2-11

Scanned by CamScanner

2.8.2 Loading the Fact tables : History and Incremental Loads 2-12
2.9 Data Quality : Issues in Data Cleansing 2-12
2.9.1 Reasons for "Dirty" Data 2-13
2.9.2 Data Cleansing 2-13
2.10 Sample ETL Tools 2-15
2.11 Need for Online Analytical Processing 2-16
✓ **Syllabus Topic :** OLTP V/s OLAP 2-16
2.12 OLTP V/s OLAP **(Dec. 2010, May 2011, May 2012, Dec. 2013)** 2-16
2.13 OLAP and Multidimensional Analysis 2-16
✓ **Syllabus Topic :** Hypercubes 2-17
2.14 Hypercube 2-18
✓ **Syllabus Topic :** OLAP Operations - Drill down, Roll up, Slice, Dice and Rotation 2-18
2.15 OLAP Operations in Multidimensional Data Model **(Dec. 2010, May 2012, Dec. 2012, May 2013, Dec. 2014, May 2016)** 2-20
✓ **Syllabus Topic :** OLAP Models - MOLAP, ROLAP 2-20
2.16 OLAP Models : MOLAP, ROLAP, HOLAP, DOLAP **(May 2016)** 2-25
2.16.1 MOLAP 2-25
2.16.2 ROLAP 2-26
2.16.3 HOLAP 2-27
2.16.4 DOLAP 2-28
2.17 Examples of OLAP 2-28
2.18 University Questions and Answers 2-28
• Chapter Ends 2-32

Module 3

Chapter 3 : Introduction to Data Mining, Data Exploration and Preprocessing **3-1 to 3-64**

Syllabus : Data Mining Task Primitives, Architecture, Techniques, KDD process, Issues in Data Mining, Applications of Data Mining, Data Exploration : Types of Attributes, Statistical Description of Data, Data Visualization, Data Preprocessing : Cleaning, Integration, Reduction : Attribute subset selection, Histograms, Clustering and Sampling, Data Transformation and Data Discretization : Normalization, Binning, Concept hierarchy generation, Concept Description : Attribute oriented Induction for Data Characterization.

3.1 What is Data Mining ? **(Dec. 2011)** 3-1
✓ **Syllabus Topic :** Data Mining Task Primitives 3-2
3.2 Data Mining Task Primitives 3-2

Scanned by CamScanner

✓ **Syllabus Topic :** Architecture 3-5

3.3 Architecture of a Typical Data Mining System **(May 2010, Dec. 2010, Dec. 2011, Dec. 2013, May 2016)** 3-5

✓ **Syllabus Topic :** Techniques 3-6

3.4 Data Mining Technique **(Dec. 2011)** 3-6

3.4.1 Statistics 3-6

3.4.2 Machine Learning 3-6

3.4.3 Information Retrieval (IR) 3-7

3.4.4 Database Systems and Data Warehouses 3-7

3.4.5 Decision Support System 3-8

✓ **Syllabus Topic :** KDD Process 3-8

3.5 Knowledge Discovery in Database (KDD) **(May 2010, Dec. 2010, May 2011, May 2012, Dec. 2012, Dec. 2013, May 2016)** 3-8

✓ **Syllabus Topic :** Issues in Data Mining 3-11

3.6 Major Issues in Data Mining 3-11

✓ **Syllabus Topic :** Applications of Data Mining 3-11

3.7 Applications of Data Mining **(Dec. 2011)** 3-11

✓ **Syllabus Topic :** Data Exploration - Types of Attributes 3-12

3.8 Types of Attributes 3-12

✓ **Syllabus Topic :** Statistical Description of Data 3-15

3.9 Statistical Description of Data 3-15

3.9.1 Central Tendency 3-15

3.9.2 Dispersion of Data 3-18

3.9.3 Graphic Displays of Basic Statistical Descriptions of Data 3-20

✓ **Syllabus Topic :** Data Visualization 3-25

3.10 Data Visualization **(Dec. 2012)** 3-25

✓ **Syllabus Topic :** Data Preprocessing 3-32

3.11 Data Preprocessing 3-32

3.11.1 Form of Data Pre-processing 3-32

✓ **Syllabus Topic :** Cleaning 3-33

3.12 Data Cleaning **(May 2016)** 3-33

3.12.1 Reasons for "Dirty" Data 3-34

3.12.2 Steps in Data Cleansing 3-34

3.12.3 Missing Values 3-35

3.12.4 Noisy Data 3-37

3.12.5 Inconsistent Data 3-41

Scanned by CamScanner

3-5
3-5
3-6
3-6
3-6
3-6
3-7
3-7
3-8
3-8
y 2011,
3-8
3-11
3-11
3-11
3-11
3-12
3-12
3-15
3-15
3-15
3-18
3-20
3-25
3-25
3-32
3-32
3-32
3-33
3-33
3-34
3-34
3-35
3-37
3-41

✓ **Syllabus Topic :** Integration 3-42

3.13 Data Integration **(May 2016)** 3-42

3.13.1 Entity Identification Problem 3-42

3.13.2 Redundancy and Correlation Analysis 3-42

3.13.3 Tuple Duplication 3-43

3.13.4 Data Value Conflict Detection and Resolution 3-46

✓ **Syllabus Topic :** Data Reduction - Attribute Subset Selection, Histograms, Clustering and Sampling 3-47

3.14 Data Reduction **(May 2016)** 3-47

3.14.1 Need for Data Reduction 3-47

3.14.2 Data Reduction Technique 3-47

3.14.2(A) Data Cube Aggregation 3-48

3.14.2(B) Dimensionality Reduction 3-48

3.14.2(C) Data Compression 3-49

3.14.2(D) Numerosity Reduction 3-50

✓ **Syllabus Topic :** Data Transformation and Data Discretization, Normalization, Binning 3-52

3.15 Data Transformation and Data Discretization 3-55

3.15.1 Data Transformation 3-55

3.15.2 Data Discretization 3-55

3.15.3 Data Transformation by Normalization 3-56

3.15.4 Discretization by Binning 3-56

3.15.5 Discretization by Histogram Analysis 3-57

✓ **Syllabus Topic :** Concept Hierarchy Generation 3-58

3.16 Concept Hierarchy Generation 3-58

✓ **Syllabus Topic :** Concept Description - Attribute Oriented Induction for Data Characterization 3-58

3.17 Concept Description : Attribute Oriented Induction for Data Characterization 3-59

3.18 Data Generalization and Summarization-based Characterization 3-60

3.18.1 Data Generalization 3-60

3.18.2 How Attribute-Oriented Induction is Performed? 3-61

3.18.2(A) Data Generalization 3-61

3.18.2(B) Attribute Generalization Control 3-61

3.18.2(C) Example of Attribute Oriented Induction 3-62

3.19 University Questions and Answers 3-63

- Chapter Ends 3-64

Scanned by CamScanner

Module 4

Chapter 4 : Classification, Prediction and Clustering 4-1 to 4-153

Syllabus : Basic Concepts, Decision Tree using Information Gain, Induction : Attribute Selection Measures, Tree pruning, Bayesian Classification : Naive Bayes, Classifier Rule - Based Classification : Using IF-THEN Rules for classification, Prediction : Simple linear regression, Multiple linear regression Model Evaluation and Selection : Accuracy and Error measures, Holdout, Random Sampling, Cross Validation, Bootstrap, Clustering : Distance Measures, Partitioning Methods (k-Means, k-Medoids), Hierarchical Methods (Agglomerative, Divisive).

✓ **Syllabus Topic :** Basic Concepts 4-1

4.1 Basic Concept : Classification 4-1

4.1.1 Classification Problem 4-2

4.1.2 Classification Example 4-2

4.1.3 Classification is a Two Step Process 4-2

4.1.4 Difference between Classification and Prediction 4-4

4.2 Classification Methods 4-5

✓ **Syllabus Topic :** Decision Tree using Information Gain 4-5

4.2.1 Decision Tree Induction 4-5

4.2.1(A) Appropriate Problems for Decision Tree Learning 4-6

4.2.1(B) Decision Tree Representation 4-6

✓ **Syllabus Topic :** Induction - Attribute Selection Measure 4-7

4.2.1(C) Attribute Selection Measure 4-7

4.2.1(D) Algorithm for Inducing a Decision Tree 4-11

✓ **Syllabus Topic :** Tree Pruning 4-13

4.2.1(E) Tree Pruning 4-13

4.2.1(F) Examples of ID3 4-13

✓ **Syllabus Topic :** Bayesian Classification - Naïve Bayes 4-51

4.2.2 Bayesian Classification : Naïve Bayes Classifier 4-51

4.2.2(A) Bayes Theorem 4-51

4.2.2(B) Basics of Bayesian Classification 4-51

4.2.2(C) Naive Bayes Classifier : Example 4-52

4.2.3 Rule based Classification 4-64

4.2.4 Other Classification Methods 4-66

Scanned by CamScanner

✓ **Syllabus Topic :** Prediction - Simple Linear Regression, Multiple Linear Regression 4-66
4.3 Prediction 4-66
4.3.1 Structure of Regression Model 4-66
4.3.2 Linear Regression 4-66
4.3.3 Multiple Linear Regression 4-67
4.3.4 Other Regression Model 4-68
✓ **Syllabus Topic :** Model Evaluation and Selection 4-68
4.4 Model Evaluation and Selection 4-69
✓ **Syllabus Topic :** Accuracy and Error Measures 4-69
4.4.1 Accuracy and Error Measures 4-69
✓ **Syllabus Topic :** Holdout 4-69
4.4.2 Holdout 4-72
✓ **Syllabus Topic :** Random Sampling 4-72
4.4.3 Random Subsampling 4-73
✓ **Syllabus Topic :** Cross-Validation 4-73
4.4.4 Cross-Validation (CV) 4-73
✓ **Syllabus Topic :** Bootstrap 4-73
4.4.5 Bootstrapping 4-75
✓ **Syllabus Topic :** Clustering 4-75
4.5 What is Clustering ? 4-75
4.5.1 What is Clustering ? **(Dec. 2010, May 2012, Dec. 2012, Dec. 2013)** 4-75
4.5.2 Categories of Clustering Methods **(May 2010, May 2012)** 4-77
4.5.3 Difference between Classification and Clustering 4-78
4.6 Types of Data 4-79
4.6.1 Interval-Scaled Variables 4-80
4.6.2 Binary Variable 4-81
4.6.3 Nominal, Ordinal, and Ratio Variables 4-83
4.6.4 Variable of Mixed Types 4-85
✓ **Syllabus Topic :** Distance Measures 4-85
4.7 Distance Measures **(Dec. 2011)** 4-85
✓ **Syllabus Topic :** Partitioning Methods (k-means, k-medoids) 4-87
4.8 Partitioning Methods **(Dec. 2010)** 4-87
4.8.1 K-means Clustering : (Centroid Based Technique) **(May 2010, Dec. 2012, May 2013, Dec. 2013, May 2014, Dec. 2014, May 2016)** 4-87
4.8.2 K-Medoids (Representative Object-based Technique) 4-103
4.8.3 Sampling Based Method 4-108

Scanned by CamScanner

✓ **Syllabus Topic : Hierarchical Methods - Agglomerative, Divisive** 4-109
4.9 Hierarchical Clustering **(May 2014, Dec. 2014)** 4-109
4.9.1 Agglomerative Hierarchical Clustering **(May 2010)** 4-109
4.9.2 Divisive Hierarchical Clustering 4-110
4.9.3 BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) 4-144
4.9.4 Advantages and Disadvantages of Hierarchical Clustering 4-145
4.10 University Questions and Answers 4-149
• Chapter Ends 4-150
..... 4-153

Module 5

Chapter 5 : Mining Frequent Patterns and Association Rules 5-1 to 5-57

Syllabus : Market Basket Analysis, Frequent Item sets, Closed Item sets and Association Rule, Frequent Pattern Mining, Efficient and Scalable Frequent Item set Mining Methods : Apriori Algorithm, Association Rule Generation, Improving the Efficiency of Apriori, FP growth, Mining frequent Itemsets using Vertical Data Format, Introduction to Mining Multilevel Association Rules and Multidimensional Association Rules.

✓ **Syllabus Topic :** Market Basket Analysis 5-1
5.1 Market Basket Analysis **(May 2012, Dec. 2013)** 5-1
5.1.1 What is Market Basket Analysis? 5-1
5.1.2 How is it Used ? 5-1
5.1.3 Applications of Market Basket Analysis 5-2
✓ **Syllabus Topic :** Frequent Item Sets, Closed Item Sets and Association Rule 5-2
5.2 Frequent Item Sets, Closed Item Sets and Association Rule 5-3
5.2.1 Frequent Itemsets 5-3
5.2.2 Closed Itemsets 5-3
5.2.3 Association Rules **(Dec. 2011, May 2012, Dec. 2013)** 5-3
✓ **Syllabus Topic :** Frequent Pattern Mining 5-5
5.3 Frequent Pattern Mining 5-6
✓ **Syllabus Topic :** Efficient and Scalable Frequent Itemset Mining Methods 5-6
5.4 Efficient and Scalable Frequent Itemset Mining Method 5-7
✓ **Syllabus Topic :** Apriori Algorithm 5-7
5.4.1 Apriori Algorithm **(Dec. 2011)** 5-7
5.4.2 Advantages and Disadvantages of Apriori Algorithm 5-7
5.4.3 Solved Examples on Apriori Algorithm 5-8
✓ **Syllabus Topic :** Association Rule Generation 5-9
5.5 Association Rule Generation 5-31
..... 5-31

Scanned by CamScanner

✓ **Syllabus Topic :** Improving the Efficiency of Apriori 5-31
5.6 Improving the Efficiency of Apriori 5-31
✓ **Syllabus Topic :** FP Growth 5-32
5.7 FP Growth **(May 2016)** 5-32
5.7.1 Definition of FP-tree 5-32
5.7.2 FP-Tree Algorithm 5-33
5.7.3 FP-Tree Size 5-34
5.7.4 Example of FP Tree 5-35
5.7.5 Mining Frequent Patterns from FP Tree 5-40
5.7.6 Benefits of the FP-Tree Structure 5-46
✓ **Syllabus Topic :** Mining Frequent Itemsets using Vertical Data Formats 5-46
5.8 Mining Frequent Itemsets using Vertical Data Formats 5-46
✓ **Syllabus Topic :** Introduction to Mining Multilevel Association Rules 5-47
5.9 Introduction to Mining Multilevel Association Rules **(May 2012, May 2013, Dec. 2013, Dec. 2014, May 2016)** 5-47
✓ **Syllabus Topic :** Multidimensional Association Rules 5-49
5.10 Mining Multidimensional (MD) Association Rules **(May 2013, Dec. 2014, May 2016)** 5-49
5.11 University Questions and Answers 5-52
• Chapter Ends 5-57

Module 6

Chapter 6 : Spatial and Web Mining 6-1 to 6-31

Syllabus : Spatial Data, Spatial Vs. Classical Data Mining, Spatial Data Structures, Mining Spatial Association and Co-location Patterns, Spatial Clustering Techniques : CLARANS Extension, Web Mining : Web Content Mining, Web Structure Mining, Web Usage mining, Applications of Web Mining.

✓ **Syllabus Topic :** Spatial Data 6-1
6.1 Spatial Data 6-1
6.1.1 Spatial Pattern 6-2
6.1.2 What is NOT Spatial Data Mining ? 6-2
6.1.3 Why Learn about Spatial Data Mining ? 6-2
6.1.4 Spatial Data Mining : Actors 6-3
6.1.5 Characteristics of Spatial Data Mining 6-3
✓ **Syllabus Topic :** Spatial Vs. Classical Data Mining 6-4
6.2 Spatial Vs. Classical Data Mining 6-4

Scanned by CamScanner

✓ **Syllabus Topic :** Spatial Data Structures 6-4
6.3 Spatial Data Structures 6-4
6.3.1 R-tree 6-4
6.3.2 R+ -tree 6-5
6.3.3 More Spatial Indexing 6-6
✓ **Syllabus Topic :** Mining Spatial Association and Co-location Patterns 6-7
6.4 Mining Spatial Association and Co-location Patterns 6-7
6.4.1 Mining Spatial Association 6-7
6.4.2 Mining Co-location Patterns 6-7
✓ **Syllabus Topic :** Spatial Clustering Techniques - CLARANS Extension 6-8
6.5 Spatial Clustering Techniques : CLARANS 6-12
✓ **Syllabus Topic :** Web Mining 6-12
6.6 Web Mining 6-13
6.6.1 How Web Mining is Different from Classical DM ? 6-13
6.6.2 Benefits of Web Data Mining 6-14
✓ **Syllabus Topic :** Web Content Mining 6-14
6.7 Web Content Mining 6-14
6.7.1 Introduction to Web Content Mining 6-14
6.7.2 Text Mining 6-14
✓ **Syllabus Topic :** Web Usage mining 6-15
6.8 Web Usage Mining 6-16
6.8.1 What is Web Usage Mining ? 6-16
6.8.2 Purpose of Web Usage Mining 6-16
6.8.3 Web Usage Mining Activities 6-17
6.8.4 Web Server Log 6-18
6.8.4(A) Structure of Web Log 6-19
6.8.4(B) Web Server Log - An Example 6-19
✓ **Syllabus Topic :** Web Structure Mining 6-19
6.9 Web Structure Mining 6-22
6.9.1 Introduction to Web Structure Mining 6-22
6.9.2 Techniques of Web Structure Mining 6-22
6.9.2(A) Page Rank Technique Used by Google 6-24
6.9.2(B) CLEVER Technique 6-24
6.10 Web Crawlers 6-25
✓ **Syllabus Topic :** Applications of Web Mining 6-29
6.11 Applications of Web Mining 6-31
- Chapter Ends 6-31
- **Appendix A : Solved University Question Paper of May 2019 ..... A-1 to A-16**

CHAPTER 1

Syllabus :

Introduction to Str
Warehouse, Data
approach. Data w
Modelling. Inform
Snowflake Schem
dimension tables

S

1.1 Introduct

- For the past fe
and large size
- Data warehou
and industrie
- Although it'
this.
- With the he
financial i
companies
decision m
- The funda
informati
- If we ne
different

Scanned by CamScanner

Module 1

CHAPTER 1

Introduction to Data Warehouse and Dimensional Modelling

Syllabus :

Introduction to Strategic Information, Need for Strategic Information, Features of Data Warehouse, Data warehouses versus Data Marts, Top-down versus Bottom-up approach. Data warehouse architecture, metadata, E-R modelling versus Dimensional Modelling, Information Package Diagram, STAR schema, STAR schema keys, Snowflake Schema, Fact Constellation Schema, Factless Fact tables, Update to the dimension tables, Aggregate fact tables.

Syllabus Topic : Introduction to Strategic Information

1.1 Introduction to Strategic Information

- For the past few years, data warehousing solution is becoming a huge trend in medium and large sized companies.
- Data warehouse is no longer just a theorized study in universities, because many business and industries have started to embrace this mainstream phenomenon.
- Although it's not applied in every company yet, we can see the stable demand growth for this.
- With the help of Data warehousing technology, every industry right from retail industry to financial institutions, manufacturing enterprises, government department, airline companies people are changing the way they perform business analysis and strategic decision making.
- The fundamental reason for the inability to provide strategic information is that strategic information was been extracted from the existing operational systems.
- If we need the strategic information, the information must be collected from altogether different types of systems.

Scanned by CamScanner

- Only specially designed decision support systems or informational systems can provide strategic information.
- The result of this is the creation of a new computing environment for the purpose of providing the strategic information for every enterprise.

Syllabus Topic : Need for Strategic Information

1.2 Need for Strategic Information

- Strategic information is required for an enterprise to decide the business strategies and establish the goals for business. Enterprise needs to monitor the results by setting some objectives.
- For examples :
  - Top 3 products sold in 2015
  - Improve the quality of top products
  - Increase sales by 10 % in the south region
  - Retain the present customers
  - Increase the productivity of particular items due to demand
- So to take decisions about these type of objectives, business people need information.

1.2.1 Desired Characteristics of Strategic Information

- **INTEGRATED :** There must be one integrated , enterprise wide view of information system.
- **DATA INTEGRITY :** Information must be accurate and must be conventional to business rules.
- **ACCESSIBLE :** Easily accessible for analysis.
- **CREDIBLE :** Every business factor must have one and only one value.
- **TIMELY :** Information system must be obtainable within the predetermined time frame.

1.2.2 Why are Operational Systems not Suitable for Providing Strategic Information?

- The fundamental reason for the inability to provide strategic information is that strategic information was been extracted from the existing operational systems.
- These operational systems such as University Record system, inventory management, claims processing, outpatient billing, and so on are not designed in a way to provide strategic information.

Scanned by CamScanner

- If we need the strategic information, the information must be collected from altogether different types of systems. Only specially designed decision support systems or informational systems can provide strategic information.

1.2.3 Operational v/s Decisional Support System

→ (MU - May 2016)

- Operational systems are tuned for known transactions and workloads, while workload is not known a priori in a data warehouse.
- Special data organization, access methods and implementation methods are needed to support data warehouse queries (typically multidimensional queries).
- Example, average amount spent on phone calls between 9AM-5PM in Pune during the month of December.

Operational System	Data Warehouse (DSS)
Application oriented	Subject oriented
Used to run business	Used to analyze business
Detailed data	Summarized and refined
Current up to date	Snapshot data
Isolated data	Integrated data
Repetitive access	Ad-hoc access
Clerical user	Knowledge user (manager)
Performance sensitive	Performance relaxed
Few records accessed at a time (tens)	Large volumes accessed at a time (millions)
Read/update access	Mostly read (batch update)
No data redundancy	Redundancy present
Database size 100 MB-100 GB	Database size 100GB - few terabytes

Syllabus Topic : Features of Data Warehouse

1.3 Introduction of Data Warehouse

→ (MU - May 2010, Dec. 2010, May 2011, May 2012, Dec. 2013)

1.3.1 Definition Data Warehouse

The term Data Warehouse was defined by Bill Inmon in 1990, in the following way: "A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". He defined the terms in the sentence as follows :

Scanned by CamScanner

Subject Oriented

Data that gives information about a particular subject instead of about a company's ongoing operations.

Integrated

Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.

Time-variant

All data in the data warehouse is identified with a particular time period.

Non-volatile

Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business.

Ralph Kimball provided a much simpler definition of a data warehouse i.e. "data warehouse is a copy of transaction data specifically structured for query and analysis".

This is a functional view of a data warehouse. Kimball did not address how the data warehouse is built like Inmondid, rather he focused on the functionality of a data warehouse.

1.3.2 Benefits of Data Warehousing

- **Potential high returns on investment and delivers enhanced business intelligence :** Implementation of data warehouse requires a huge investment in lakhs of Rs. But it helps the organization to take strategic decisions based on past historical data and organization can improve the results of various processes like marketing segmentation, inventory management and sales.
- **Competitive advantage :** As previously unknown and unavailable data is available in data warehouse, decision makers can access that data to take decisions to gain the competitive advantage.
- **Saves Time :** As the data from multiple sources is available in integrated form, business users can access data from one place. There is no need to retrieve the data from multiple sources.
- **Better enterprise intelligence :** It improves the customer service and productivity.
- **High quality data :** Data in data warehouse is cleaned and transferred into desired format. So data quality is high.

Scanned by CamScanner

1.3.3 Features of a Data Warehouse

Characteristics/ Features of a Data Warehouse → **(MU - Dec. 2010, Dec. 2014)**

A common way of introducing data warehousing is to refer to the characteristics of a data warehouse :

1.	Subject Oriented	2.	Integrated
3.	Nonvolatile	4.	Time Variant

1. Subject Oriented

- Data warehouses are designed to help analyze data. For example, to learn more about banking data, a warehouse can be built that concentrates on transactions, loans, etc.
- This warehouse can be used to answer questions like "Which customer has taken maximum loan amount for last year?" This ability to define a data warehouse by subject matter, loan in this case, makes the data warehouse subject oriented.

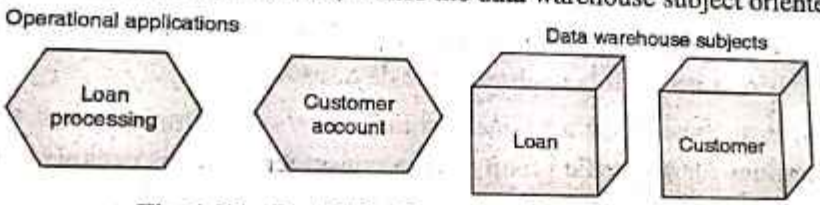


Fig. 1.3.1 : Data Warehouse is subject Oriented

2. Integrated

- A data warehouse is constructed by integrating multiple, heterogeneous data sources like, relational databases, flat files, on-line transaction records. The data collected is cleaned and then data integration techniques are applied, which ensures consistency in naming conventions, encoding structures, attribute measures etc. among different data sources.

Example :

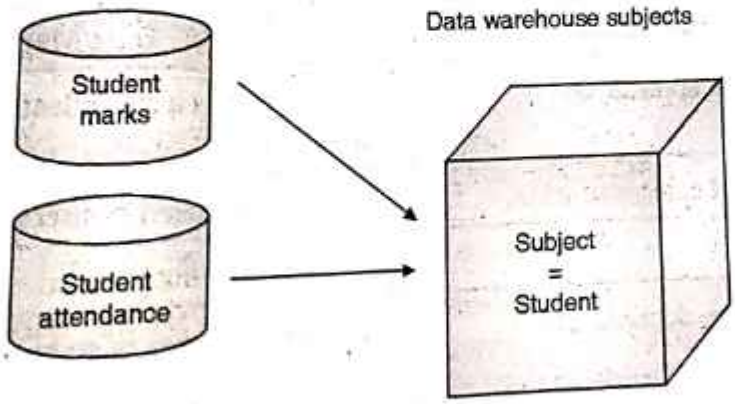


Fig. 1.3.2 : Integrated Data Warehouse

Scanned by CamScanner

3. **Non-volatile**

Nonvolatile means that, once data entered into the warehouse, it cannot be removed or changed because the purpose of a warehouse is to analyze the data.

4. **Time Variant**

A data warehouse maintains historical data. For e.g. A customer record has details of his job, a data warehouse would maintain all his previous jobs (historical information) when compared to a transactional system which only maintains current job due to which its not possible to retrieve older records.

Syllabus Topic : Data Warehouses Versus Data Marts

1.4 Data Warehouses versus Data Marts

→ **(MU - May 2013, May 2016)**

Data Mart defined

- A data mart is oriented to a specific purpose or major data subject that may be distributed to support business needs. It is a subset of the data resource.
- A data mart is a repository of a business organization's data implemented to answer very specific questions for a specific group of data consumers such as organizational divisions of marketing, sales, operations, collections and others.
- A data mart is typically established as one dimensional model or star schema which is composed of a fact table and multi-dimensional table.
- A data mart is a small warehouse which is designed for the department level.
- It is often a way to gain entry and provide an opportunity to learn.
- **Major problem :** If they differ from department to department, they can be difficult to integrate enterprise-wide.

Table 1.4.1 : Differences between Data Warehouse and Data Mart

Data Warehouse	Data Mart
A data warehouse is application independent.	A data mart is a dependent on specific DSS application.
It is centralized, and enterprise wide.	It is decentralized by user area.
It is well planned.	It is possibly not planned.
The data is historical, detailed and summarized.	The data consists of some history, detailed and summarized.

Scanned by CamScanner

Data Warehouse	Data Mart
It consists of multiple subjects.	It consists of a single subject of concern to the user.
It is highly flexible.	It is restrictive.
Implementation takes months to year.	Implementation is done usually in months.
Generally size is from 100 GB to 1TB.	Generally size is less than 100GB.

Syllabus Topic : Top-Down versus Bottom-Up Approach

1.5 Top-Down versus Bottom-Up Approach

Data Warehousing Design Strategies or Approaches for Building a Data Warehouse.

1.5.1 The Top Down Approach : The Dependent Data Mart Structure

→ (MU - May 2010, Dec. 2010, May 2011, May 2012)

- The data flow in the top down OLAP environment begins with data extraction from the operational data sources.

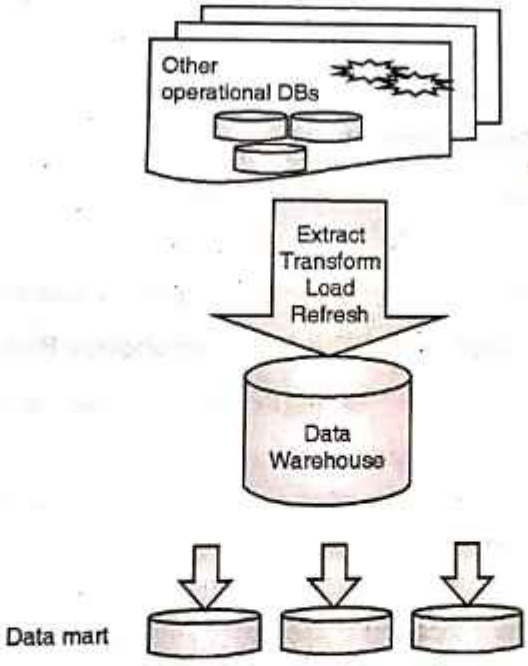


Fig. 1.5.1 : Top Down Approach

- This data is loaded into the staging area and validated and consolidated for ensuring a level of accuracy and then transferred to the Operational Data Store (ODS).
  - The ODS stage is sometimes skipped if it is a replication of the operational databases.

Scanned by CamScanner

- Data is also loaded into the Data warehouse in a parallel process to avoid extracting it from the ODS.
  - Detailed data is regularly extracted from the ODS and temporarily hosted in the staging area for aggregation, summarization and then extracted and loaded into the Data warehouse.
  - The need to have an ODS is determined by the needs of the business.
  - If there is a need for detailed data in the Data warehouse then, the existence of an ODS is considered justified.
  - Else organizations may do away with the ODS altogether.
  - Once the Data warehouse aggregation and summarization processes are complete, the data mart fresh cycles will extract the data from the Data warehouse into the staging area and perform a new set of transformations on them.
  - This will help organize the data in particular structures required by data marts.
- Then the data marts can be loaded with the data and the OLAP environment becomes available to the users.

Advantages of top down Approaches

- It is not just a union of disparate data marts but it is inherently architected.
- The data about the content is centrally stored and the rules and control are also centralized.
- The results are obtained quickly if it is implemented with iterations.

Disadvantages of top down Approaches

- Times consuming process with an iterative method.
- The failure risk is very high.
- As it is integrated a high level of cross functional skills are required.

1.5.2 The Bottom-Up Approach : The Data Warehouse Bus Structure

→ (MU - May 2010, Dec. 2010, May 2011, May 2012)

- This architecture makes the data warehouse more of a virtual reality than a physical reality. All data marts could be located in one server or could be located on different servers across the enterprise while the data warehouse would be a virtual entity being nothing more than a sum total of all the data marts.
- In this context even the cubes constructed by using OLAP tools could be considered as data marts. In both cases the shared dimensions can be used for the conformed dimensions.
- The bottom-up approach reverses the positions of the Data warehouse and the Data marts.

Scanned by CamScanner

- Data marts are directly loaded with the data from the operational systems through the staging area.
- The ODS may or may not exist depending on the business requirements. However, this approach increases the complexity of process coordination.
- The data flow in the bottom up approach starts with extraction of data from operational databases into the staging area where it is processed and consolidated and then loaded into the ODS.
- The data in the ODS is appended to or replaced by the fresh data being loaded.
- After the ODS is refreshed the current data is once again extracted into the staging area and processed to fit into the Data mart structure.
- The data from the Data mart then is extracted to the staging area aggregated, summarized and so on and loaded into the Data Warehouse and made available to the end user for analysis.

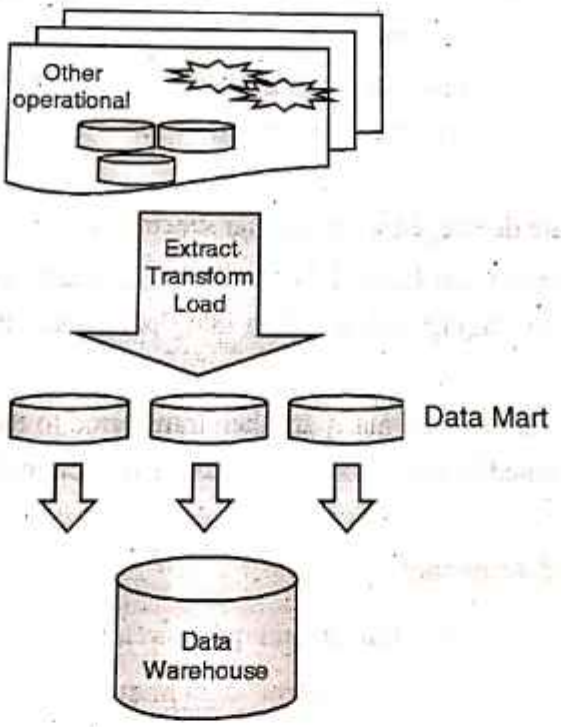


Fig. 1.5.2 : Bottom Up Approach

Advantages of bottom up Approach

- This model strikes a good balance between centralized and localized flexibility.
- Data marts can be delivered more quickly and shared data structures along the bus eliminate the repeated effort expended when building multiple data marts in a non-architected structure.

Scanned by CamScanner

- The standard procedure where data marts are refreshed from the ODS and not from the operational databases ensures data consolidation and hence it is generally recommended approach.
- Manageable pieces are faster and are easily implemented.
- Risk of failure is low.
- Allows one to create important data mart first.

Disadvantages of bottom up Approach

- Allows redundancy of data in every data mart.
- Preserves inconsistent and incompatible data.
- Grows unmanageable interfaces.

1.5.3 Hybrid Approach

- The Hybrid approach aims to harness the speed and user orientation of the Bottom up approach to the integration of the top-down approach.
- The Hybrid approach begins with an Entity Relationship diagram of the data marts and a gradual extension of the data marts to extend the enterprise model in a consistent, linear fashion.
- These data marts are developed using the star schema or dimensional models.
- The Extract, Transform and Load (ETL) tool is deployed to extract data from the source into a non persistent staging area and then into dimensional data marts that contain both atomic and summary data.
- The data from the various data marts are then transferred to the data warehouse and query tools are reprogrammed to request summary data from the marts and atomic data from the Data Warehouse.

Advantages of Hybrid Approach

- Provides rapid development within an enterprise architecture framework.
- Avoids creation of renegade "independent" data marts.
- Instantiates enterprise model and architecture only when needed and once data marts deliver real value.
- Synchronizes meta data and database models between enterprise and local definitions.
- Back filled DW eliminates redundant extracts.

Scanned by CamScanner

Disadvantages of Hybrid Approach

- Requires organizations to enforce standard use of entities and rules.
- Back filling a DW is disruptive, requiring corporate commitment, funding, and application rewrites.
- Few query tools can dynamically query atomic and summary data in different databases.

1.5.4 Federated Approach

- This is a hub-and-spoke architecture often described as the "architecture of architectures". It recommends an integration of heterogeneous data warehouses, data marts and packaged applications that already exist in the enterprise.
- The goal is to integrate existing analytic structures wherever possible and to define the "highest value" metrics, dimensions and measures and share and reuse them within existing analytic structures.
- This may result in the creation of a common staging area to eliminate redundant data feeds or building of a data warehouse that sources data from multiple data marts, data warehouses or analytic applications.
- Hackney-a vocal proponent of this architecture claims that it is not an elegant architecture but it is an architecture that is in keeping with the political and implementation reality of the enterprise.

Advantages of Federated Approach

- Provides a rationale for "band aid" approaches that solve real business problems.
- Alleviates the guilt and stress data warehousing managers might experience by not adhering to formalized architectures.
- Provides pragmatic way to share data and resources.

Disadvantages of Federated Approach

- The approach is not fully articulated.
- With no predefined end-state or architecture in mind, it may give way to unfettered chaos.
- It might encourage rather than dominant in independent development and perpetuate the disintegration of standards and controls.

1.5.5 A Practical Approach

The Steps in the Practical Approach are :

1. The first step is to do Planning and defining requirements at the overall corporate level.
2. An architecture is created for a complete warehouse.
3. The data content is conformed and standardized.

Scanned by CamScanner

4. Consider the series of supermarts one at a time and implement the data warehouse.
5. In this practical approach, first the organizations needs are determined. The key to this approach is that planning is done first at the enterprise level. The requirements are gathered at the overall level.
6. The architecture is established for the complete warehouse. Then the data content for each supermart is determined. Supermarts are carefully architected data marts. Supermart is implemented one at a time.
7. Before implementation checks the data types, field length etc. from the various supermarts, which helps to avoid spreading of different data across several data marts.
8. Finally a data warehouse is created which is a union of all data marts. Each data mart belongs to a business process in the enterprise, and the collection of all the data marts form an enterprise data warehouse.

Syllabus Topic : Data Warehouse Architecture

1.6 Data Warehouse Architecture

1.6.1 The Information Flow Mechanism

- The Building Blocks of a Data Warehouse Or information flow Mechanism as shown in Fig. 1.6.1.

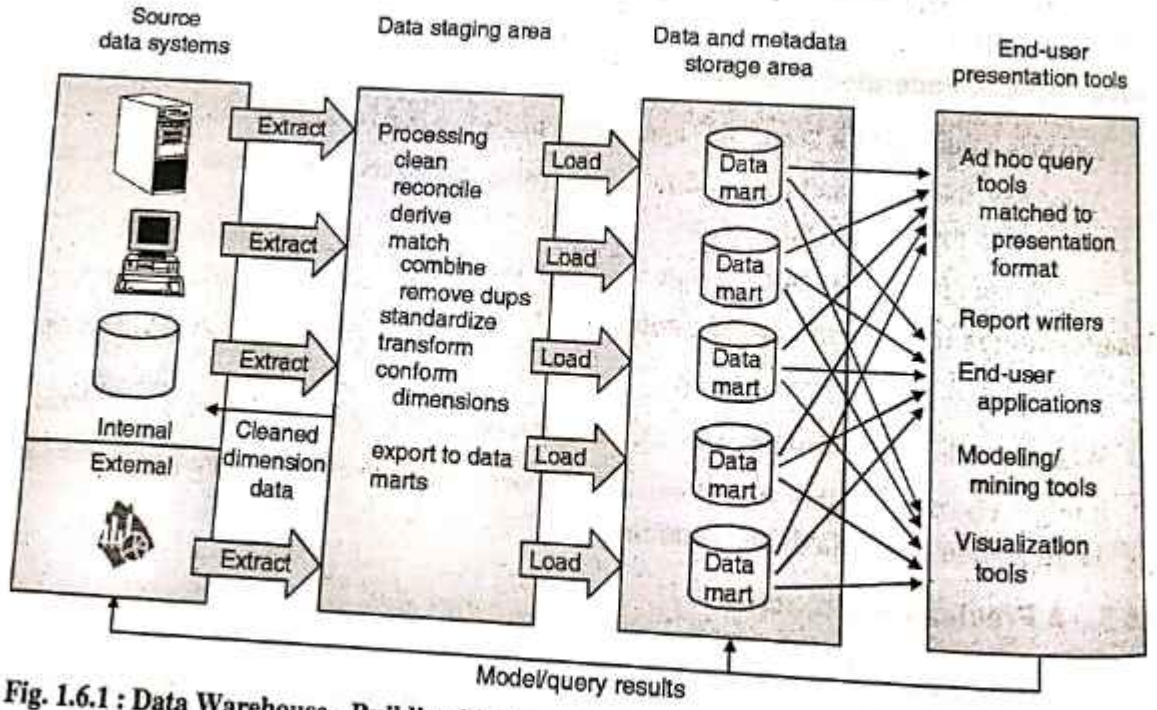


Fig. 1.6.1 : Data Warehouse - Building blocks or components Or Information Flow Mechanism

Scanned by CamScanner

- In Fig. 1.6.1, the source data component is shown on the left, shows various internal and external source data system.
- The next building block is the data staging area where the data is extracted from source data system and then transformation and loading of data is done.
- In the middle, data storage component is placed which manages the data warehouse data.
- A meta data repository is used to keep a track of the data.
- On the rightmost side there is information delivery component. This component is responsible to deliver the information in different ways to the end users.

1. Source Data Component

- The source data component provides the necessary data for the data warehouse.
- Typically, the source data for the warehouse comes from the operational applications.
- This Operational data and processing is completely separated from data warehouse processing.
- The information repository is surrounded by a number of key components designed to make the entire environment functional, manageable and accessible by both the operational systems that source data into the warehouse and by end-user query and analysis tools.

2. Data Staging Component

- As soon as the data arrives into the data staging area, it is cleaned up, and using transformation function it is converted into an integrated structure and format.
- There are different types of transformation functions that may be applied over the data, they are conversion, summarization, filtering and condensation of data.
- As the warehouse maintains historical data, the amount of data is very huge. The warehouse must be able to hold and maintain such a large volume as well as different data structures for the same.
- The functionality includes :
  - Removing unwanted data from operational databases.
  - Converting to common data names and definitions.
  - Establishing defaults for missing data.
  - Accommodating source data definition changes.

3. Data Storage Component

- The central data warehouse database is the cornerstone of the data warehousing environment.
- For implementation, mostly a Relational database management system (RDBMS) technology is used.

Scanned by CamScanner

- However, this kind of implementation is often constrained by the fact that traditional RDBMS products are optimized for transactional database processing.
- Some of the Data warehouse features like very large database size, ad hoc query processing and the need for flexible user view creation including aggregates, multi-table joins and drill-downs have become drivers for different technological approaches to the data warehouse database.

4. Information Delivery Component

- The main purpose of data warehouse is to provide strategic information to business users for decision-making. These users interact with the data warehouse using different tools.
- The information delivery component is used for providing data warehouse information to one or more destinations according to some user specified scheduling algorithm.
- Delivery of information may be based on time of day or on the completion of an external event.

5. Metadata Component

- Meta data is data about data that describes the data warehouse. It is used for building, maintaining, managing and using the data warehouse. Meta data can be classified into :
  - Technical meta data is used by warehouse designers and administrators for warehouse development and management tasks.
  - With Business meta data, users get an easy to understand perspective of the information stored in the data warehouse.

6. Management and Control Component

- Data warehouses are very large as much as 4 times the size of an operational database, reaching terabytes in size depending on how much history needs to be saved.
- Managing the data warehouse environment is very crucial as the data warehouse products include gateways to access multiple enterprise data sources without having to rewrite applications to interpret and utilize the data. The various databases reside on disparate systems thus the need for internetworking tools arises.

Scanned by CamScanner

1.6.2 Data Warehouse Architecture

→ **(MU - May 2010, Dec. 2010, May 2011, Dec. 2011, May 2012, May 2013, May 2014, Dec. 2014, May 2016)**

- The data in a data warehouse comes from operational systems of the organization as well as from other external sources. These are collectively referred to as ***source systems***.
- The data ***extracted*** from source systems is stored in an area called ***data staging area***, where the data is cleaned, ***transformed***, combined, and duplicated to prepare the data in the data warehouse.
- The data staging area is generally a collection of machines where simple activities like sorting and sequential processing takes place. The data staging area does not provide any query or presentation services.
- As soon as a system provides query or presentation services, it is categorized as a ***presentation server***.
- A presentation server is the target machine on which the data is ***loaded*** from the data staging area organized and stored for direct querying by end users, report writers and other applications.
- The three different kinds of systems that are required for a data warehouse are :
  (i) Source Systems (ii) Data Staging Area
  (iii) Presentation servers
- The data travels from source systems to presentation servers via the data staging area. The entire process is popularly known as ETL (extract, transform, and load) or ETT (extract, transform, and transfer). Oracle's ETL tool is called Oracle Warehouse Builder (OWB) and MS SQL Server's ETL tool is called Data Transformation Services (DTS).
- A typical architecture of a data warehouse is shown in Fig.1.6.2 :

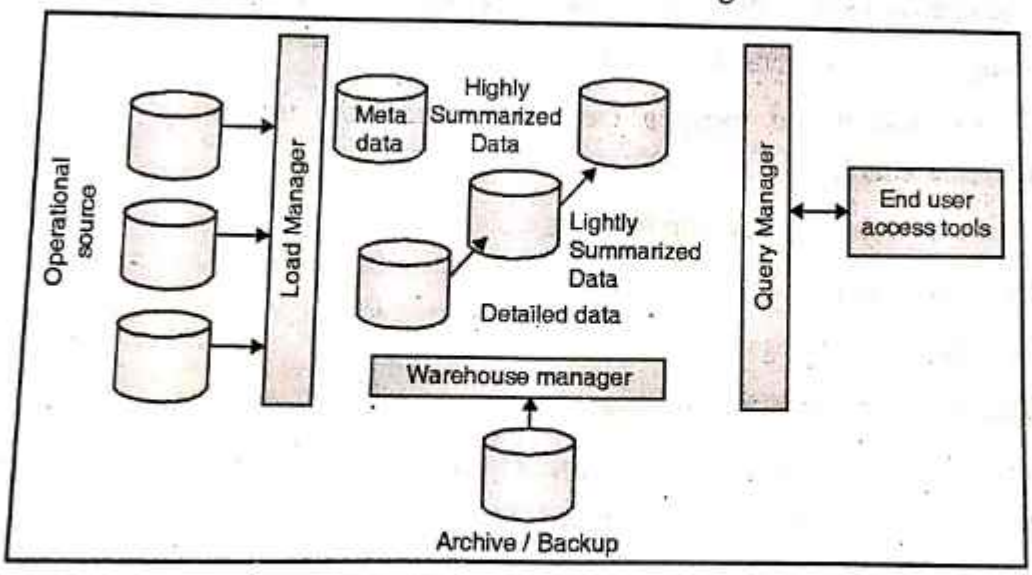


Fig. 1.6.2 : Data Warehouse Architecture

Scanned by CamScanner

Each component and the tasks performed by them are explained below :

1. Operational Source

The sources of data for the data warehouse are supplied from :

- The data from the mainframe systems in the traditional network and hierarchical format.
- Data can also come from the relational DBMS like Oracle, Informix.
- In addition to these internal data, operational data also includes external data obtained from commercial databases and databases associated with supplier and customers.

2. Load Manager

- The load manager performs all the operations associated with extraction and loading data into the data warehouse.
- These operations include simple transformations of the data to prepare the data for entry into the warehouse.
- The size and complexity of this component will vary between data warehouses and may be constructed using a combination of vendor data loading tools and custom built programs.

3. Warehouse Manager

- The warehouse manager performs all the operations associated with the management of data in the warehouse.
- This component is built using vendor data management tools and custom built programs.
- The operations performed by warehouse manager include.
- Analysis of data to ensure consistency.
- Transformation and merging the source data from temporary storage into data warehouse tables.
- Create indexes and views on the base table.
- Denormalization.
- Generation of aggregation.
- Backing up and archiving of data.
- In certain situations, the warehouse manager also generates query profiles to determine which indexes and aggregations are appropriate.

Scanned by CamScanner

4. Query Manager

- The query manager performs all operations associated with management of user queries.
- This component is usually constructed using vendor end-user access tools, data warehousing monitoring tools, database facilities and custom built programs.
- The complexity of a query manager is determined by facilities provided by the end-user access tools and database.

5. Detailed Data

- This area of the warehouse stores all the detailed data in the database schema.
- In the majority of cases detailed data is not stored online but aggregated to the next level of details.
- The detailed data is added regularly to the warehouse to supplement the aggregated data.

6. Lightly and Highly Summarized Data

- This stores all the predefined lightly and highly summarized (aggregated) data generated by the warehouse manager.
- This area of the warehouse is transient as it will be subject to change on an ongoing basis in order to respond to the changing query profiles.
- The main goal of the summarized information is to speed up the query performance.
- As the new data is loaded into the warehouse, the summarized data is updated continuously.

7. Archive and Back up Data

- The detailed and summarized data are stored for the purpose of archiving and back up.
- The data is transferred to storage archives such as magnetic tapes or optical disks.

8. Meta Data

- The data warehouse also stores all the Meta data (data about data) definitions used by all processes in the warehouse.
- It is used for variety of purpose including :
  - The extraction and loading process - Meta data is used to map data sources to a common view of information within the warehouse.
  - The warehouse management process - Meta data is used to automate the production of summary tables.

Scanned by CamScanner

- As part of Query Management process - Meta data is used to direct a query to the most appropriate data source.

- The structure of Meta data will differ in each process, because the purpose is different.

9. End-User Access Tools

- The main purpose of a data warehouse is to provide information to the business managers for strategic decision-making.
- These users interact with the warehouse using end user access tools.
- Some of the examples of end user access tools can be :
  - Reporting and Query Tools
  - Application Development Tools
  - Executive Information Systems Tools
  - Online Analytical Processing Tools
  - Data Mining Tools

1.6.3 Three Tier/ Multi-tier Data Warehouse Architecture

→ **(MU - Dec. 2013)**

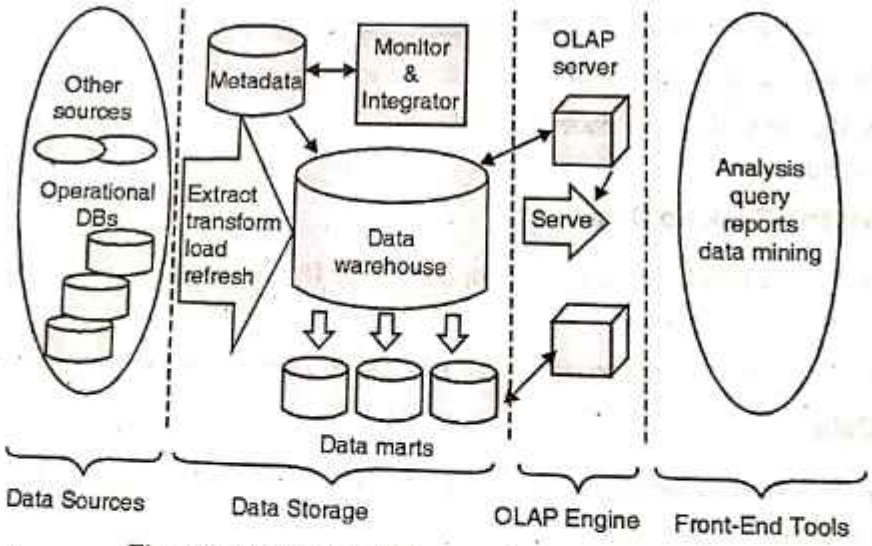


Fig. 1.6.3 : Multi-tier Data warehouse Architecture

1. Bottom Tier (Data Sources and Data Storage)

- It is a warehouse database server, that is generally a RDBMS.
- Using Application Program interfaces (called as gateways), data is extracted from operational and external sources.

Scanned by CamScanner

- Gateways like, ODBC(Open Database connection), OLE-DB (Open linking and embedding for database), JDBC (Java Database Connection) is supported by underlying DBMS.

2. Middle Tier (OLAP Engine)

OLAP Engine is either implemented using ROLAP (Relational online Analytical Processing) or MOLAP(Multidimensional OLAP).

3. Top Tier (Front End Tools)

- This tier is a client which contains query and reporting tools, Analysis tools, and /or data mining tools.
- From the Architecture Point of view there are three data warehouse Models :

I. Enterprise Warehouse

The information of the entire organization is collected related to various subjects in enterprise warehouse.

II. Data Mart

- A subset of Warehouse that is useful to a specific group of users.
- It can be categorized as Independent vs. dependent data mart.

III. Virtual warehouse

- A set of views over operational databases.
- Only some of the possible summary views may be materialized.

Syllabus Topic : Metadata

1.7 Metadata

→ (MU - Dec. 2010, Dec. 2012, May 2013, May 2014, Dec. 2014)

1.7.1 Definition

→ (MU - May 2010, Dec. 2011, May 2012, Dec. 2013)

Data warehouse metadata are pieces of information stored in one or more special-purpose *metadata repositories* that include.

(a) Information on the contents of the data warehouse, their location and their structure,

(b) Information on the processes that take place in the data warehouse back-stage, concerning the refreshment of the warehouse with clean, up-to-date, semantically and structurally reconciled data,

(c) Information on the implicit semantics of data (with respect to a common enterprise model), along with any other kind of data that aids the end-user exploit the information of the warehouse,

Scanned by CamScanner

(d) Information on the infrastructure and physical characteristics of components and the sources of the data warehouse and,

(e) Information including security, authentication, and usage statistics that aids the administrator, tunes the operation of the data warehouse as appropriate.

1.7.2 Describe Metadata of a Book Store

Metadata can also be considered as an equivalent of Amazon book store. If we consider each data element as a book, the meta-data will contain.

- Name of the book,
- Summary of the book,
- Assessments about the book,
- The date of publication,
- High level description of what it contains,
- Who are the publishers,
- How you can find the book,
- Author of the book,
- Whether the book is available OR not.
- This information helps you to :
  - Search for the book
  - Access the book
  - Understand about the book before you access OR buy it.

1.7.3 Data Warehouse Metadata

→ **(MU - May 2010, May 2012, Dec. 2013)**

Data warehousing has specific metadata requirements. Metadata that describes tables typically includes :

- Physical Name
- Logical Name
- Type : Fact, Dimension, Bridge
- Role : Legacy, OLTP, Stage,
- DBMS : DB2, Informix, MS SQL Server, Oracle, Sybase
- Location
- Definition
- *Notes*

Scanned by CamScanner

Metadata describes columns within tables

Physical Name	Logical Name
Order in Table	Data type
Length	Decimal Positions
Nullable / Required	Default Value
Edit Rules	Definition
Notes	

Example of Meta data for Customer sales data warehouse

- Entity Name : *Customer*
- Alias Names : Account, Client
- Definition : A person or an organization that purchases goods or services from the company.
- Remarks : Customer entity includes regular, current, and past customers.
- Source Systems : Finished Goods Orders, Maintenance Contracts, Online Sales.
- Create Date : June 15, 2007
- Last Update Date : *June 21, 2009*
- Update Cycle : *Weekly*
- Last Full Refresh Date : *December 29, 2008*
- Full Refresh Cycle : *Every six months*
- Data Quality Reviewed : *July 25, 2009*
- Last Deduplication : *June 05, 2009*
- Planned Archival : *Every six months*
- Responsible User : *Pallavi H.*

1.7.4 Classification of Metadata or Types of Metadata in Data Warehouse

→ (MU - May 2010, Dec. 2011, May 2012, Dec. 2013)

Operational Meta data

- In an Enterprise, data for the data warehouse comes from various operational systems.
- Different source systems contain different data structures having different field lengths and data types.
- So the information of operational data source is given by Operational Meta Data.

Scanned by CamScanner

Extraction and Transformation Meta data

- This contains the information about the extraction of data from heterogeneous sourc system.
- It also contains information about data transformation in data staging area.

End User Meta data

This particular category provides the end user the flexibility of looking for information in their own way.

Syllabus Topic : E-R Modelling versus Dimensional Modelling

1.8 E-R Modelling versus Dimensional Modelling

1.8.1 What is Dimensional Modeling ?

→ **(MU - Dec. 2011, May 2012)**

- It is a logical design technique used for data warehouses.
- Dimensional model is the underlying data model used by many of the commercial OLAP products available today in the market.
- Dimensional model uses the relational model with some important restrictions.
- It is one of the most feasible technique for delivering data to the end users in a data warehouse.
- Every dimensional model is composed of at least one table with a multipart key called the *fact table* and a set of smaller tables called *dimension tables*.

1.8.2 Difference between Data Warehouse Modeling and Operational Database Modeling

Operational Database Modeling	Data Warehouse Modeling
Current Values are the Data Content.	Data is Archived, Derived, Summarized.
Data structure is Optimized for transactions.	Data structure is Optimized for complex queries.
Access frequency is High.	Access frequency is Medium to low.
Data Access type is Read, Update, Delete.	Data access type is only Read.
Usage is Predictable, Repetitive.	Usage is Ad hoc, Random, Heuristic.
Response time is in Sub - seconds.	Response time is in Several seconds to minutes.
Large Number of users.	Relatively small number of users.

Scanned by CamScanner

1.8.3 Comparison Database and Data Warehouse Database

1. Used for Online Transactional Processing (OLTP) but can be used for other purposes such as Data Warehousing. This records the data from the user for history.
2. The tables and joins are complex since they are normalized (for RDMS). This is done to reduce redundant data and to save storage space.
3. Entity - Relational modeling techniques are used for RDMS database design.
4. Optimized for write operation.
5. Performance is low for analysis queries.

Data Warehouse

1. Used for Online Analytical Processing (OLAP). This reads the historical data for the Users for business decisions.
2. The Tables and joins are simple since they are de-normalized. This is done to reduce the response time for analytical queries.
3. Data - Modeling techniques are used for the Data Warehouse design.
4. Optimized for read operations.
5. High performance for analytical queries.
6. Is *usually* a Database.

1.8.4 Comparison between Dimensional Model and ER model

Sr. No.	Dimensional Model	ER Model
1.	Support ad-hoc querying for business analyst and complex analyzes (data warehouse and multidimensional database)	Support for OLTP and ODS(Operational Data store)
2.	Entities are linked through a series of joins.	Entities are linked through a series of joins.
3.	Simplify the view of the data model. You can rotate the data cube to see different views of the data.	The data model has only one dimension.
4.	It is asymmetric.	It is symmetric. All tables look the same.
5.	Permit redundancy.	Remove the redundancy in data.
6.	It is extensible to accommodate unexpected new data elements and new design decisions. The application is not changed.	If the model is modified, the applications are modified.

Scanned by CamScanner

Sr. No.	Dimensional Model	ER Model
7.	It is robust. The dimensional model design can be done independent of expected query patterns.	It is variable in structure and very vulnerable to changes in the user's querying habits.
8.	The model is easy and understandable.	The model for enterprise is very hard for people to visualize and keep in their heads.
9.	The model really models a business. It is a body of standard approaches for handling common modeling situations in the business world.	The model does not really model a business. It models the micro relationships among data elements.

Syllabus Topic : Information Package Diagram

1.9 Information Package Diagram

- Information package diagram is the approach to determine the requirement of data warehouse.
- It gives the metrics which specifies the business units and business dimensions.
- The information package diagram defines the relationship between the subject or dimension matter and key performance measures (facts).
- The information package diagram shows the details that users want so its effective for communication between the user and technical staff.

Table 1.9.1 : Information Package for Hotel Occupancy

Hotel	Room Type	Time	Room Status
Hotel Id	Room id	Time id	Status id
Branch Name	room type	Year	Status Description
Branch Code	room size	Quarter	
Region	number of beds	Month	
Address	type of bed	Date	
city/stat/zip	max occupants	day of week	
construction year	Suite	day of month,	
renovation year		holiday flag	

Scanned by CamScanner

Facts

(a) Occupied Rooms (b) Vacant Rooms (c) Unavailable Rooms
(d) No of occupants (e) Revenue

Syllabus Topic : Star Schema

1.10 The Star Schema

→ **(MU - Dec. 2010, May 2012, Dec. 2014)**

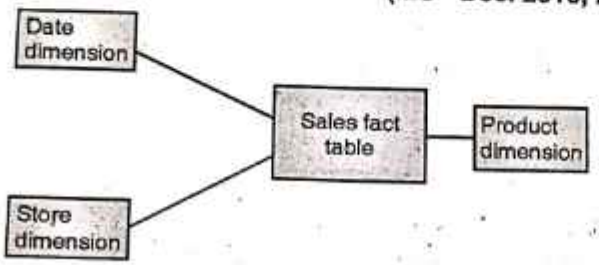


Fig. 1.10.1 : Sales Star Schema

- Star Schema is the most popular schema design for a data warehouse.
- Dimensions are stored in a Dimension table and every entry has its own unique identifier.
- Every Dimension table is related to one or more fact tables.
- All the unique identifiers (primary keys) from the dimension tables make up for a composite key in the fact table.
- The fact table also contains facts. For example a combination of store_id, date_key and product_id giving the amount of a certain product sold on a given day at a given store.
- Foreign keys for the dimension tables are contained in a fact table. For eg.(date key, product id and store id) are all three foreign keys.
- In a dimensional modeling fact tables are normalised, whereas dimension tables are not.
- The size of the fact tables is large as compared to the dimension tables.
- The Facts in the star schema can be classified into three types :

(i) **Fully-additive :** Additive facts are facts that can be summed up through all of the dimensions in the fact table.

(ii) **Semi-additive :** Semi-additive facts are facts that can be summed up for some of the dimensions in the fact table, but not the others.

Example : Bank Balances : You can take a bank account as Semi- Additive since a current balance for the account can't be summed as time period; but if you want see current balance of a bank you can sum all accounts current balance.

Scanned by CamScanner

(iii) **Non-additive :** Non-additive facts are facts that cannot be summed up for any of the dimensions present in the fact table.

E.g. : Ratios, Averages and Variance.

Advantages of Star Schema

- A star schema describes aspects of a business. It is made up of multiple dimension tables and one fact table. For e.g. if you have a book selling business, some of the dimension tables would be customer, book, catalog and year. The fact table would contain information about the books that are ordered from each catalog by each customer during a particular year.
- Reduced Joins, Faster Query Operation.
- It is fully denormalized schema.
- Simplest DW schema.
- Easy to understand.
- Easy to Navigate between the tables due to less number of joins.
- Most suitable for Query processing.

Example :

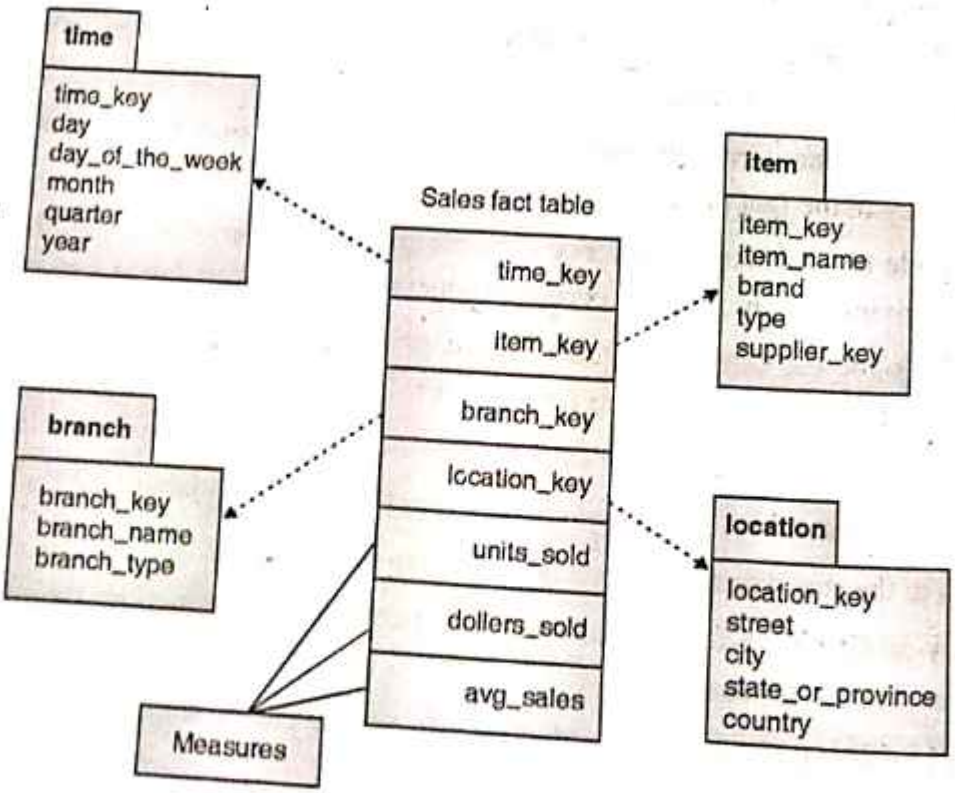


Fig. 1.10.1 : Sales Star Schema

Scanned by CamScanner

Syllabus Topic : STAR Schema Keys

1.11 STAR schema Keys

(i) Primary Keys : The primary key of the dimension table is one of the attribute value which identifies each row in a dimension table uniquely.

Example :

In Professor_Dimdimension, Prof_id is primary key of dimension Professor_Dim.

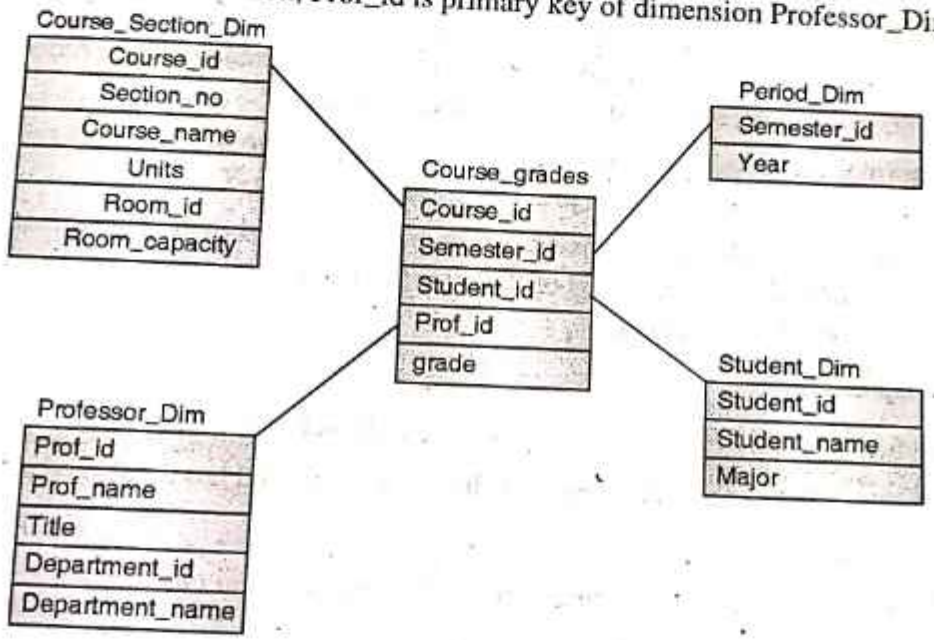


Fig. 1.11.1

(ii) Surrogate Keys

- These keys are system generated sequence numbers.
- They do not have any built in meanings.
- All data warehouse keys must be meaningless surrogate keys.
- One must not use the original production keys.
- A four byte integer makes a good surrogate key.

(iii) Foreign Keys

Every Dimension table has one-to-many relationship with the fact table. So the primary key of each dimension table is a foreign key in the fact table.

Scanned by CamScanner

Syllabus Topic : Snowflake Schema

1.12 The Snowflake Schema

➔ **(MU - May 2011, Dec. 2011, May 2012, Dec. 2012, May 2013, May 2014, Dec. 2014**

- A snowflake schema is used to remove the low cardinality i.e attributes having low distinct values, textual attributes from a dimension table and placing them in a secondary dimension table.
- For e.g. in Sales Schema, the product category in the product dimension table can be removed and placed in a secondary dimension table by normalizing the product dimension table. This process is carried out on large dimension tables.
- It is a normalization process carried out to manage the size of the dimension tables. But this may affect its performance as joins needs to be performed.
- In a star schema, if all the dimension tables are normalised then this schema is called as snowflake schema, and if only few of the dimensions in a star schema are normalised then it is called as star flake schema.

Star Flake Schema

- It is a hybrid structure (i.e. star schema + snowflake schema).
- Every fact points to one tuple in each of the dimensions and has additional attributes.
- Does not capture hierarchies directly.
- Straightforward means of capturing a multiple dimension data model using relations.

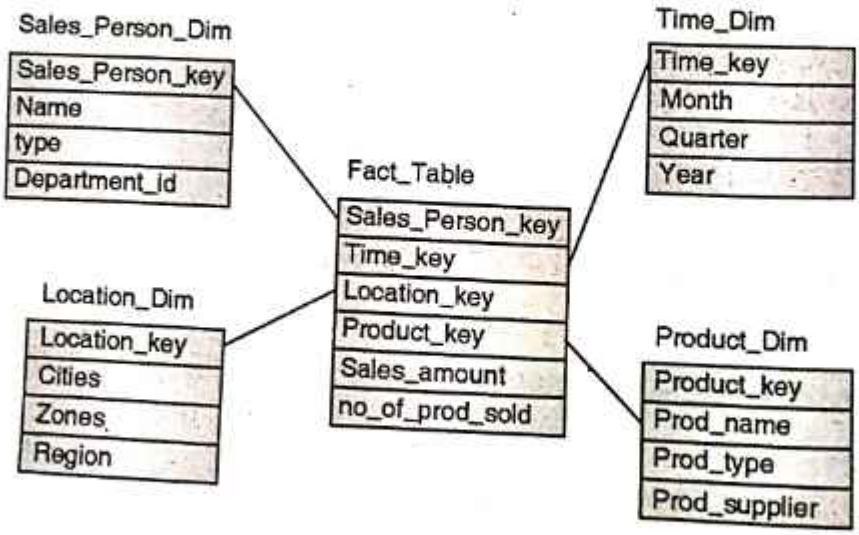


Fig. 1.12.1 Contd.....

Scanned by CamScanner

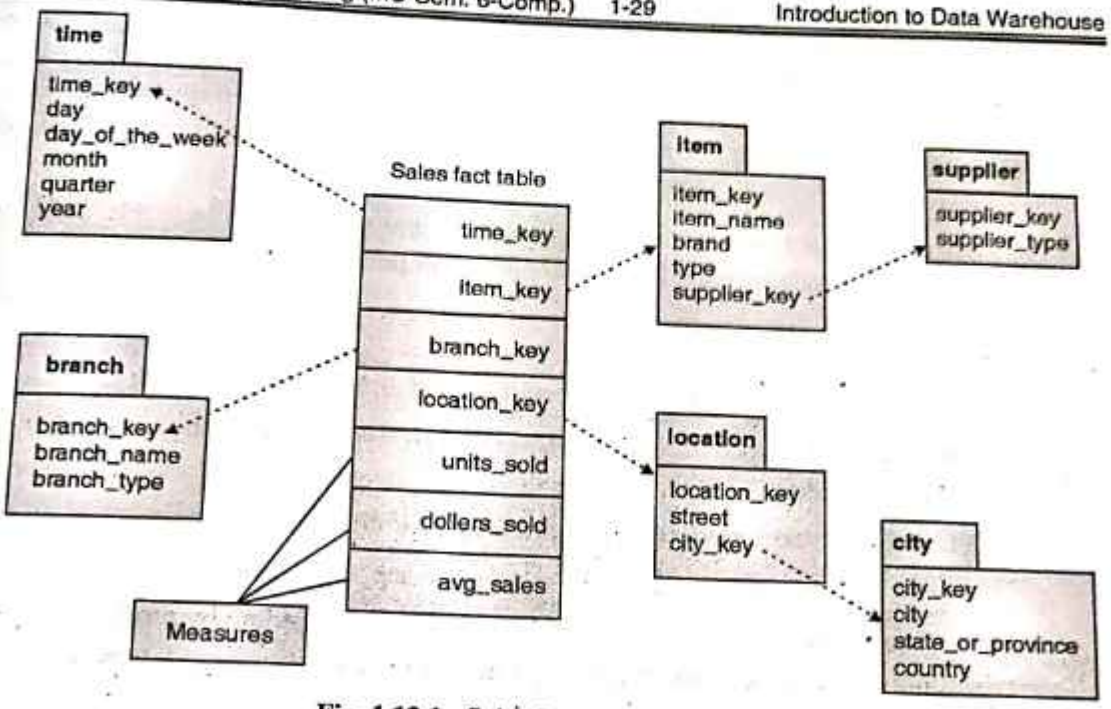


Fig. 1.12.1 : Sales Snowflake Schema

1.12.1 Differentiate between Star Schema and Snowflake Schema

Sr. No.	Star Schema	Snowflake Schema
1.	Star schema contains the dimension tables mapped around one or more fact tables.	A Snowflake schema contains in-depth joins because the tables are split in to many pieces.
2.	It is a de-normalized model.	It is the normalised form of Star schema.
3.	No need to use complicated joins.	Have to use complicated joins, since it has more tables.
4.	Queries results fast.	There will be some delay in processing the Query.
5.	Star Schemas are usually not in BCNF form. All the primary keys of the dimension tables are in the fact table.	In Snowflake schema, dimension tables are in 3NF, so there are more dimension tables which are linked by primary – foreign key relation.

1.12.2 Steps of Designing a Dimensional Model

Step1 : Select the Business Process

The first step in the design is to decide what business process(es) to model by combining an understanding of the business requirements with an understanding of the available data.

Scanned by CamScanner

Step 2 : Declare the Grain

- Once the business process has been identified, the data warehouse team faces a serious decision about the granularity. What level of detail must be made available in the dimensional model?
- The grain of a fact table represents the level of information detail in a fact table. Declaring the grain means specifying exactly what an individual fact table record represents.
- It is recommended that the most atomic information captured by a business process. Atomic data is the most detailed information collected. The more detailed and atomic the fact measurements are, the more we know and we can analyze the data better.
- In the star schema discussed above, the most detailed data would be transaction line item detail in the sale receipt.

  (date, time, product code, product_name, price/unit, number_of_units, amount)

  = (18-SEP-2002, 11.02, p1, dettol soap, 15, 2, 30)
- But in the above dimensional model we provide sales data rolled up by product(all records corresponding to the same product are combined) in a store on a day. A typical fact table record would look like this :

  18-SEP-2002, Product1, Store1, 150, 600
- This record tells us that on 18th Sept. 150 units of Product1 was sold for Rs. 600 from Store1.

Step 3 : Choose the Dimensions

- Once the grain of the fact table has been chosen, the date, product, and store dimensions are readily identified.
- It is often possible to add more dimensions to the basic grain of the fact table, where these additional dimensions naturally take on only one value under each combination of the primary dimensions.
- If the additional dimension violates the grain by causing additional fact rows to be generated, then the grain must be revised to accommodate this dimension.

Step 4 : Identify the Facts

- The first step in identifying fact tables is where we examine the business, and identify the transaction that may be of interest.
- In our example the Electronic Point of Sale (EPOS) transactions give us two facts, quantity sold and sale amount.

Syllabus

1.13 Fact Constellation S

Fact Constellation

- As its name implies, it is sl
- This schema is more comp it contains multiple fact ta
- This allows dimension ta
- A schema of this type sophistication.
- For each star schema o schema.
- That solution is very fl
- The main disadvantag because many variants
- In a fact constellati dimensions, which ar
- This may be useful i and other facts with
- Use of that model s details down to the which is calculated
- In that case using through a fact con

Family of stars

Scanned by CamScanner

Syllabus Topic : Fact Constellation Schema

1.13 Fact Constellation Schema or Families of Star

→ **(MU - May 2012, May 2014)**

Fact Constellation

- As its name implies, it is shaped like a constellation of stars (i.e., star schemas).
- This schema is more complex than star or snowflake varieties, which is due to the fact that it contains multiple fact tables.
- This allows dimension tables to be shared amongst the fact tables.
- A schema of this type should only be used for applications that need a high level of sophistication.
- For each star schema or snowflake schema it is possible to construct a fact constellation schema.
- That solution is very flexible, however it may be hard to manage and support.
- The main disadvantage of the fact constellation schema is a more complicated design because many variants of aggregation must be considered.
- In a fact constellation schema, different fact tables are explicitly assigned to the dimensions, which are for given facts relevant.
- This may be useful in cases when some facts are associated with a given dimension level and other facts with a deeper dimension level.
- Use of that model should be reasonable when for example, there is a sales fact table (with details down to the exact date and invoice header id) and a fact table with sales forecast which is calculated based on month, client id and product id.
- In that case using two different fact tables on a different level of grouping is realized through a fact constellation model.

Family of stars

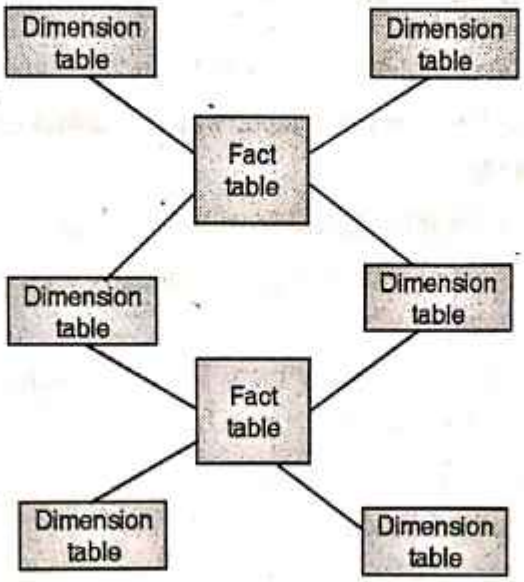


Fig. 1.13.1 : Family of stars

Scanned by CamScanner

Example :

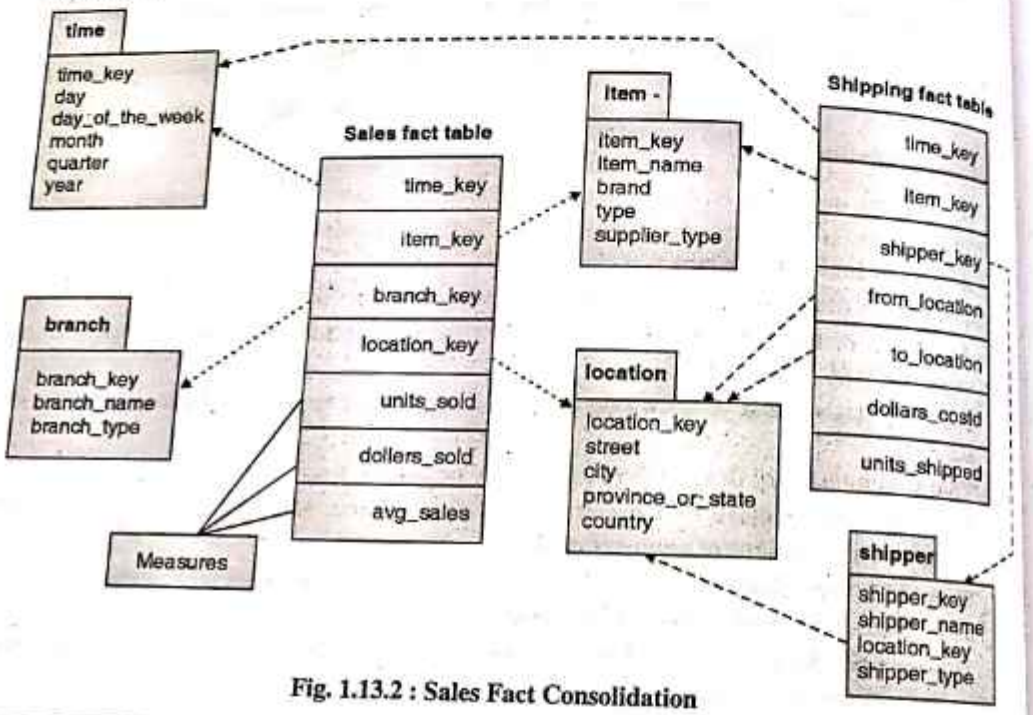


Fig. 1.13.2 : Sales Fact Consolidation

Syllabus Topic : Factless Fact Tables

1.14 Factless Fact Tables

1.14.1 Fact Tables and Dimension Tables

- A dimensional model consists of Fact tables and dimension tables.
- Each dimensional model has a primary table which is a fact table that is meant to contain the business measurements.
- Numeric and additive are the most useful facts.
- A fact table has many to many relationships; it contains a set of two or more foreign keys that join to a dimension table.
- A fact in the fact table depends on many facts. For eg. In sales schema, sales_amount fact depends on Product, Location and time. These factors are called as dimensions.
- Dimensions are factors on which a given fact depends. The sales_amount fact can also be thought of as a function of three variables.

$$Sales_amount = (product, location, time)$$

- Likewise in a sales fact table we
- Dimension tables are companion
- Each dimension table is defined integrity with any given fact ta textual information.
- To understand the concepts of following scenario :
- Imagine standing in the marke down the quantity sold and the
- Note that a measurement need product, and store).
- The information gathered can

- The facts are **Sales_Unit,** additive), which depend o dimensions are stored in di

1.14.2 Factless Fact Tab

- Factless table means only
- Used only to put relation b
- Are useful to describe something has/has not hap
- Often used to represent m
- The only thing they cont event which is identifie tables.

Scanned by CamScanner

- Likewise in a sales fact table we may include other facts like sales_unit and cost.
- Dimension tables are companion tables to a fact table in a star schema.
- Each dimension table is defined by its primary key that serves as the basis for referential integrity with any given fact table to which it is joined. Most dimension tables contain textual information.
- To understand the concepts of facts, dimension, and star schema, let us consider the following scenario :
- Imagine standing in the marketplace and watching the products being sold and writing down the quantity sold and the sales amount each day for each product in each store.
- Note that a measurement needs to be taken at every intersection of all dimensions (day, product, and store).
- The information gathered can be stored in the fact table as shown in Table 1.14.1 :

Table 1.14.1 : Fact Table

Sales Fact Table
Date Key
Product Key
Store Key
Sales_Unit
Sales_Amount
Cost

- The facts are **Sales_Unit, Sales_Amount,** and **Cost** (note that all are numeric and additive), which depend on dimensions **Date, Product**, and **Store**. The details of the dimensions are stored in dimension tables.

1.14.2 Factless Fact Table

➔ **(MU - Dec. 2012, May 2013)**

- Factless table means only the key available in the Fact there is no measures available.
- Used only to put relation between the elements of various dimensions.
- Are useful to describe events and coverage, i.e. the tables contain information that something has/has not happened.
- Often used to represent many-to-many relationships.
- The only thing they contain is a concatenated key, they do still however represent a focal event which is identified by the combination of conditions referenced in the dimension tables.

Scanned by CamScanner

of

- An Example of Factless fact table can be seen in the Fig. 1.14.1.

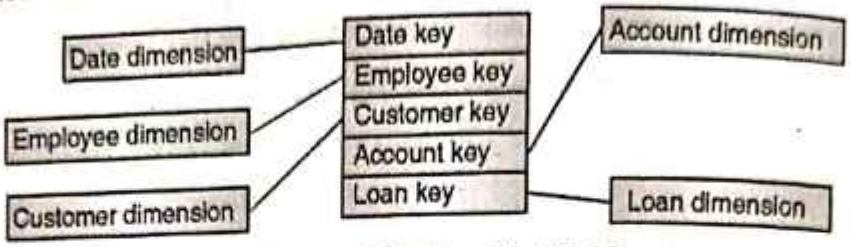


Fig. 1.14.1 : A Factless Fact Table

- There are two main types of factless fact tables :

1. Event tracking tables

- Use a factless fact table to track events of interest to the organization. For example attendance at a cultural event can be tracked by creating a fact table containing the following foreign keys (i.e. links to dimension tables) : event identifier, speaker/entertainer identifier, participant identifier, event type, date.This table can then be queried to find out information, such as which cultural events or event types are the most popular.
- Following example shows factless fact table which records every time a student attends a course or Which class has the maximum attendance? Or What is the average number of attendance of a given course?
- All the queries are based on the COUNT() with the GROUP BY queries. So we can first count and then apply other aggregate functions such as AVERAGE, MAX, MIN.

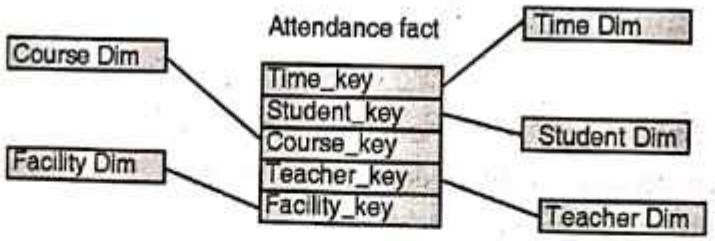


Fig. 1.14.2 : Example of Event Tracking Tables

2. Coverage Tables

The other type of factless fact table is called Coverage table by Ralph. It is used to support negative analysis report. For example a Store that did not sell a product for a given period. To produce such report, you need to have a fact table to capture all the possible combinations. You can then figure out what is missing.

Common examples of factless fact table :

- Ex-Visitors to the office.
- List of people for the web click.
- Tracking student attendance or registration events.

Scanned by CamScanner

1.15 Update to the Dimension Tables

→ (MU - May 2016)

- Every day more and more sales take place, so more and more rows are added to the fact table.
- Updation due to change in fact table happens very rarely.
- Dimension tables are more stable as compared to the fact tables.
- Dimension table changes due to the change in attributes themselves but not because of increase in number of rows.

1.15.1 Slowly Changing Dimensions

- Dimensions are generally constant over time but if not constant then changes slowly.
- The customer ID of the record remain same but the marital status or location of customer may change over time.
- In OLTP system, whenever such change in attribute value happens, the old values replace the new value by overwriting old ones.
- But in data warehouse, overwriting of attributes is not the solution as historical data for analysis is required always.
- So making such changes in attributes have 3 different types :

(i) Type 1 Changes,
(ii) Type 2 Changes,
(iii) Type 3 Changes

(i) Type 1 Changes

Principles of Type 1 Changes

- Type 1 changes are related to the correction of errors in source systems.
- Example, In customer dimension if name changes from Anand Prasad - >Anandraj Prasad
- Sometimes changes has no importance, so old values can be overwritten.
- Changes need not be preserved in DW.
- How to apply Type 1 Changes to the Data Warehouse :
- Replace the old value with new value by overwriting the attribute value in the dimension table row.
- No need to preserve the old value.

Scanned by CamScanner

evised Syllabus (Rev - 2016) of
ity
3-2019
redit and Grading System)

- No need to do other changes in dimension table row.
- The key values are not affected for dimension table.
- Type 1 changes are simple and easy to implement.

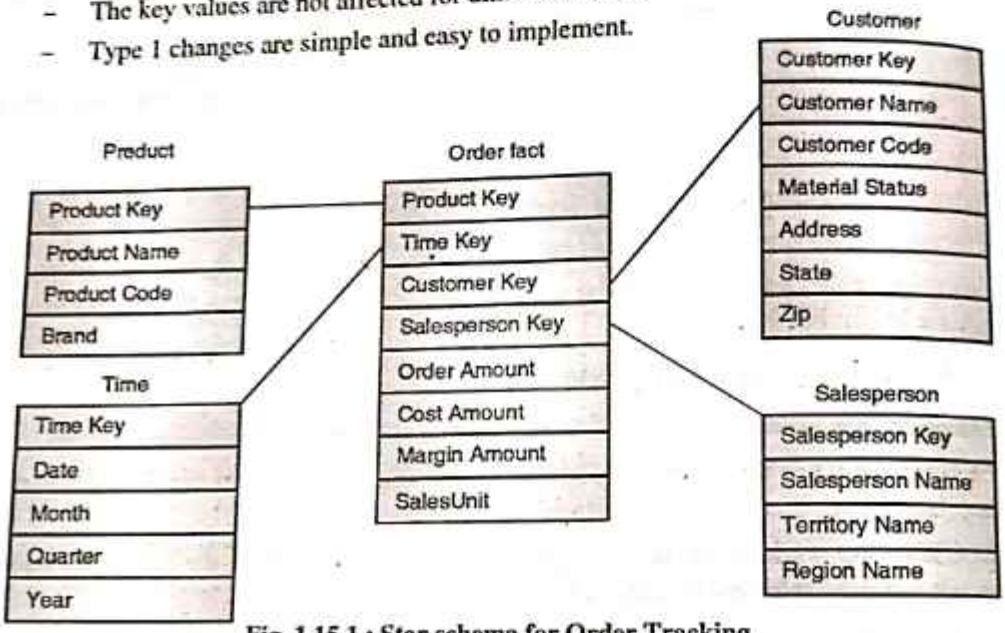


Fig. 1.15.1 : Star schema for Order Tracking

- From Fig. 1.15.1 star Schema, If Customer Name of Customer Dimension or Product Name of Product dimension changes, then it comes under Type 1 change as there is no need to preserve the old values. It can be overwritten with new names without changing any Key Value.

(ii) Type 2 Changes

- Check the attribute Marital Status of Customer Dimension.
- This attribute value can be changed from Single to Married over time.
- *If the Marital status of Anand Prasad Changes from Single to Married, then old value can not be overwritten as there may be a need to track orders by marital Status.*
- *If Anand Prasad Changes his state from Maharashtra to Gujarat, then also old value of state has to be preserved .*
- *This type of changes are called Type 2 Changes where historical Values are preserved instead of overwriting .*

General principles for type 2 change

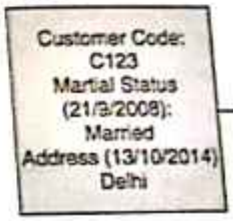
- *Type 2 changes are related to the true changes in source systems.*
- *History of data must be preserved in the data warehouse.*

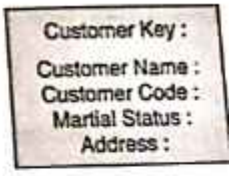
- This type of change part
- If any change for same

How to apply Type 2 Cha

- Add a new tuple in dim
- Include date field on w
- Original row should n
- The new row is insert

Example :





(iii) Type 3 Changes

- Type 3 changes
- Complex queri
  - Hard to im
  - Time-con
  - Hard to m
- In Type 2 cha that cut-off "Married" af groups.
- In case, if th be handled attribute for

Scanned by CamScanner

- This type of change partitions the history in the data warehouse.
- If any change for same attribute, then it must be preserved.

How to apply Type 2 Changes to the Data Warehouse ?

- Add a new tuple in dimension table with new value of the attribute.
- Include date field on which the change occurred in dimension table.
- Original row should not be changed and the key value of that row is not affected.
- The new row is inserted with a new surrogate key.

Example :

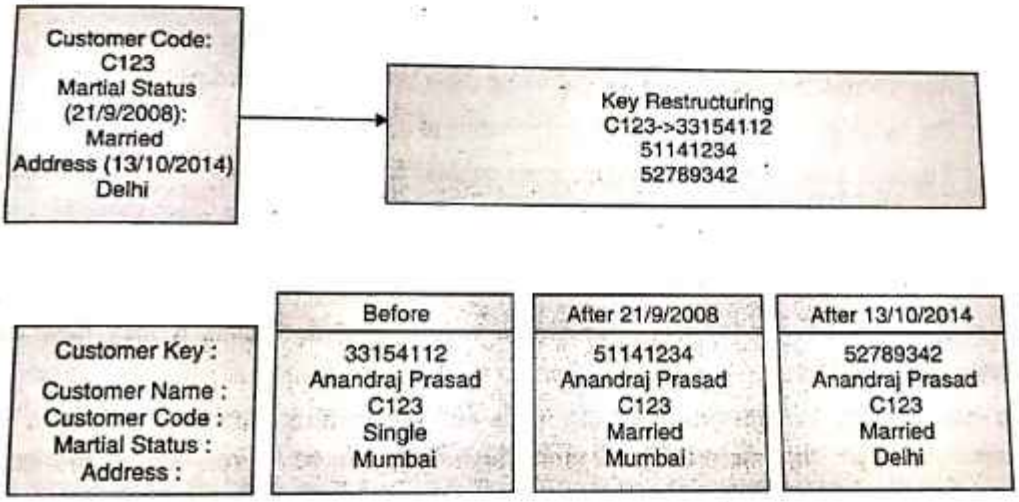


Fig. 1.15.2

(iii) Type 3 Changes

- Type 3 changes are not common at all
- Complex queries on type 2 changes may be :
  - Hard to implement
  - Time-consuming
  - Hard to maintain
- In Type 2 changes, date is a cut-off point. If someone track the orders before or after that cut-off date , then customer either fall in "single" group prior to that date and "Married" after that date. But for any period of time, it cannot be counted in both groups.
- In case, if the requirement is to count the orders on or after cut-off date then it cannot be handled as a type 2 change. There is a need to track both old and new value of attribute for a certain period. This type of change is called Type 3 change.

Scanned by CamScanner

General Principles of Type 3 Change

- Type 3 change usually consists of "soft" or tentative changes in the source system.
- The change in the attribute value needs to be preserved, so a history needs to be maintained to keep a track of the old and new values.
- They can be used for performance comparison across the transition.
- With type 3 change its possible to track forward and backward.

How to apply Type 3 Changes to the Data Warehouse

- No new dimension row is needed.
- The existing queries will seamlessly switch to the current value.
- Any queries that need to use the old value must be revised accordingly.
- The technique works best for one soft change at a time.
- If there is a succession of changes, more sophisticated techniques must be advised.

1.15.2 Large Dimension Tables

- Large dimension tables are very deep and wide.
- Deep means it has a very large number of rows and wide means it may have many attributes or columns.
- To handle large dimensions, one can take out some mini dimensions from a large dimension as per the interest. These mini dimensions can be represented in the form of star schema.
- For example, the above mentioned order analysis star schema is one of the mini dimension of manufacturing company in which the marketing department of the company is interested.
- Customer and Product are generally large dimensions.
- Large dimensions are generally slow and inefficient due to its size.
- They tend to have multiple hierarchies to perform various OLAP operations like drill down or roll up.

1.15.3 Rapidly Changing or Large Slowly Changing Dimensions

- In type 2 changes, new row is created with the new value of changed attribute. This preserves the history or old values of attributes.
- If again there is a change in same attribute, then again a new dimension table row is created with new value.

Scanned by CamScanner

- This is feasible if the dimension changes infrequently like once or twice a year. For example, Product dimension which has rows in thousands changes rarely so it is manageable.
- But in case of customer dimensions, where number of rows are millions and changes infrequently, then type 2 changes are feasible and not much difficult. If customer dimension changes rapidly then Type 2 changes are problematic and difficult.
- If dimension table is rapidly changing and large then break that dimension table into one or more smaller dimension tables.

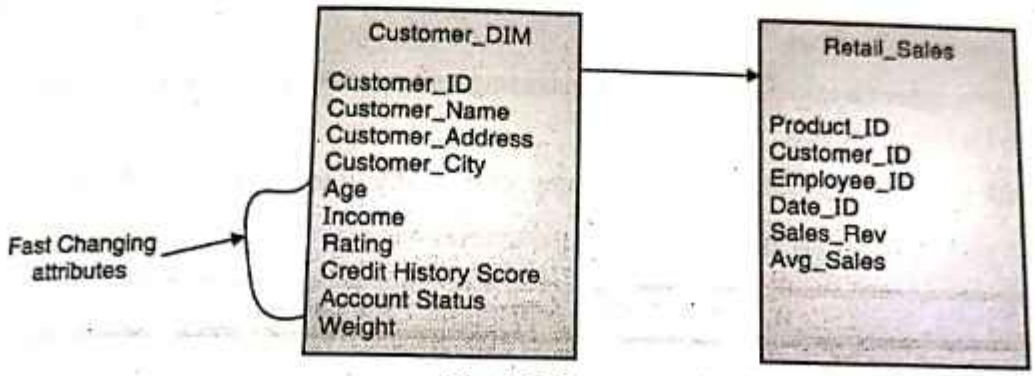


Fig. 1.15.3

- Move the rapidly changing attribute in another dimension table and leave the original dimension table with slowly changing attributes.
- Example, Consider Customer dimension, then following attributes changes rapidly Age, Income, Rating, Credit history score, Customer account status, Weight.
- From Customer_Dim, take out the fast changing attribute and create new mini dimension as shown in Fig. 1.15.4.

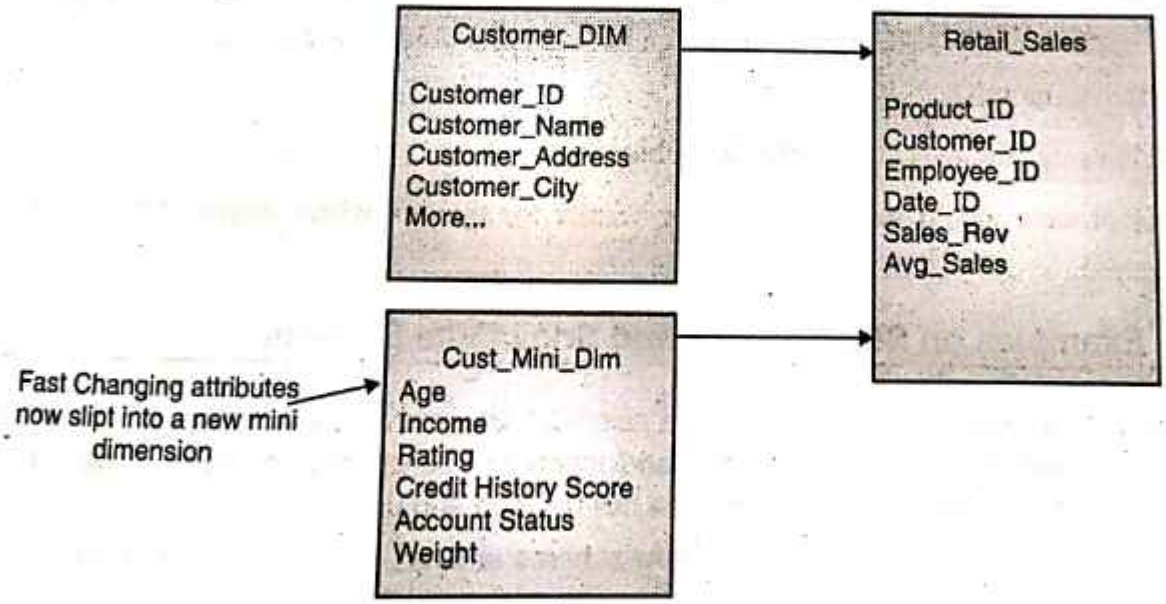


Fig. 1.15.4

Scanned by CamScanner

1.15.4 Junk Dimensions

- From source legacy systems, some textual data or flags cannot be the significant fields in major dimensions. But at the same time, it cannot be discarded. There are some options which can be used to consider such fields.
- Don't consider such texts and flags. Discard all while creating major dimensions. But by doing this you may be losing some useful information.
- Keep such texts and flags as it is in fact table. But due to this size of fact table increases without any advantage.
- Create a separate dimension table for each flag and text. With this, number of dimension table increases greatly.
- Create a single "Junk" dimension and keep all meaningful texts and flags into it.
- Such junk dimensions are useful to fire the queries based on flag or text values.

Syllabus Topic : Aggregate Fact Tables

1.16 Aggregate Fact Tables

→ **(MU - May 2014)**

- A fact table contains measurements or so called as facts.
- For e.g In a Sales data warehouse, a sales fact table contains units sold for a transaction as one of the fact.
- Addition of some level of aggregation to the fact table, For e.g. we could have aggregated the sales fact table by month. The number of units sold to a particular customer in a particular month.
- Addition of summary data to the fact table.
- This process would help in retrieving results for queries where aggregation would be required.

1.17 Examples on Star Schema and Snowflake Schema

Ex. 1.17.1 : All electronics company have sales department. Sales consider four dimensions namely time, item, branch and location. The schema contains a central fact table sales with two measures dollars_sold and unit_sold.

Design star schema, snowflake schema and fact constellation for same.

Scanned by CamScanner

Soln. :

(a) **Star Schema**

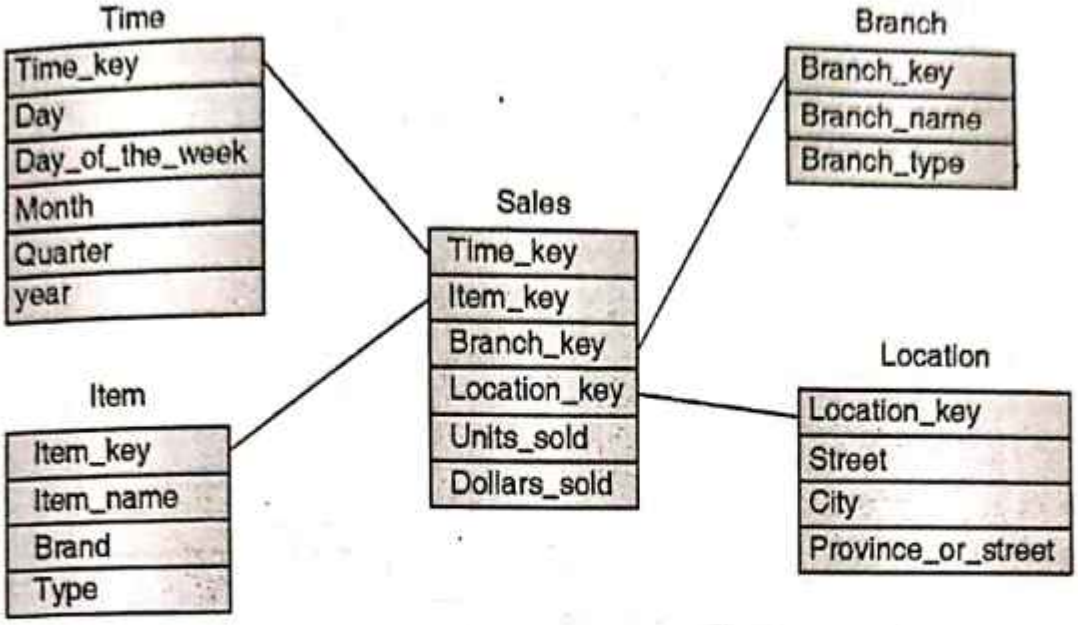


Fig. P. 1.17.1 : Sales Star Schema

(b) **Snowflake Schema**

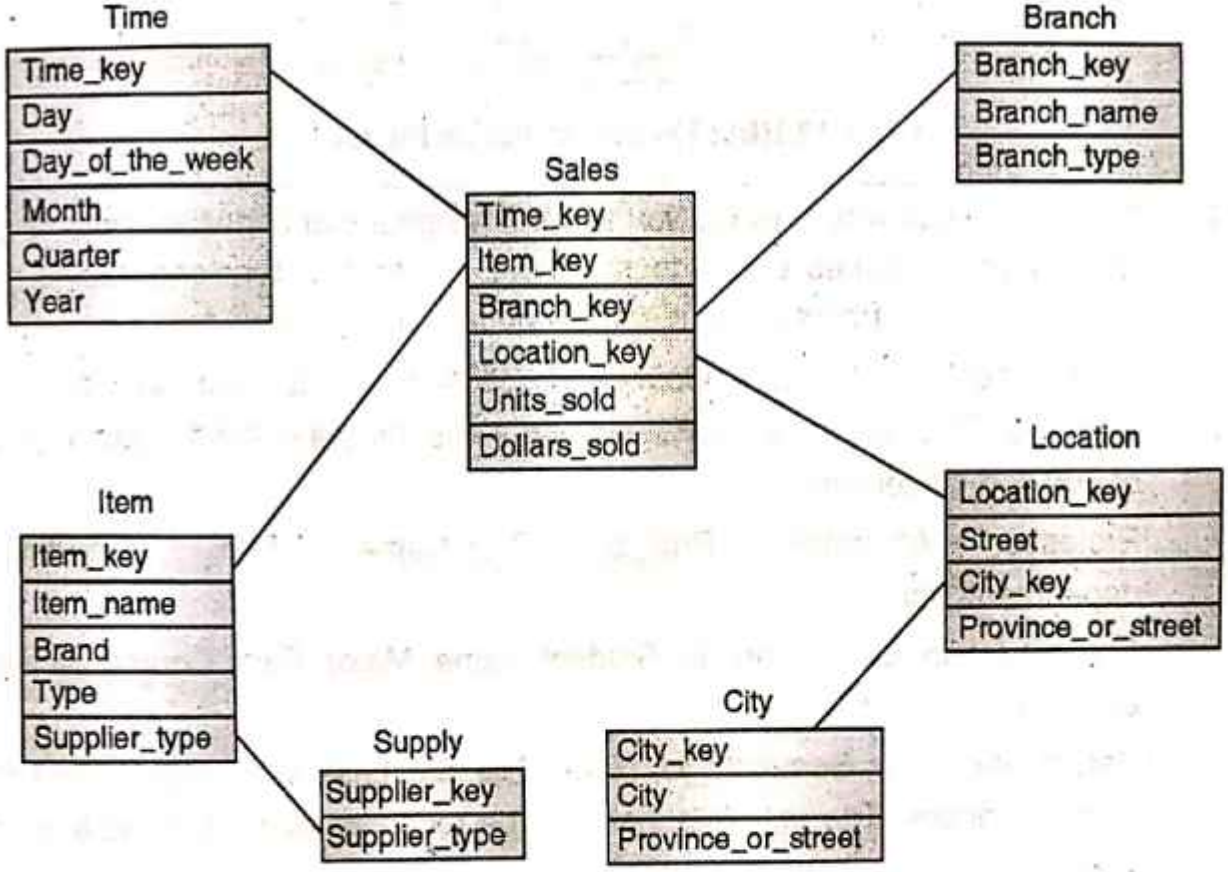


Fig. P. 1.17.1(a) : Sales Snowflake Schema

Scanned by CamScanner

(c) Fact Constellation

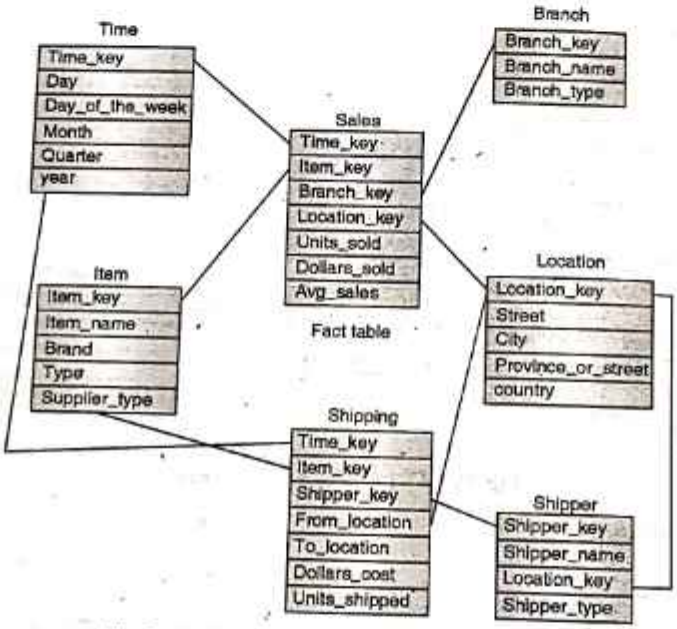


Fig. P. 1.17.1(b) : Fact constellation for sales

Ex. 1.17.2 : The Mumbai university wants you to help design a star schema to record grades for course completed by students. There are four dimensional tables namely course_section, professor, student, period with attributes as follows :

Course_section Attributes: Course_Id, Section_number, Course_name, Units, Room_id, Roomcapacity. During a given semester the college offers an average of 500 course sections

Professor Attributes: Prof_id, Prof_Name, Title, Department_id, department_name

Student Attributes: Student_id, Student_name, Major. Each Course section has an average of 60 students

Period Attributes: Semester_id, Year. The database will contain Data for 30 months periods. The only fact that is to be recorded in the fact table is course Grade

Answer the following Questions

(a) Design the star schema for this problem.

(b) Estimate the number of rows in the fact table, using the assumptions stated above and also estimate the total size of the fact table (in bytes) assuming that each field has an average of 5 bytes.

Scanned by CamScanner

(c) Can you convert this star schema to a snowflake schema ? Justify your answer and design a snowflake schema if it is possible.

Soln. :

(a) Star Schema

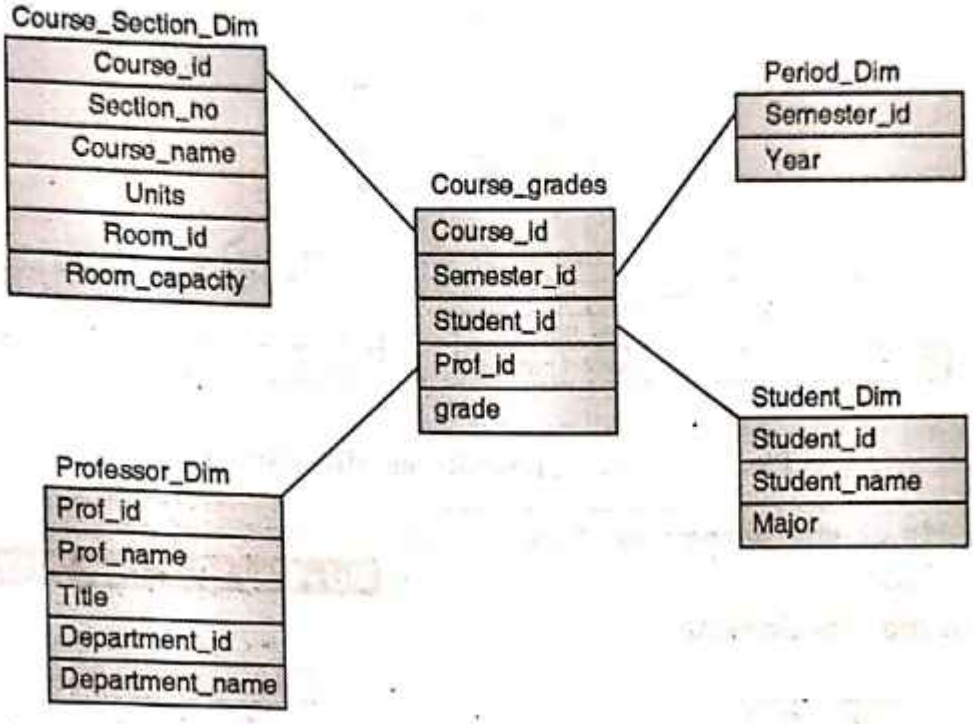


Fig. P. 1.17.2 : University Star Schema

(b) Total Courses Conducted by university = 500

Each Course has average students = 60

University stores data for 30 months

Total Student in University for all courses in 30 months = 500*60 = 30000

Time Dimension = 30 months = 5 Semesters (Assume 1 semester = 6 months)

Now, Number of rows of fact table = 30000*5 = 150000 (one student has 5 grades for 5 semesters)

(c) Snowflake Schema

- Yes, the above star schema can be converted to a snowflake schema, considering the following assumptions.
- Courses are conducted in different rooms, so course dimension can be further normalized to rooms dimension as shown in the Fig. P. 1.17.2(a).
- Professor belongs to a department, and department dimension is not added in the star schema, so professor dimension can be further normalized to department dimension.
- Similarly students can have different major subjects, so it can also be normalized as shown in the Fig. P. 1.17.2(a).

Scanned by CamScanner

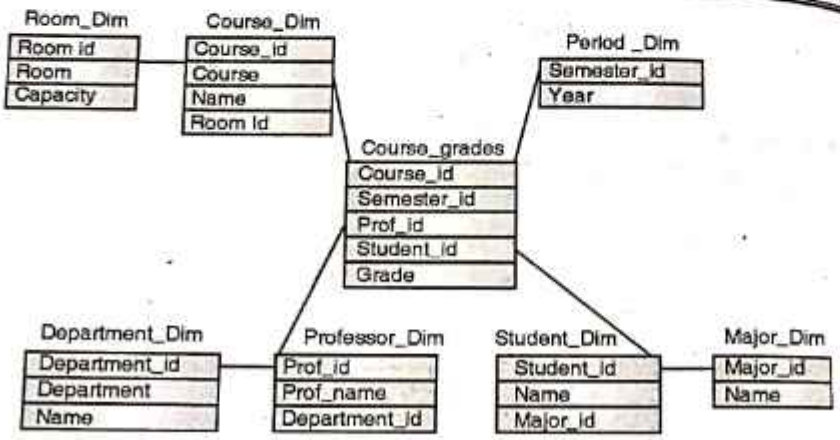


Fig. P. 1.17.2(a) : University Snowflake Schema

Ex. 1.17.3 : **Draw star schema** for "Hotel Occupancy" considering dimensions like Time, Hotel etc. **MU - May 2012, Dec. 2013, 10 Marks**

Soln. : Draw the Star Schema

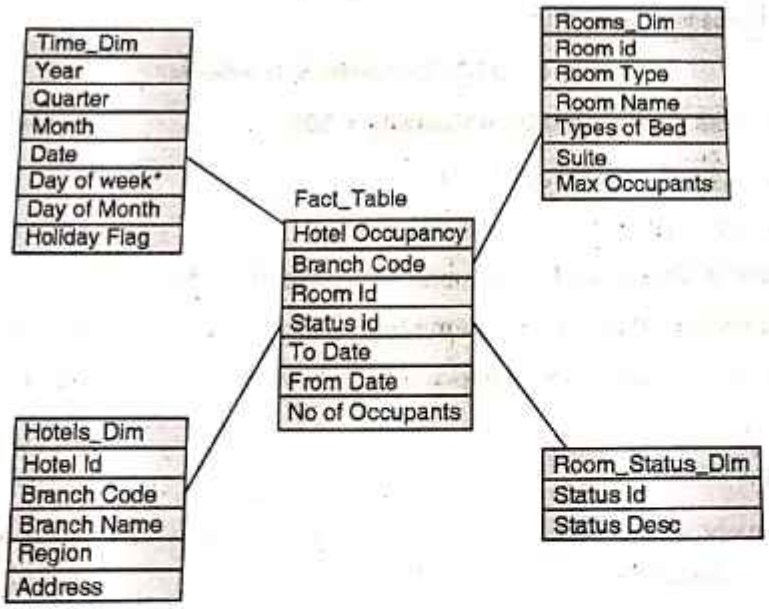


Fig. P. 1.17.3 : Hotel Occupancy Star Schema

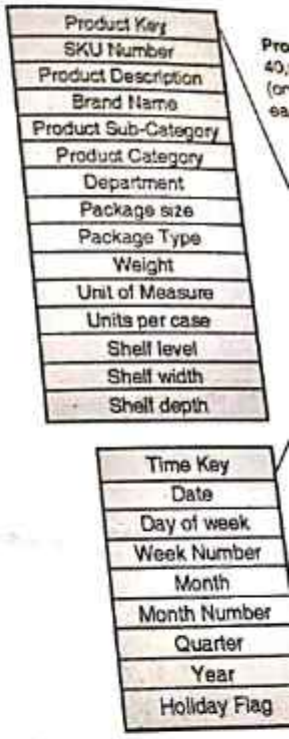
Ex. 1.17.4 : For a Supermarket Chain consider the following dimensions, namely Product, store, time , promotion. The schema contains a central fact tables sales facts with three measures unit_sales, dollars_sales and dollar_cost.

Design star schema and calculate the maximum number of base fact table records for the values given below :

Time period : 5 years
Store : 300 stores reporting
Product : 40,000 products
Promotion : a sold item ma

Soln. :

(a) Star schema :



(b) Time period = 5 years

There are 300 stores,

Each stores daily sale

Promotion = 1

Maximum number o

Ex. 1.17.5 : Draw a S

Scanned by CamScanner

Time period : 5 years

Store : 300 stores reporting daily sales

Product : 40,000 products in each store(about 4000 sell in each store daily)

Promotion : a sold item may be in only one promotion in a store on a given day

MU - May 2013, May 2016, 10 Marks, Dec. 2014, 5 Marks

Soln. :

(a) Star schema :

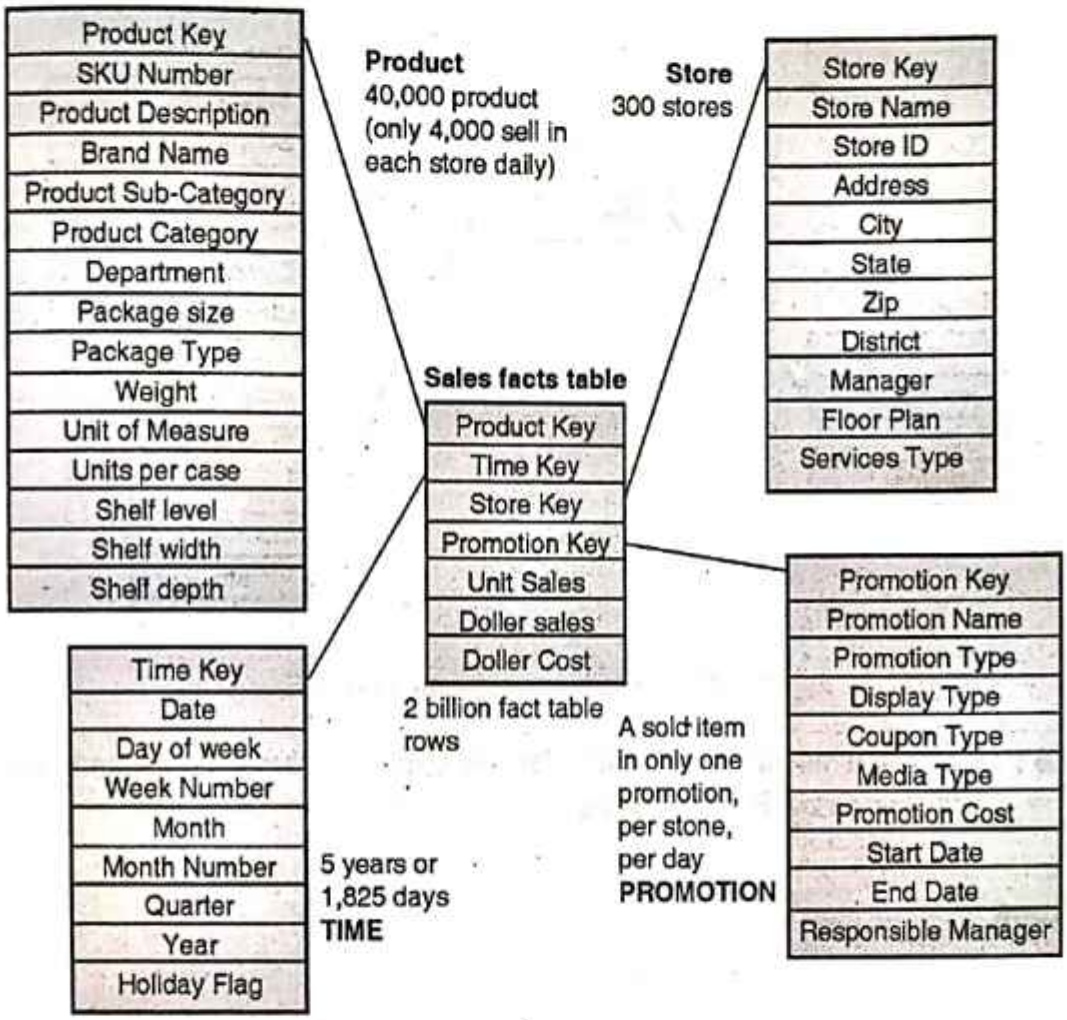


Fig. P. 1.17.4 : Sales Promotion Star Schema

(b) Time period = 5 years × 365 days = 1825

There are 300 stores,

Each stores daily sale = 4000

Promotion = 1

Maximum number of fact table records: 1825 × 300 × 4000 × 1= 2 billion

Ex. 1.17.5 : Draw a Star Schema for Student academic fact database.

Scanned by CamScanner

Soln. :

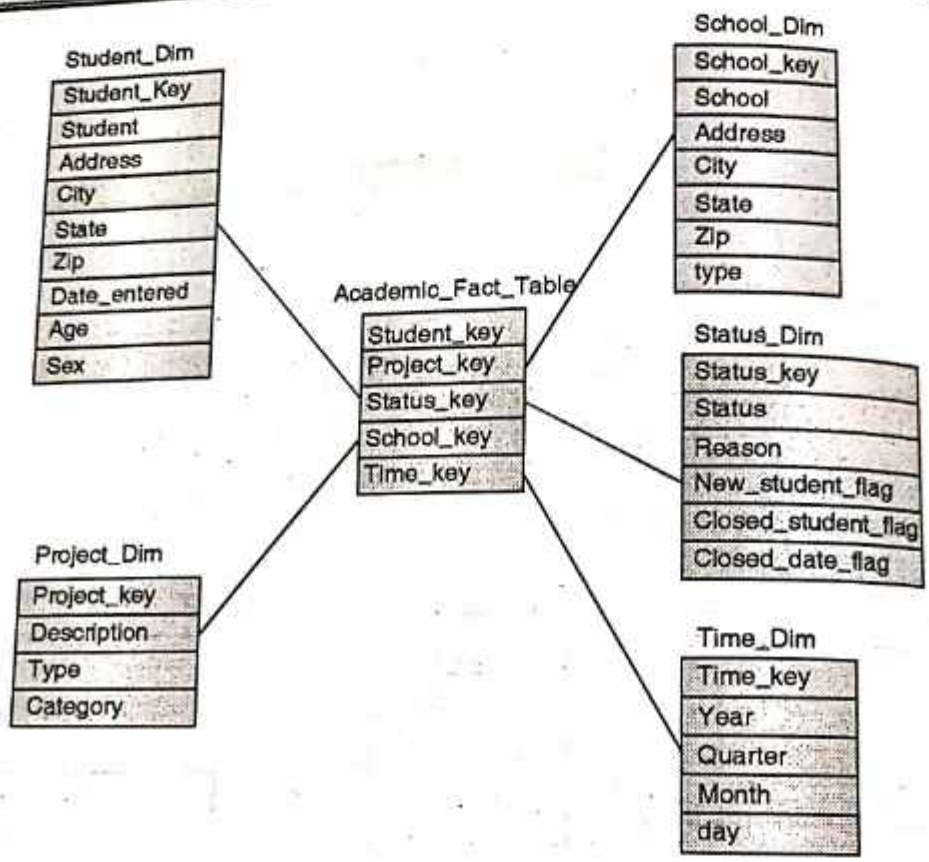


Fig. P. 1.17.5 : Student Academic Star Schema

Ex. 1.17.6 : List the dimensions and facts for the Clinical Information System and Desig Star and Snow Flake Schema.

Soln. :

Dimensions

1. Patient
2. Doctor
3. Procedure
4. Diagnose
5. Date of Service
6. Location
7. Provider

Facts

1. Adjustment
2. Charge
3. Age

Scanned by CamScanner

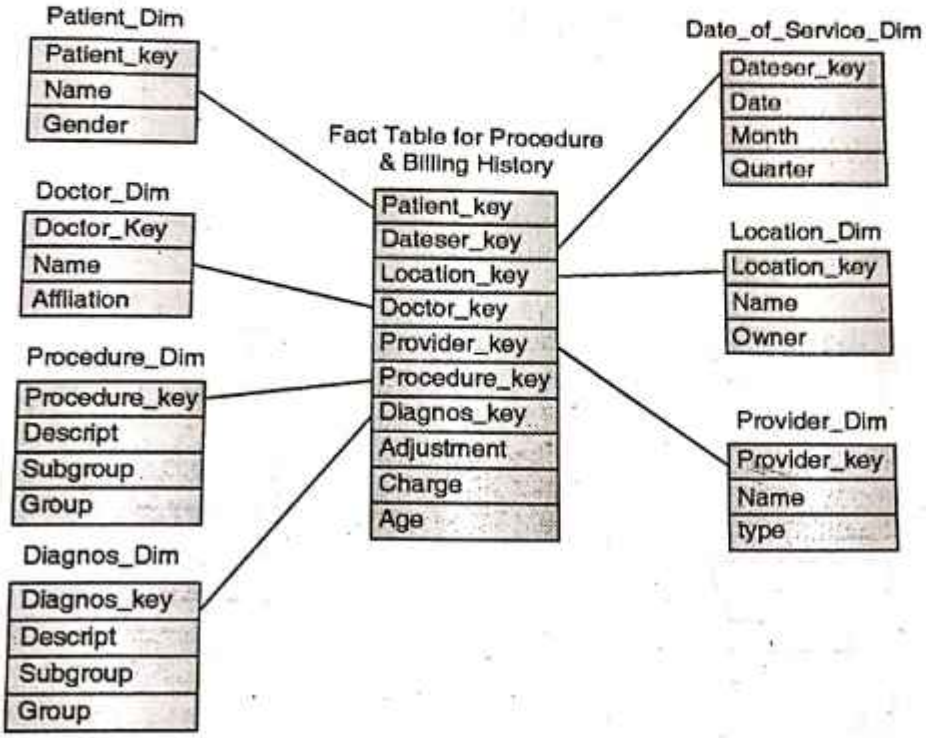


Fig. P. 1.17.6 : Clinical Information Star Schema

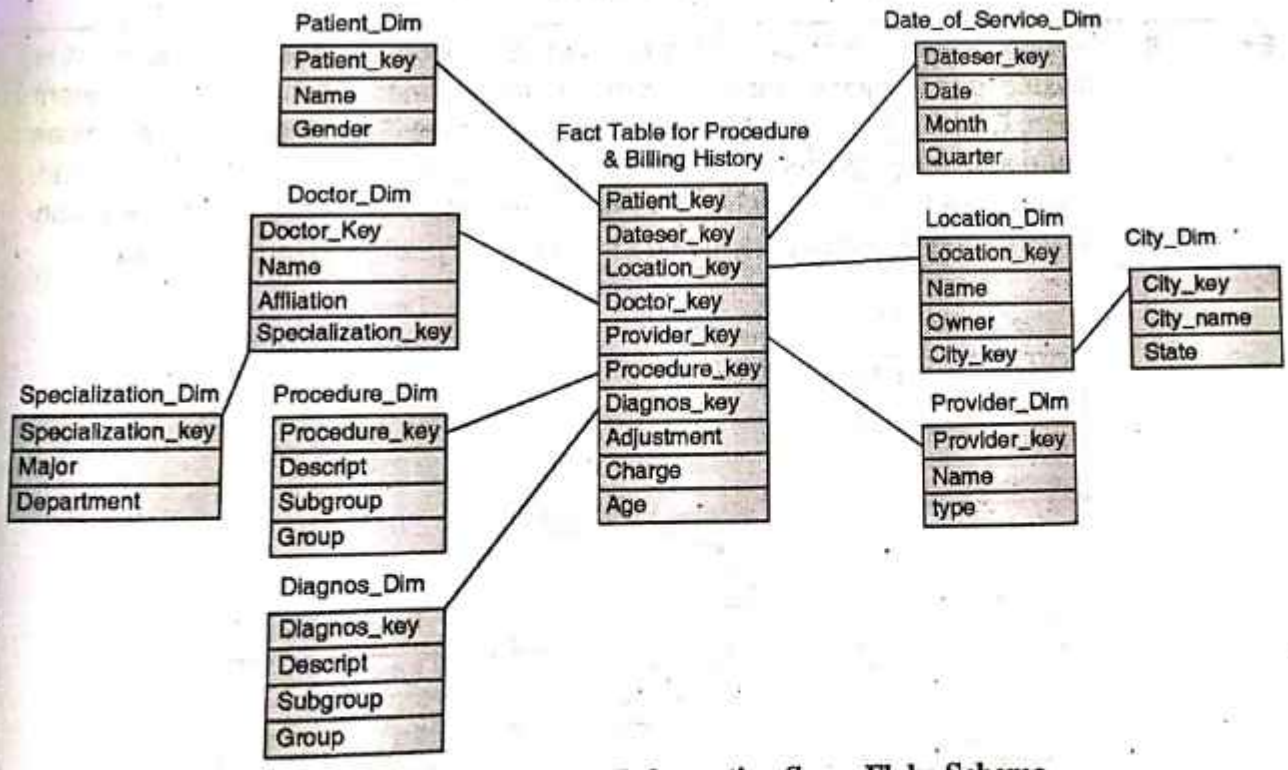


Fig. P. 1.17.6(a) : Clinical Information Snow Flake Schema

Scanned by CamScanner

Ex. 1.17.7 : Draw a Star Schema for Library Management.

Soln. :

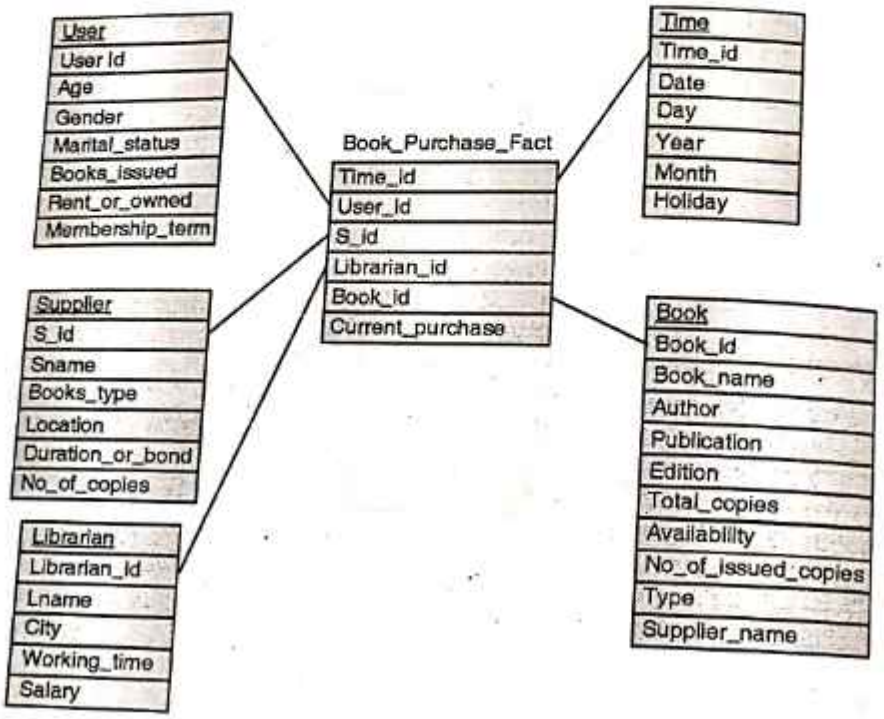


Fig. P. 1.17.7 : Star schema for Library

Ex. 1.17.8 : A manufacturing company has a huge sales network. To control the sales, it is divided in the regions. Each region has multiple zones. Each zone has different cities. Each sales person is allocated different cities. The object is to track sales figure at different granularity levels of region. Also to count no. Of products sold. Create data warehouse schema to take into consideration of above granularity levels for region, sales person and the quarterly, yearly and monthly sales.

Soln. :

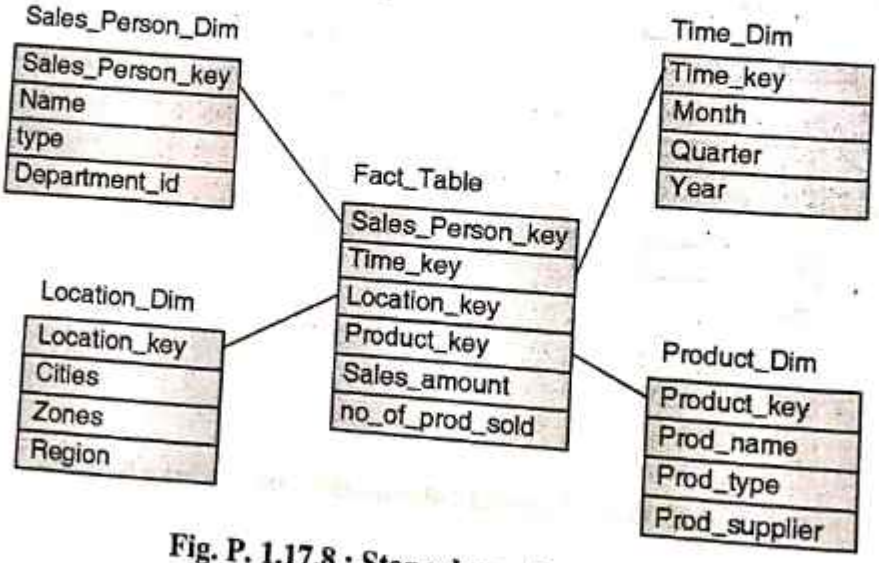


Fig. P. 1.17.8 : Star schema for Sales

Scanned by CamScanner

Ex. 1.17.9 : A bank wants to develop a data warehouse for effective decision –making about their loan schemes. The bank provides loans to customers for various purposes like House Building loan, car loan, educational loan, personal loan etc. The whole country is categorized into a number of regions, namely, North, South, East, West. Each region consists of a set of states; loan is disbursed to customers at interest rates that change from time to time. Also, at any given point of time, the different types of loans have different rates. That data warehouse should record an entry for each disbursement of loan to customer. With respect to the above business scenario.

(i) Design an information package diagram. Clearly explain all aspects of the diagram.

(ii) Draw a star schema for the data warehouse clearly identifying the fact tables, dimension tables, their attributes and measures.

MU - May 2014, 10 Marks

Soln. :

(i)

Time	Customer	Branch	Location
Time_key	Customer_key	Branch_key	Location_key
Day	Account_number	Branch_Area	Region
Day_of_week	Account_type	Branch_home	State
Month	Loan_type		City
Quarter			Street
Year			
Holiday_flag			

(ii) Star Schema for a Bank

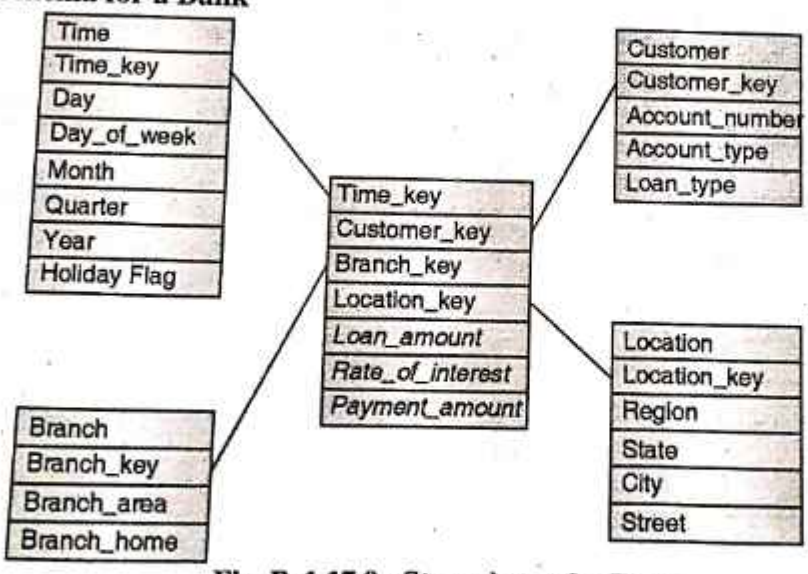
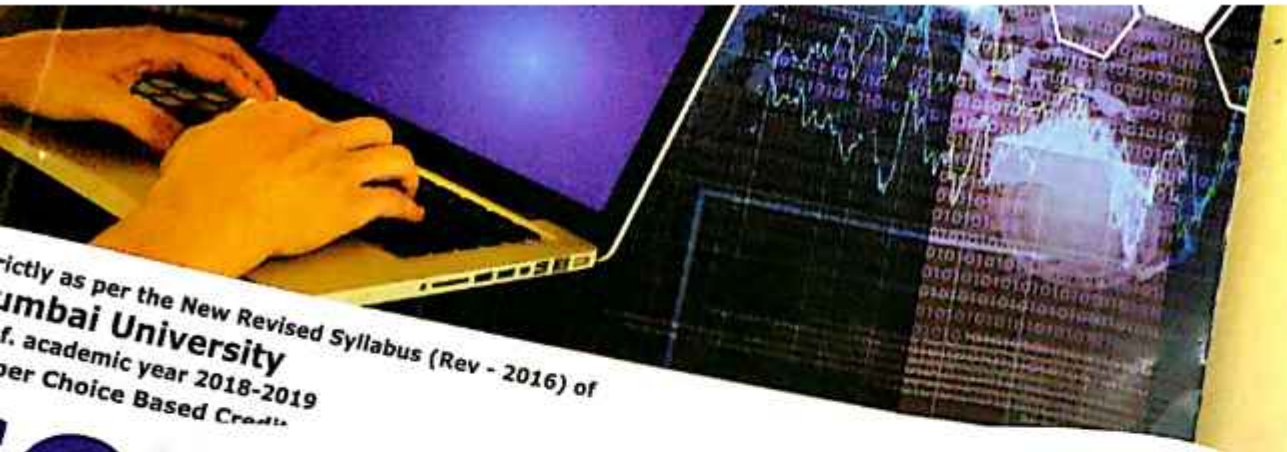


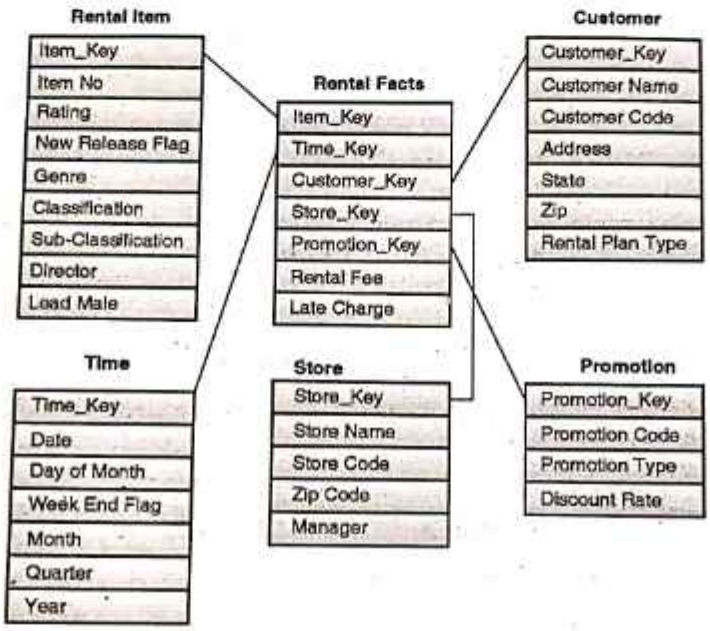
Fig. P. 1.17.9 : Star schema for Bank

Scanned by CamScanner



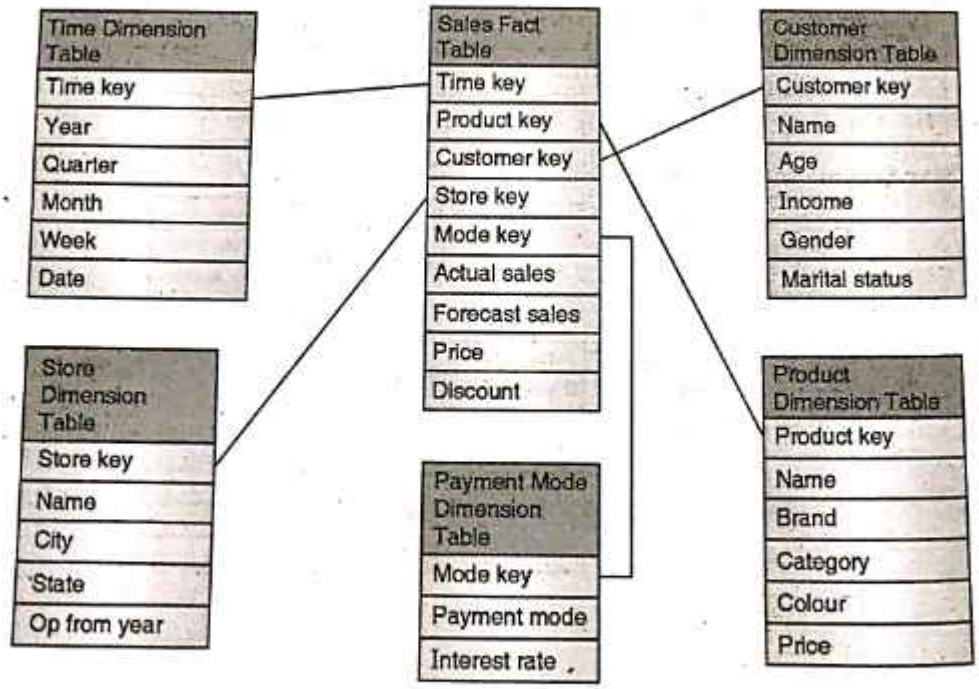
Ex. 1.17.10 : Draw star schema for video Rental

Soln. :



Ex. 1.17.11 : Star schema for retail chain.

Soln. :

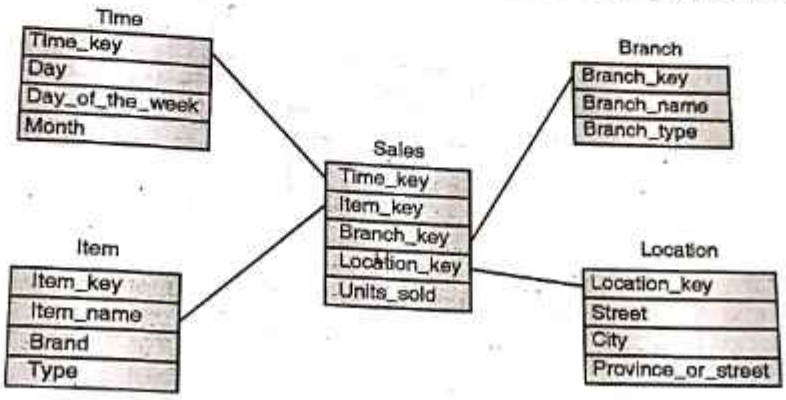


Scanned by CamScanner

Ex. 1.17.12 : Design Star schema for autosales analysis of company.

MU - Dec. 2011, 10 Marks

Soln. :



Ex. 1.17.13 : Consider the following database for a chain of bookstores.

BOOKS (Booknum, Primary_Author, Topic, Total_Stock, Price)

BOOKSTORE (Storenum, City, State, Zip, inventory_Value)

STOCK (Storenum, Booknum, QTY)

With respect to the above business scenario, answer the following questions. Clearly state any reasonable assumptions you make.

(a) Design an information package diagram.

(b) Design a star schema for the data warehouse clearly identifying the fact tables(s), Dimension table(s), their attributes and measures.

MU - Dec. 2012, 10 Marks

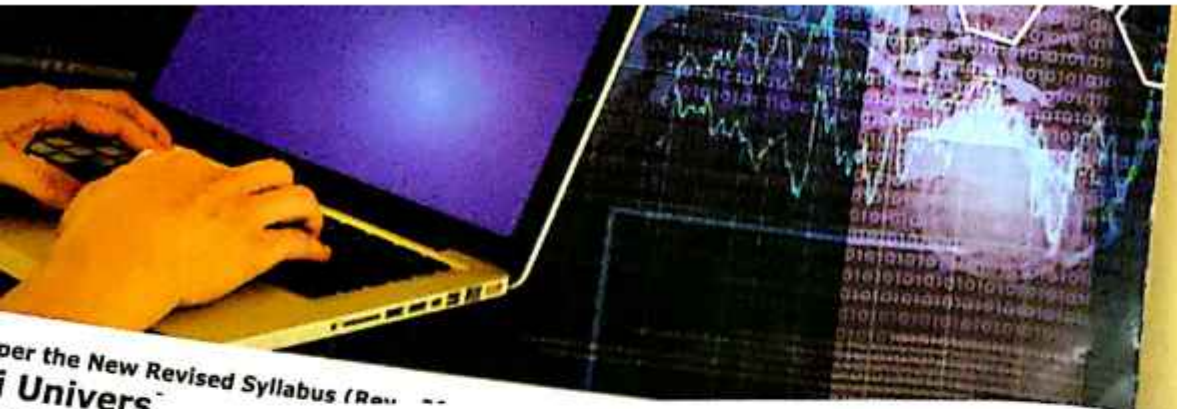
Soln. :

(a) Information Package Diagram

BOOKS	BOOKSTORE
Booknum	Storenum
Primary_author	City
Topic	State
	Zip

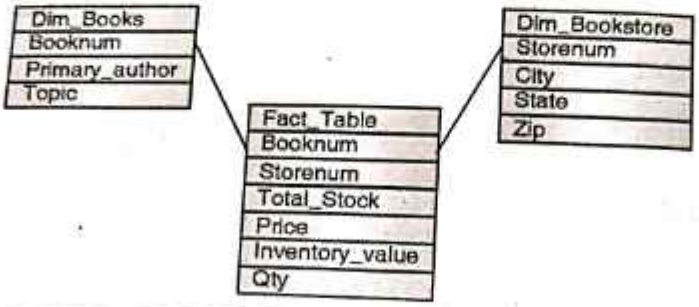
Facts : Total_Stock , Price, Inventory_value , Qty

Scanned by CamScanner



per the New Revised Syllabus (Rev...
i Univers
mic year
ice B

(b) Star Schema



Ex. 1.17.14 : One of India's large retail departmental chains, with annual revenues touching $ 2.5 billion mark and having over 3600 employees working at diverse locations, was keenly interested in a business intelligence solution that can bring clear insights on operations and performance of departmental stores across the retail chain. The company needed to support a data warehouse that exceeds daily sales data from Point of Sales (POS) across all locations, with 80 million rows and 71 columns.

(a) List the dimensions and facts for above application.

(b) Design star schema and snow flake schema for the above application.

(c) Design a BI application which will provide Retail Chain company with features and performance that meet their objectives using any data mining technique.

Soln. :

(a) Dimensions

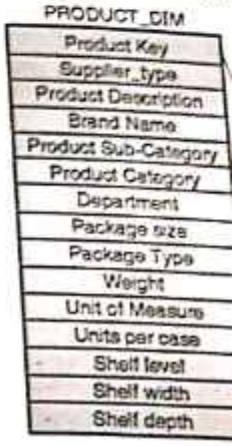
Product, Store, Time, Location .

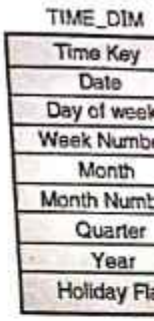
Facts :

Unit Sales, Dollar Sales, Dollar Cost

Data Warehousing & Mining (M

(b) Star Schema and snowfla





Scanned by CamScanner

(b) Star Schema and snowflake Schema

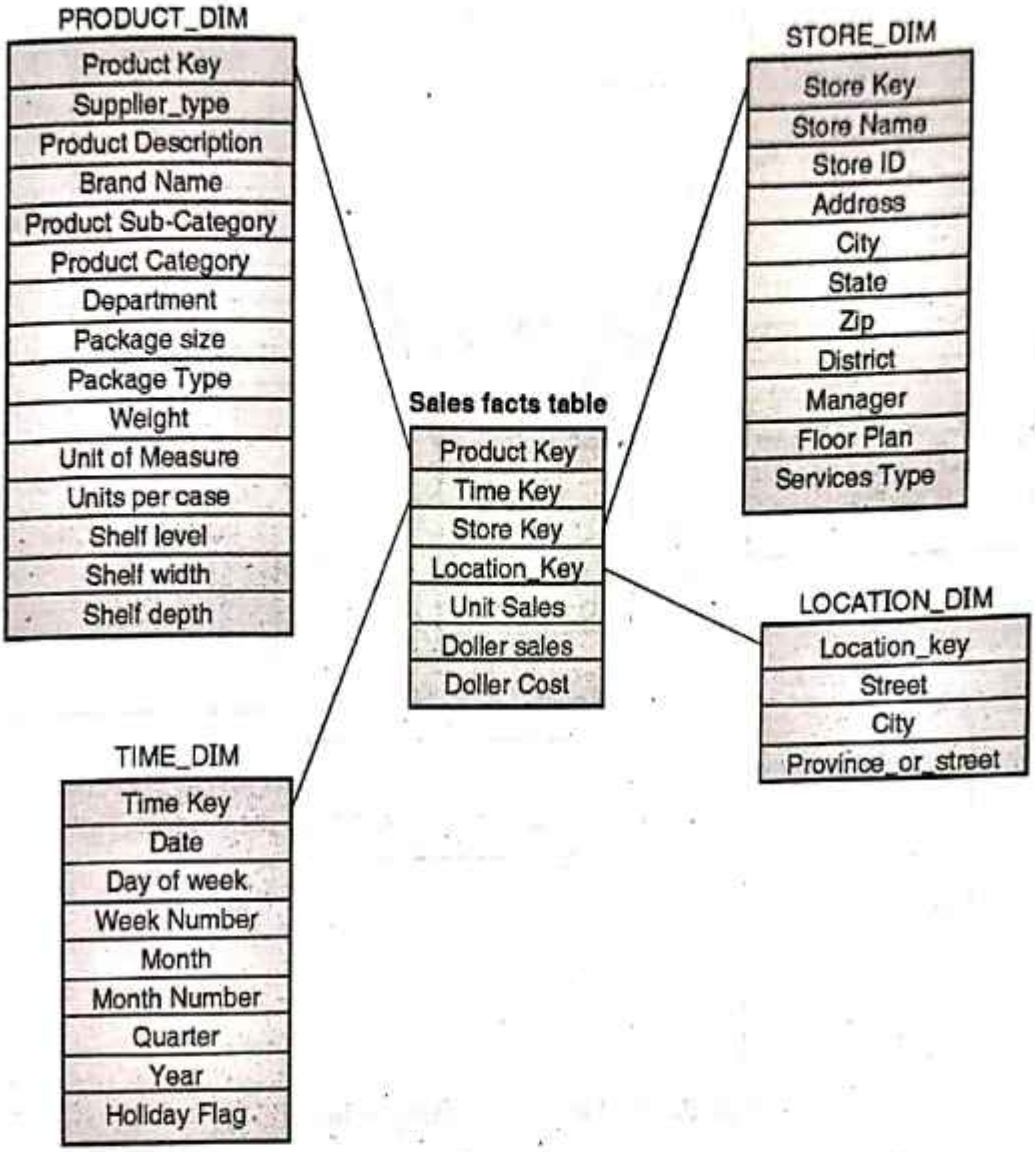


Fig. P. 1.17.14 : Star Schema

Scanned by CamScanner

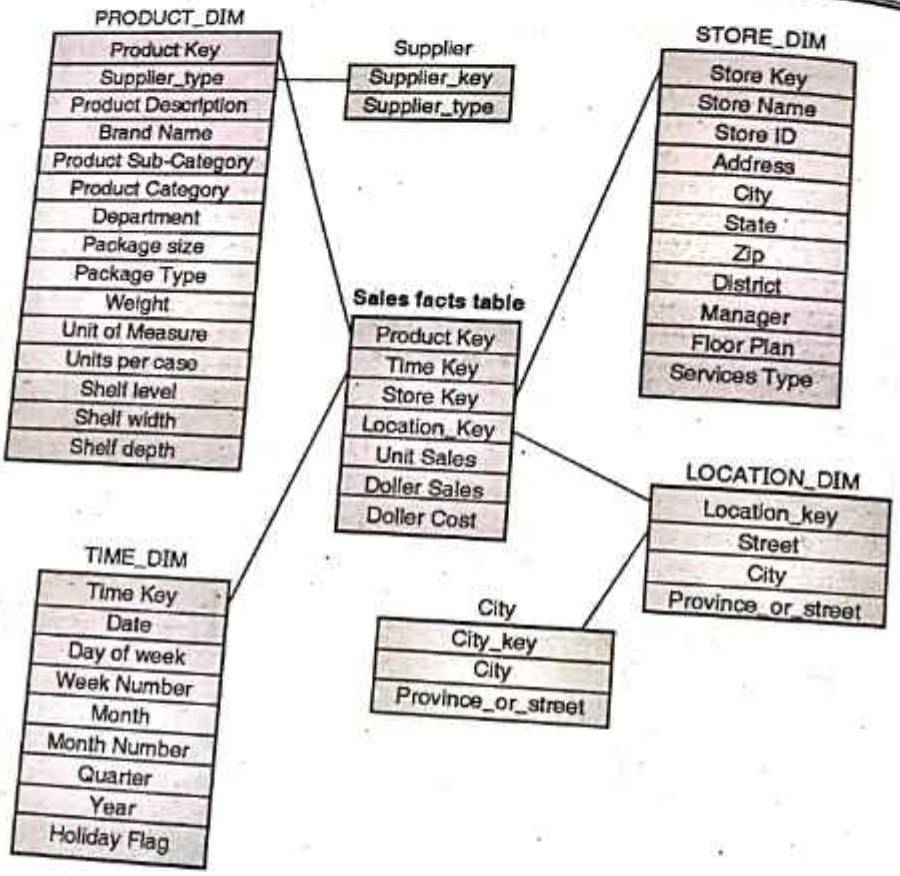


Fig. P. 1.17.14(a) : Snowflake Schema

(c) Need for BI Solution for Retail Business

- Whether shopping for daily necessities or for luxury goods, today's empowered consumers are looking for clear and compelling differences among retailers.
- This challenges for developing dynamic, customer-centric business models driven by fast, unfettered access to customer, product and other information.
- Sources throughout an enterprise and beyond will need to provide this information. And whether any particular bit of information flows from one customer's actions or from thousands of aggregated transactions, its value must be extracted and exploited as quickly as possible and at all points up and down the supply chain to have the maximum impact on your bottom line.
- To cope with these challenges, many retailers are building unified repositories of data known as data warehouses.

Scanned by CamScanner

- Linked to the right tools, and tied to the right business processes, data warehouses can drive faster, better decision making, deeper and more revealing understandings of customer behaviour, and precise-even prescient-control over every aspect of the business.
- Data warehousing can serve as the foundation for building strategic and sustainable business advantages from the point-of-sale all the way up to the executive office.

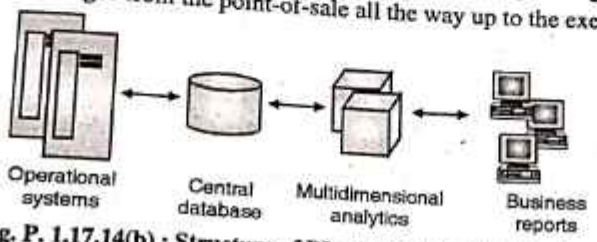


Fig. P. 1.17.14(b) : Structure of BI solution for Retail Business

Advantages of BI Solution for Retail Company

- The Retail Business Intelligence Solution has the scope and flexibility to enable better, more timely decision making at all levels of an enterprise.
- Promoting unmatched customer-centricity, and more effective and profitable cross-channel selling.
- The Solution helps you follow through on customer-centric strategies by integrating both product and customer information to yield deeper, actionable insights into not just what is selling, but to whom it is selling and why.
- Given as an input about customer, operational and competitive data, the Solution can provide real-time monitoring, localized analyses and automated exceptions management to respond to fast-changing trends and business conditions.

1.18 University Questions and Answers

May 2010

Q. 1 Define Data Warehouse. Explain the architecture of data warehouse with suitable block diagram. *(Ans. : Refer sections 1.3 and 1.6.2)* **(10 Marks)**

Q. 2 Define Metadata. What are the different types of metadata stored in a data warehouse ? Illustrate with a simple customer sales data warehouse. *(Ans. : Refer sections 1.7.1, 1.7.3 and 1.7.4)* **(10 Marks)**

Q. 3 How are top-down and bottom-up approaches for building data warehouse differ ? Discuss the merits and limitation of each approach. *(Ans. : Refer sections 1.5.1 and 1.5.2)* **(10 Marks)**

Scanned by CamScanner

Dec. 2010

Q. 4 Explain the characteristics of the data present in the data warehouse. *(Ans. : Refer section 1.3.3)* (10 Ma

Q. 5 Define Data Warehouse. Explain the architecture of data warehouse with su block diagram. *(Ans. : Refer sections 1.3 and 1.6.2)* (10 Ma

Q. 6 Distinguish between Top-Down and Bottom-Up Approach. *(Ans. : Refer sections 1.5.1 and 1.5.2)* (5 Ma

Q. 7 Write short notes on Metadata. *(Ans. : Refer section 1.7)* (5 Ma

Q. 8 A dimension table is wide; the fact table is deep. Explain. What is STAR schema a its advantages. *(Ans. : Refer section 1.10)* (10 Ma

Ans. :

Dimension Table

Dimension table has got all the detail information of their respective table f e.g., customer dimension table will contain all the related info about customers whereas fa table contains the main data, which contains the surrogate keys of every dimensions alo with other measures.

A dimension table contains a higher granular information so have less no of records and needs to have all the necessary details (More columns) related to the grain of the table. On the other side A fact table has the lowest level grain of a subject area. Lower grain cause more number of rows in the Fact table.

May 2011

Q. 9 Define a data warehouse. Explain the architecture of data warehouse with suitable block diagram. *(Ans. : Refer sections 1.3 and 1.6.2)* (10 Marks)

Q. 10 Write short notes on Top-down and Bottom-up approaches in data warehousing. *(Ans. : Refer sections 1.5.1 and 1.5.2)* (5 Marks)

Q. 11 Write short note on snowflake schema. *(Ans. : Refer section 1.12)* (5 Marks)

Dec. 2011

Q. 12 Explain Architecture of data warehouse. *(Ans. : Refer section 1.6.2)* (10 Marks)

Q. 13 Explain in detail metadata and its various types. *(Ans. : Refer sections 1.7.1 and 1.7.4)* (10 Marks)

Q. 14 Explain Snowflake Schema with example. *(Ans. : Refer section 1.12)* (10 Marks)

Q. 15 What is dimensional modeling ? Explain in detail. *(Ans. : Refer section 1.8.1)* (10 Marks)

Q. 16 Design Star Schema f *(Ans. : Refer Ex. 1.17.*

May 2012

Q. 17 Define a data wareho and hence explain its

Q. 18 Differentiate betwe warehouse. Explain *(Ans. : Refer sectio*

Q. 19 What is meant by meta data stored in *(Ans. : Refer sectio*

Q. 20 State and explain each of them. *(An*

Q. 21 Define what is m requirements for information packa schema. *(Ans. :*

Q. 22 Explain dimensio

Dec. 2012

Q. 23 What is the role *(Ans. : Refer se*

Q. 24 Consider the f BOOKS (Book BOOKSTORE STOCK (Stor With respect state any rea

(a) Desig

(b) Desig tables

(Ans.

Q. 25 Define the

(a) Fact

(b) Sno

Scanned by CamScanner

Q. 16 Design Star Schema for autosales analysis of the company. *(Ans. : Refer Ex. 1.17.12)*

(10 Marks)

May 2012

Q. 17 Define a data warehouse. Explain what is the need for developing a data warehouse and hence explain its architecture. *(Ans. : Refer sections 1.3 and 1.6.2)* **(10 Marks)**

Q. 18 Differentiate between top down and bottom-up approaches for building a data warehouse. Explain the advantages and disadvantages of each of them. *(Ans. : Refer sections 1.5.1 and 1.5.2)* **(10 Marks)**

Q. 19 What is meant by meta data ? Explain with an example. Explain the different types of meta data stored in a data warehouse. *(Ans. : Refer sections 1.7.1, and 1.7.3 and 1.7.4)* **(10 Marks)**

Q. 20 State and explain the various schemas used in data warehousing with examples for each of them. *(Ans. : Refer sections 1.10, 1.12 and 1.13)* **(10 Marks)**

Q. 21 Define what is meant by information package diagram. For recording the information requirements for "hotel occupancy" having dimensions like time, hotel etc, give the information package diagram for the same, also draw the star schema and snow flake schema. *(Ans. : Refer Ex. 1.17.3)* **(10 Marks)**

Q. 22 Explain dimension modeling in detail. *(Ans. : Refer section 1.8.1)* **(10 Marks)**

Dec. 2012

Q. 23 What is the role of Meta data in a data warehouse? Illustrate with examples. *(Ans. : Refer section 1.7)* **(10 Marks)**

Q. 24 Consider the following database for a chain of bookstores.

BOOKS (Booknum, Primary_author, Topic, Total_Stock, price)

BOOKSTORE (Storenum, City, state, Zip, Inventory_value)

STOCK (Storenum, Booknum, Qty)

With respect to the above business scenario, answer the following questions. Clearly state any reasonable assumptions you make.

(a) Design an information package diagram.

(b) Design a star schema for the data warehouse clearly identifying the Fact tables (s), Dimension table (s), their attributes and measures. *(Ans. : Refer Ex. 1.17.13)* **(10 Marks)**

Q. 25 Define the following terms by giving examples :

(a) Fact less Fact tables *(Ans. : Refer section 1.14.2)*

(b) Snowflake Schema *(Ans. : Refer section 1.12)* **(10 Marks)**

Scanned by CamScanner

May 2013

Q. 26 What are differences between Data Warehouse and Data Mart
(Ans. : Refer section 1.4)

Q. 27 Explain the role of Meta data in a data warehouse. **(5 Marks)**
(Ans. : Refer section 1.7)

Q. 28 Write detailed notes on Data Warehouse Architecture. **(10 Marks)**
(Ans. : Refer section 1.6.2)

Q. 29 For a Supermarket Chain consider the following dimensions, namely Product, store, time, promotion. The schema contains a central fact table, sales facts with three measures unit sales, dollars_sales and dollar_cost. Design star schema for this application. **(10 Marks)**

Calculate the maximum number of base fact table records for warehouse with the following values given below :

Time period: 5 years

Store: 300 stores reporting daily sales

Product: 40,000 products in each store (about 4000 sell in each store daily)
(Ans. : Refer Ex. 1.17.3)

Q. 30 Define the following terms by giving examples **(10 Marks)**

(a) Factless fact tables *(Ans. : Refer section 1.14.2)* **(5 Marks)**

(b) Snowflake Schema *(Ans. : Refer section 1.12)* **(5 Marks)**

Dec. 2013

Q. 31 What is a Data ware house? Explain the three tier architecture of a Data Ware house with a block diagram. *(Ans. : Refer sections 1.3 and 1.6.3)* **(10 Marks)**

Q. 32 What is meant by meta data? Explain with example. Explain the different types of meta data stored in a data ware house. Illustrate with examples.
(Ans. : Refer sections 1.7.1, 1.7.3 and 1.7.4) **(10 Marks)**

Q. 33 What is meant by Information Package Diagram, For recording the information requirements for "Hotel Occupancy" having dimensions like time, hotel etc., give the information package diagram for; the same, also draw the star schema and snowflake schema. *(Ans. : Refer Ex. 1.17.3)* **(10 Marks)**

May 2014

Q. 34 Explain clearly the role of meta data in a data warehouse **(10 Marks)**
(Ans. : Refer section 1.7)

Q. 35 Write detailed notes on Data Warehouse Architecture. **(10 Marks)**
(Ans. : Refer section 1.6.2)

(10 Marks)

Scanned by CamScanner

Q. 36 A bank wants to develop a data warehouse for effective decision-making about their loan schemes. The bank provides loans to customers for various purposes like house building loan, car loan, educational loan, personal loan, etc. The whole country is categorized into a number of regions, namely, north, south, east and west. Each region consists of a set of states. Loan is disbursed to customers at interest rates that change from time to time. Also, at any given point of time, the different types of loans have different rates. The data warehouse should record an entry for each disbursement of loan to customer.

(a) Design an information package diagram for the application.

(b) Design a star schema for the data warehouse clearly identifying the fact tables (s), dimensional table (s), their attributes and measures.

(Ans. : Refer Ex. 1.17.9) **(10 Marks)**

Q. 37 Define the following terms by giving examples :

(a) Fact constellation *(Ans. : Refer section 1.13)*

(b) Snowflake schema *(Ans. : Refer section 1.12)*

(c) Aggregate fact tables *(Ans. : Refer section 1.16)* **(15 Marks)**

Dec. 2014

Q. 38 What are the different characteristics of a Data Warehouse ?
(Ans. : Refer section 1.3.3) **(5 Marks)**

Q. 39 Explain the role of meta data in a data warehouse.
(Ans. : Refer section 1.7) **(10 Marks)**

Q. 40 Write detailed notes on Data Warehouse Architecture.
(Ans. : Refer section 1.6.2) **(10 Marks)**

Q. 41 For a supermarket chain consider the following dimensions, namely product, store, time, promotion. The schema contains a central fact table, sales facts with three measures unit_sales, dollars_sales and doller_cost. Design star schema for this application. *(Ans. : Refer Ex. 1.17.4)* **(5 Marks)**

Q. 42 Define the following terms : **(10 Marks)**

(a) Dimension tables *(Ans. : Refer section 1.10)*

(b) Snowflake schema *(Ans. : Refer section 1.12)*

May 2016

Q. 43 Illustrate the architecture of a typical DW system. Differentiate DW and data mart.
(Ans. : Refer section 1.6.2 and 1.4) **(10 Marks)**

Q. 44 Write short note on : Operational Vs. decision support system.
(Ans. : Refer section 1.2.3) **(5 Marks)**

Scanned by CamScanner

Q. 45 For a super market chain, consider the following dimensions namely product, store time and promotion. The schema contains a central fact table for sales.

i. Design star schema for the above application.

ii. Calculate the maximum number of base fact table records for warehouse with the following values given below :

Time period-5 years

Store - 300 stores reporting daily sales

Product - 40,000 products in each store (about 4000 sell in each store daily)

(Ans. : Refer Ex. 1.17.4) **(10 Marks)**

Q. 46 Write short note on : Updates to dimension tables.

(Ans. : Refer section 1.15) **(5 Marks)**

❑❑❑

Chapter Ends

CHAPTER 2

E

Syllabus :

Major steps in ETL proce
tasks, Major transformat
OLAP definition, Dimens
up, Slice, Dice and Rotat

2.1 An Introduction

2.1.1 What is ETL Too

- An ETL tool is a tool
  it is compatible with a
- ETL functions reshap
  to be stored in the da

2.1.2 Desired Featu

- Automated data mo
  warehouse, data ma
- Extensible, robust,
- Standardization of
- Reusabi

Scanned by CamScanner

CHAPTER 2

Module 2

ETL Process and OLAP

Syllabus :

Major steps in ETL process, Data extraction : Techniques, Data transformation : Basic tasks, Major transformation types, Data Loading : Applying Data, OLTP Vs OLAP, OLAP definition, Dimensional Analysis, Hypercubes, OLAP operations : Drill down, Roll up, Slice, Dice and Rotation, OLAP models : MOLAP, ROLAP.

2.1 An Introduction to ETL Process

→ **(MU - May 2012)**

2.1.1 What is ETL Tool?

- An ETL tool is a tool that reads data from one or more sources, transforms the data so that it is compatible with a destination and loads the data to the destination.
- ETL functions reshape the relevant data from the source systems into useful information to be stored in the data warehouse.

2.1.2 Desired Features

- Automated data movement across data stores and the analytical area in support of a data warehouse, data mart or ODS effort.
- Extensible, robust, scalable infrastructure.
- Standardization of ETL processes across the enterprise.
- Reusability of custom and pre-build functions.

Scanned by CamScanner

- Better utilization of existing hardware resources.
- Faster change-control and management.
- Integrated meta-data management.
- Complete development environment, "work as you think" design metaphor.

Syllabus Topic : Major Steps in ETL Process

2.2 Major Steps in ETL Process

→ (MU - May 20

Following are the major steps in the ETL process, these steps gives a good insight into the ETL process.

- Find out all the target data needed in the data warehouse.
- Find out all the data sources (internal and external).
- Prepare data mapping for target data elements from sources.
- Establish comprehensive data extraction rules.
- Determine data transformation and cleansing rules.
- Plan for aggregate tables.
- Organise data staging area and test tools.
- Write procedures for all data loads.
- ETL for dimension tables.
- ETL for fact tables.

Syllabus Topic : Data Extraction Techniques

2.3 Data Extraction

→ (MU - May 2016)

- Data extraction and conversions differ in a data warehouse as compared to operational systems. Two major factors differentiate them they are :
  1. Data is extracted from many disparate sources for a data warehouse.
  2. Data needs to be extracted on an ongoing incremental loads and one time initial full load.

- These two factor
warehouse. This
house program or

List of Data extracti

- **Source identific**
- **Method of extr**
each of the sourc
- **Extraction freq**
be carried out o
- **Time window**
- **Job Sequencin**
job has finishe
- **Exception Ha**

2.4 Identific

- List each dat
- List each dir
- For each tar
- If there are
- Identify mu
- Identify sir
- Ascertain
- Inspect so

2.5 Data

Source s
of data extra
data falls.

Scanned by CamScanner

- These two factors increase the complexity of the data extraction process in a data warehouse. This data extraction can be carried out using a third party tool along with in house program or scripts.

List of Data extraction Issues

- **Source identification :** The Source Application and the source structures are identified.
- **Method of extraction :** The extraction process can be either manual or tool based for each of the source systems.
- **Extraction frequency :** Data extraction frequency needs to be defined whether it should be carried out on a daily basis, weekly or quarterly.
- **Time window :** Time window needs to be decided for each data source.
- **Job Sequencing :** find out whether the beginning of one job has to wait until the previous job has finished completely.
- **Exception Handling :** The input records that cannot be extracted have to be handled upon

2.4 Identification of Data Sources

→ (MU - May 2016)

- List each data item of metrics or facts needed for analysis in fact tables.
- List each dimension attribute from all dimensions.
- For each target data item, find the source system and source data item.
- If there are multiple sources for one data element, choose the preferred source.
- Identify multiple source fields for a single target field and form consolidation rules.
- Identify single source field for multiple target fields and establish splitting rules.
- Ascertain default values.
- Inspect source data for missing values.

2.5 Data in Operational Systems

→ (MU - May 2010, Dec. 2010, May 2011, Dec. 2011, May 2012, Dec. 2012, May 2013, Dec. 2013, May 2014, Dec. 2014, May 2016)

Source systems usually store data in two ways, current value and Periodic status. The type of data extraction techniques to be used has to be decided based on the type of category the data falls.

Scanned by CamScanner

Current Value

- The value of an attribute represents the value at this moment of time.
- Values are transient or transitory.
- The change of value in this category is dependent on the business transaction.
- Change in value or how long the value will stay cannot be predicted.
- Some of the e.g. are Customer name, address, account balance and outstanding amoun on an individual order.

Periodic Status

In Periodic status, every time a change occurs the value of the attribute needs to b preserved.

Storing current value

Attribute : Customer's state of residence

6/1/2000	Value : OH
9/15/2000	Changed to CA
1/22/2001	Changed to NY
3/1/2001	Changed to NJ

Storing Periodic status

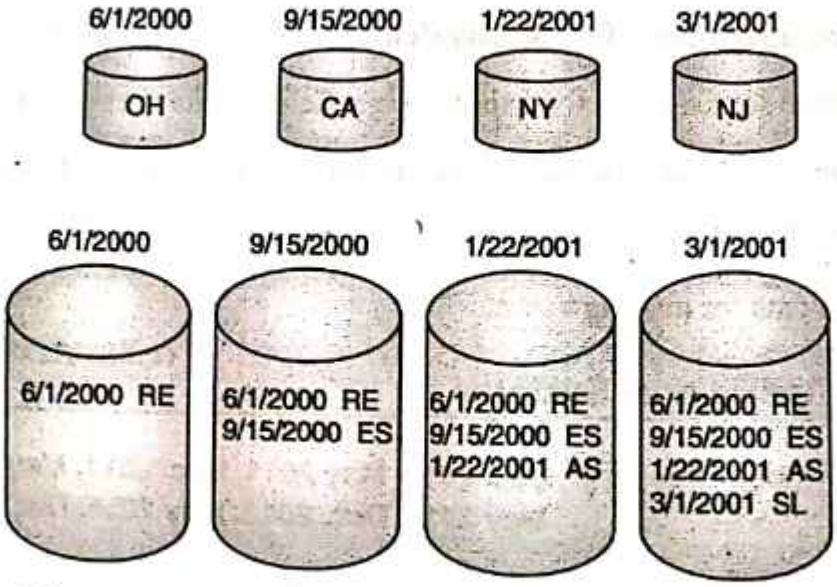


Fig. 2.5.1 : Example of Current Value and Periodic Status

Scanned by CamScanner

Attribute : Status of Property consigned to an auction house for sale.

6/1/2000	Value : RE (property receipted)
9/15/2000	Changed to ES (value estimated)
1/22/2001	Changed to AS (assigned to auction)
3/1/2001	Changed to SL (property sold)

2.5.1 Immediate Data Extraction

- Immediate Data extraction is real time, it occurs as the transaction happens.
- There are three options for immediate data extraction : Capture through transaction logs, Capture through database triggers, and capture in source applications.

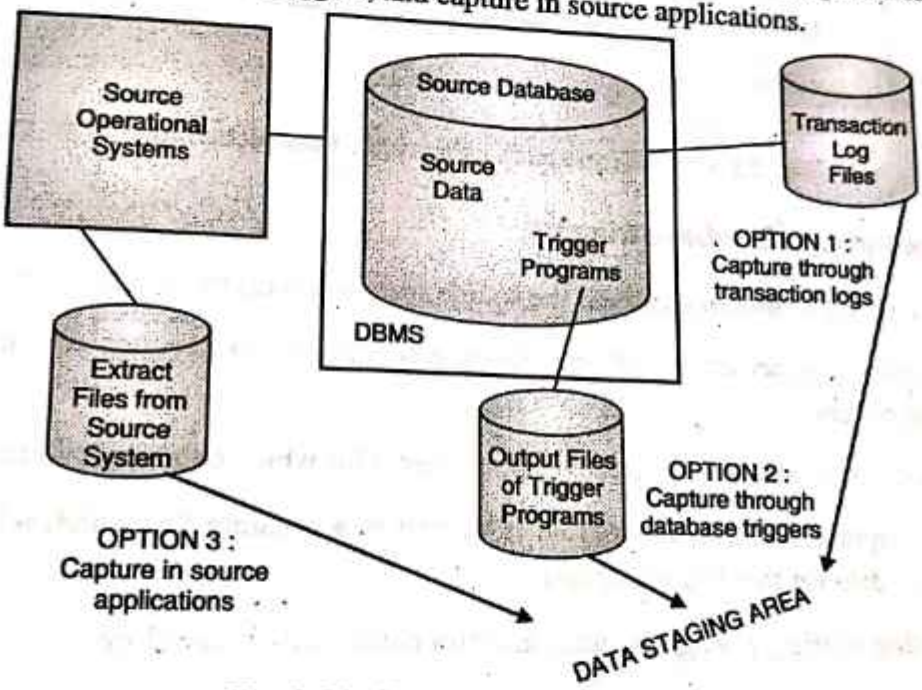


Fig. 2.5.2 : Immediate data extraction

1. Capture through transaction logs

- In this option, the transactional log maintained by DBMS is used for data extraction.
- All the committed transactions are selected from the transactional log.
- All the transactions have to be extracted before the log file is refreshed.
- This option is not feasible if the data source is other than the database applications like for e.g. indexed or flat files.
- Data replication is a method for creating copies in a distributed environment.

Scanned by CamScanner

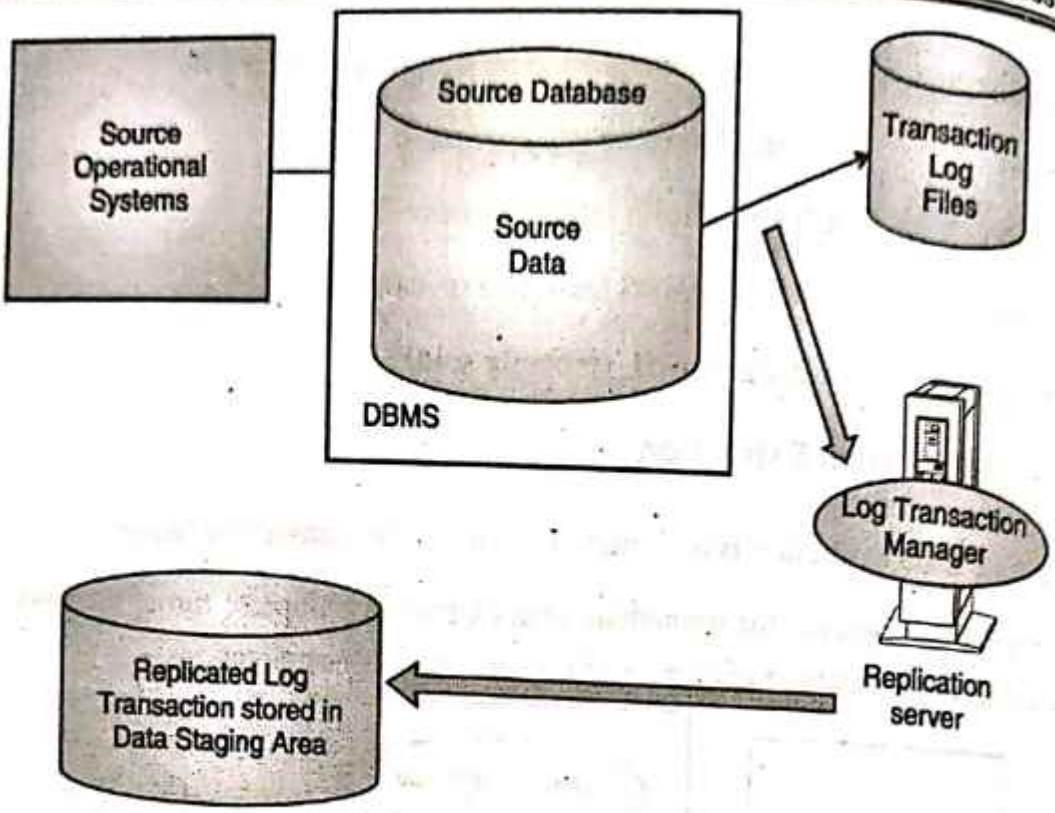


Fig. 2.5.3 : Data extraction using replication technology

2. Capture through Database triggers

- This option is also applicable if the source system is a database application.
- Database triggers are stored procedures which are invoked when certain predefined event occurs.
- Trigger programs may be created for all events for which data needs to be captured.
- The output of the trigger program is written in a separate file which will be used to extract data for the data warehouse.
- Building of trigger programs adds an extra burden on the development effort.

3. Capture in source applications

- This option is also known as Application assisted data capture.
- Relevant application programs needs to be modified that write to source files and databases.
- Other extract programs may use the separate file containing the changes to the source data.
- This option can be used for any type of source system, for eg., A database application, indexed file or flat file.

Scanned by CamScanner

- Revision and maintenance of the programs in the source operational system is needed.
- This option may degrade the performance due to additional processing needed to capture the changes.

2.5.2 Deferred Data Extraction

- This option does not capture the changes in real time.
- There are two options : (i) capture based on date (ii) time stamp and capture by comparing files.

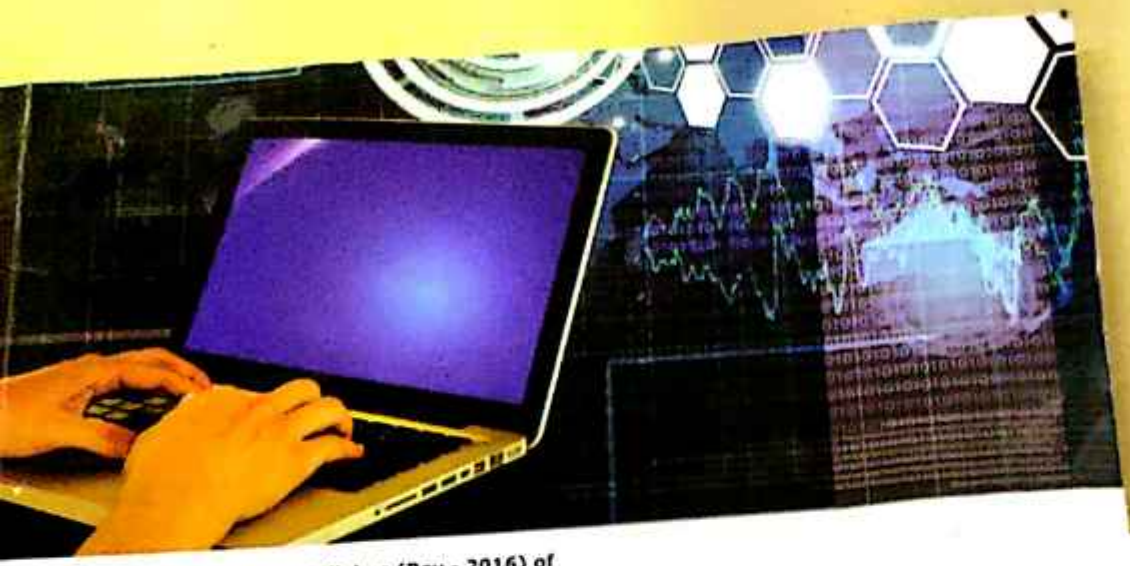
1. Capture based on date and time stamp

- Time stamp is considered as a basis for selecting the records for data extraction.
- A time stamp is marked to the source record whenever it is created or updated.
- This technique works well if the number of revised records are small.
- This technique works for any type of source file, provided the source file contains a date and time stamp. .

2. Capture by comparing files

- If all of the above techniques are not feasible for the source file then this option is considered.
- This technique is also called snapshot differential technique as it compares two snapshots of the source data.
- This technique captures the changes by comparing the two copies of data.
- Although this technique is simple, comparing full rows in a large file can be very inefficient.
- However this technique is feasible when the legacy source systems do not have transaction logs, time stamp on source records.

Scanned by CamScanner



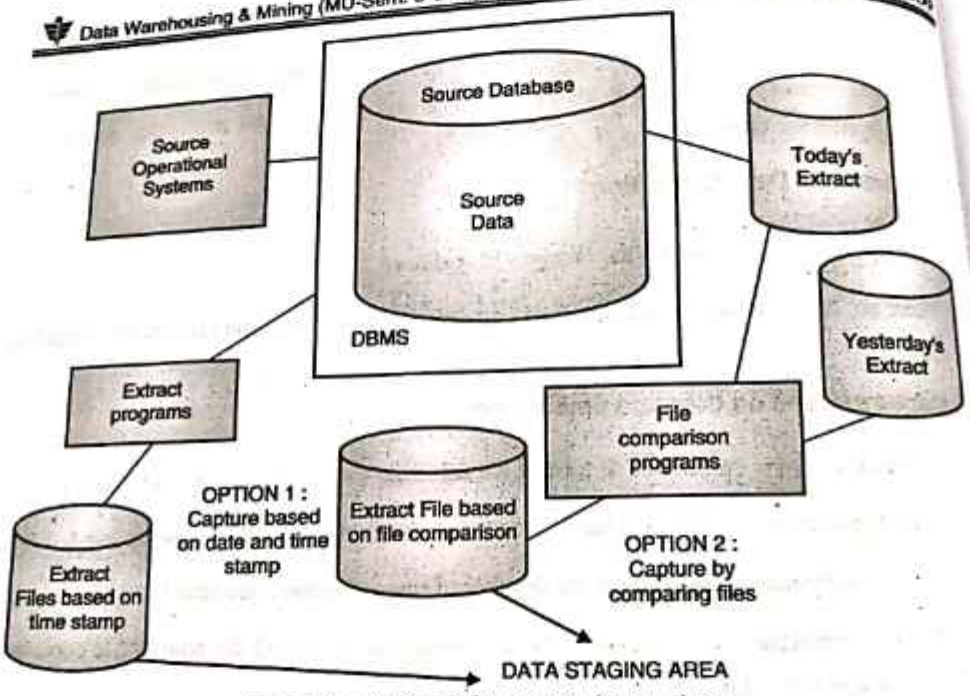


Fig. 2.5.4 : Deferred data extraction: options

Syllabus Topic : Data Transformation - Basic Tasks, Major Transformation Types

2.6 Data Transformation : Tasks Involved in Data Transformation

→ (MU - May 2010, Dec. 2010, May 2011, Dec. 2011, May 2012, Dec. 2012, May 2013, Dec. 2013, May 2014, Dec. 2014, May 2016)

- The operational databases developed can be based on any set of priorities, which keeps changing with the requirements.
- Therefore those who develop data warehouse based on these databases are typically faced with inconsistency among their data sources.
- Transformation process deals with rectifying any inconsistency (if any).
- One of the most common transformation issues is **'Attribute Naming Inconsistency'**. It is common for the given data element to be referred to by different data names in different databases.

Scanned by CamScanner

- **Example :** Employee Name may be EMP_NAME in one database and ENAME in the other.
- Thus one set of Data Names are picked and used consistently in the data warehouse.
- Once all the data elements have right names, they must be converted to common formats.
- The conversion may encompass the following :
  - Characters must be converted ASCII to EBCDIC or vice versa.
  - Mixed Text may be converted to all uppercase for consistency.
  - Numerical data must be converted in to a common format.
  - Data Format has to be standardized.
  - Measurement may have to convert. (Rs/ $)
  - Coded data (Male/ Female, M/F) must be converted into a common format.
- All these transformation activities are automated and many commercial products are available to perform the tasks. DataMAPPER from Applied Database Technologies is one such comprehensive tool.

Data transformation can have the following activities

- **Smoothing :** It removes noise from the data.
- **Aggregation :** It uses summarization, data cube construction.
- **Generalization :** It uses concept hierarchy to replace the data by higher level concepts.
- **Normalization :** Attributes are scaled to fall within a small, specified range.
- Scaling attribute values to fall within a specified range.

  Example : To transform V in [min, max] to V′ in [0,1], apply

  $$V' = (V\text{-Min}) / (\text{Max-Min})$$

- Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers) :

  $$V' = (V\text{-Mean}) / \text{Std. Dev.}$$

- **Attribute/feature construction :** New attributes are constructed from the given ones and added for data mining process.

Scanned by CamScanner

2.6.1 The Set of Basic Tasks

1. **Selection of data**

   Select the desired data from the source system; it can be either whole records or par several records.

2. **Splitting/joining**

   - It is the data manipulation on selected data records.
   - It can be either a Split or a join operation of the selected parts during t transformation.

3. **Conversion**

   This step includes conversions of single fields to standardize the data extraction forma from different source systems and understandable to the users.

4. **Summarization**

   - Sometimes the lowest granularity for analysis or querying of data in data warehous may not be needed. So, in this case, the data transformation function include summarization of data.
   - For example, for a credit card company to analyze sales patterns, it may not be necessary to store in the data warehouse every single transaction on each credit card. Instead, you may want to summarize the daily transactions for each credit card and store the summary data instead of storing the most granular data by individual transactions.

5. **Enrichment**

   It is the rearranging and simplifying of individual fields to make them more useful in the ETL process.

2.7 Data Integration and Consolidation

→ (MU - May 2010, Dec. 2010, May 2011, Dec. 2011, May 2012, Dec. 2012, May 2013, Dec. 2013, May 2014, Dec. 2014, May 2016)

- Integration of data is combining all the relevant operational data into coherent data structures to be made ready to be loaded into a data warehouse.

Scanned by CamScanner

- It can be considered as a pre-process type before major transformation techniques are applied.
- Standardisation of names, data representations and resolving any discrepancies in the way the same data is represented in different source systems is required.

Syllabus Topic : Data Loading - Applying Data

2.8 Data Loading : Techniques of Data Loading

(MU - May 2010, Dec. 2010, May 2011, Dec. 2011, May 2012, Dec. 2012, May 2013, Dec. 2013, May 2014, Dec. 2014, May 2016)

Loading process is the physical movement of the data from the computer systems storing the source database(s) to that which will store the data warehouse database.

The whole process of moving data into the data warehouse repository is referred to in the following ways :

(1) **Initial Load :** For the very first time loading all the data warehouse tables.

(2) **Incremental Load :** Periodically applying ongoing changes as per the requirement.

(3) **Full Refresh :** Deleting the contents of a table and reloading it with fresh data.

Data Refresh versus Update

After the initial load, the data warehouse needs to be maintained and updated and this can be done by the following two methods :

- Update-application of incremental changes in the data sources.
- Refresh-complete reload at specified intervals.

Data Loading

- Data are physically moved to the data warehouse.
- The loading takes place within a "load window".
- The trend is to near real time updates of the data warehouse as the warehouse is increasingly used for operational applications.

2.8.1 Loading the Dimension Tables

- Procedure for maintaining the dimension tables includes two functions, initial loading of the tables and thereafter applying the changes on an ongoing basis.

Scanned by CamScanner

ctly as per the New Revised Syllabus (Rev - 2016) of
mbai University
academic year 2018-2019
r Choice Based Cre

- System generated keys are used in a data warehouse.
- The records in the source system have their own keys.
- Therefore before an initial load or an ongoing load the production keys must be converted to system generated keys in the data warehouse.
- Another issue is related to the application of Type 1, Type 2 and Type 3 dimension changes to the data warehouse. Fig. 2.8.1 shows how to handle it.

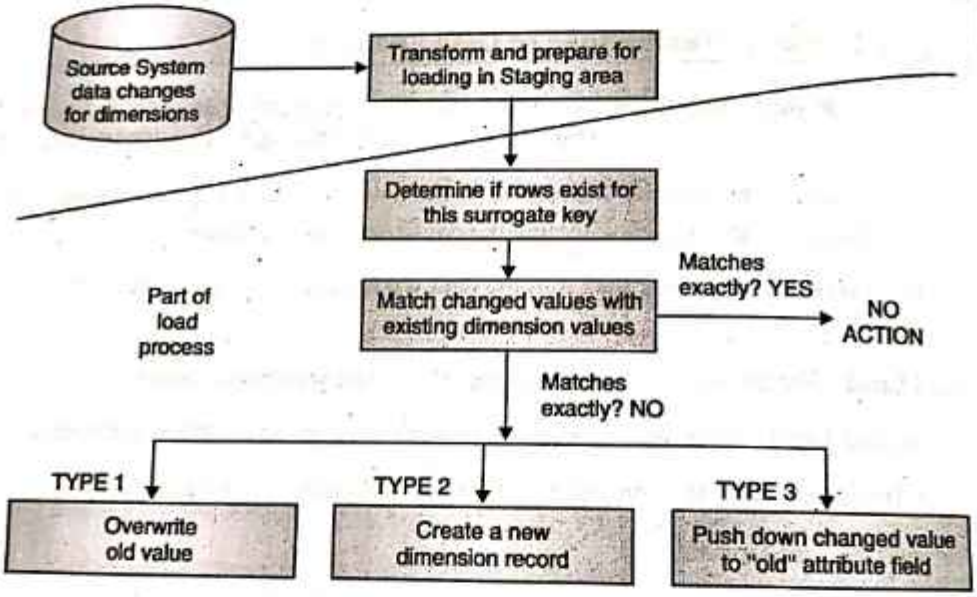


Fig. 2.8.1 : Loading changes to dimension tables

2.8.2 Loading the Fact tables : History and Incremental Loads

- The key in the fact table is the concatenation of keys from the dimension tables.
- Therefore for this reason dimension records are loaded first.
- A concatenated key is created from the keys of the corresponding dimension tables.

2.9 Data Quality : Issues in Data Cleansing

Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g. due to misspellings during data entry, missing information or other invalid data.

Scanned by CamScanner

2.9.1 Reasons for "Dirty" Data

- Dummy Values,
- Absence of Data,
- Multipurpose Fields,
- Cryptic Data,
- Contradicting Data,
- Inappropriate Use of Address Lines,
- Violation of Business Rules,
- Reused Primary Keys,
- Non-Unique Identifiers,
- Data Integration Problems.

2.9.2 Data Cleansing

- Source systems contain "dirty data" that must be cleansed.
- Specialized data cleansing software is often used. Important for performing name and address correction and house holding functions.
- Leading data cleansing vendors include Vality (Integrity), Harte-Hanks (Trillium), and First logic.

Steps in Data Cleansing

1. **Parsing**
   - The individual data elements are located and identified in the source files and then these elements are isolated in the target files.
   - For e.g. Name can be parsed as the first, middle, and last name, another example, Address can be parsed as street number and street name, city and state.
2. **Correcting**
   - Sophisticated data algorithm and secondary data sources are used for correction of individual data components.
   - For e.g. replace a vanity address and addition of a zip code.

Scanned by CamScanner

3. **Standardizing**

- Standardizing applies conversion routines to transform data into its preferred (and consistent) format using both standard and custom business rules.
- Examples include adding a pre name, replacing a nickname, and using a preferred street name.

4. **Matching**

- Searching and matching records within and across the parsed, corrected and standardized data based on predefined business rules to eliminate duplications.
- Examples include identifying similar names and addresses.

5. **Consolidating**

The matched records are analysed and identified for relationships and then consolidating and merging them into ONE representation.

6. **Data Cleansing must deal with many types of possible errors**

- These include missing data and incorrect data at one source.
- Inconsistent data and conflicting data when two or more sources are involved.

7. **Data Staging**

- Often used as an interim step between data extraction and later steps.
- Accumulates data from asynchronous sources using native interfaces, flat files, FTP sessions, or other processes.
- At a predefined cut off time, data in the staging file is transformed and loaded to the warehouse.
- There is usually no end user access to the staging file.
- An operational data store may be used for data staging.
- Information quality is the key consideration in determining the value of the information. The developer of the data warehouse is not usually in a position to change the quality of its underlying historic data, though a data warehousing project can put spotlight on the data quality issues and lead to improvements for the future. It is, therefore, usually necessary to go through the data entered into the data warehouse and make it as error free as possible. This process is known as ***Data Cleansing***.

Scanned by CamScanner

0 Sample ETL Tools

The ETL Process is shown in the Fig. 2.10.1.

Teradata Warehouse Builder from Teradata.

Data Stage from Accentual Software.

SAS System from SAS Institute.

Power Mart/Power Center from Informatica.

Sagent Solution from Sagent Software.

Hummingbird Genio Suite from Hummingbird Communications.

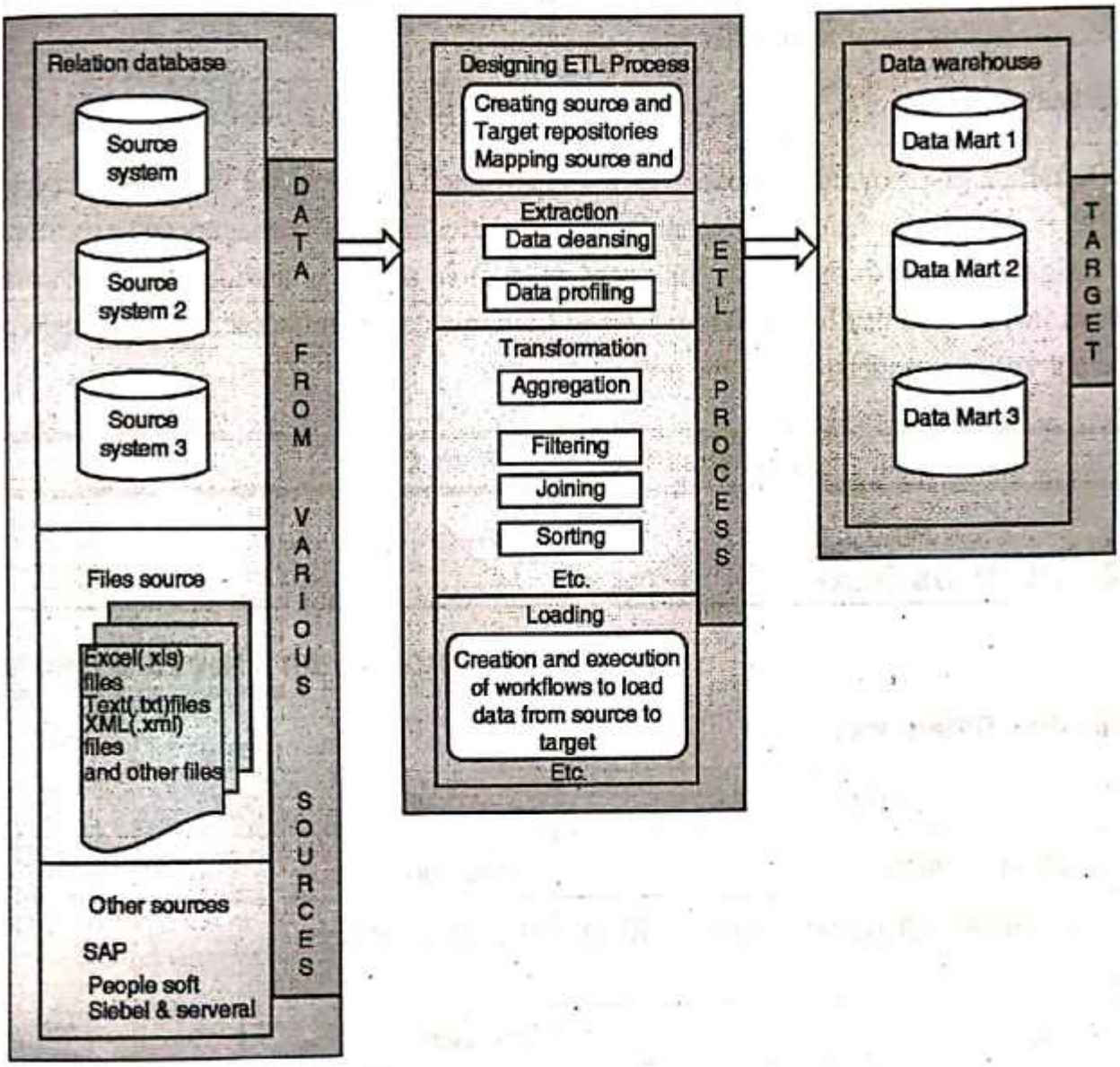


Fig. 2.10.1 : ETL Process

Scanned by CamScanner

2.11 Need for Online Analytical Processing

- OLAP or the On Line Analytical supports the multidimensional view of data.
- OLAP provides fast, steady, and proficient access to the various views of information.
- The complex queries can be processed.
- It's easy to analyze information by processing complex queries on multidimensional views of data.
- Data warehouse is generally used to analyse the information where huge amount of historical data is stored.
- Information in data warehouse is related to more than one dimension like sales, market trends, buying patterns, supplier, etc.

Definition

Definition given by OLAP council (www.olapcouncil.org) On-Line Analytical Processing (OLAP) is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access in a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user.

Syllabus Topic : OLTP V/s OLAP

2.12 OLTP V/s OLAP

→ (MU - Dec. 2010, May 2011, May 2012, Dec. 2013)

Application Differences

OLTP	OLAP
Transaction oriented	Subject oriented
High Create/Read/Update/ Delete (CRUD) activity	High Read activity
Many users	Few users
Continuous updates – many sources	Batch updates – single source
Real-time information	Historical information
Tactical decision-making	Strategic planning

OL
Controlled, customize
RDBMS
Operational database

Modeling Objective

High transaction v at a time.
Balancing needs batch processing.
Highly volatile da
Data redundancy
Few levels of gra
Complex datal personnel.

Model Differen

Single purpose System.
Full set of En
Eliminate red
Natural or su
Validate M Analysis.
Technical requiremen
This mome

2.13 OL

- Multi
- Multi

Scanned by CamScanner

OLTP	OLAP
Controlled, customized delivery	"Uncontrolled", generalized delivery
RDBMS	RDBMS and/or MDBMS
Operational database	Informational database

Modeling Objectives Differences

OLTP	OLAP
High transaction volumes using few records at a time.	Low transaction volumes using many records at a time.
Balancing needs of online v/s scheduled batch processing.	Design for on-demand online processing.
Highly volatile data.	Non-volatile data.
Data redundancy – *BAD*.	Data redundancy – *GOOD*.
Few levels of granularity.	Multiple levels of granularity.
Complex database designs used by IT personnel.	Simpler database designs with business-friendly constructs.

Model Differences

OLTP	OLAP
Single purpose model – supports Operational System.	Multiple models – support Informational Systems.
Full set of Enterprise data.	Subset of Enterprise data.
Eliminate redundancy.	Plan for redundancy.
Natural or surrogate keys.	Surrogate keys.
Validate Model against business Function Analysis.	Validate Model against reporting requirements.
Technical metadata depends on business requirements.	Technical metadata depends on data mapping results.
This moment in time is important.	Many moments in time are essential elements.

2.13 OLAP and Multidimensional Analysis

- Multidimensional analysis is done in data warehouse environment.
- Multidimensional analysis is identical with online analytical processing (OLAP).

Scanned by CamScanner

- To perform multidimensional analysis, OLAP is required which performs complex calculations and displays the desired information.
- OLAP gives multiple views of data to perform multidimensional analysis.

Syllabus Topic : Hypercubes

2.14 Hypercube

- Let us assume that there are two business dimensions Product and Time. Business users wish to analyse metrics as fixed cost, variable cost, indirect sales, direct sales, and profit margin. Then for analysis purpose, data can be displayed on spreadsheet which shows metrics as columns and time as row for one product per page as shown in Table 2.14.1.

Table 2.14.1 : Display of Spreadsheet for Product Trousers where Time as row and Metrics as columns

	Fixed Cost	Variable Cost	Indirect Sales	Direct Sales	Profit Margin
Jan.	340	110	230	320	100
Feb.	270	90	200	260	100
Mar.	310	100	210	270	70
April	340	110	210	320	80
May	330	110	230	300	90
June	260	90	150	300	100
July	310	100	180	300	70
Aug.	380	130	210	360	60
Sep.	300	100	180	290	70
Oct.	310	100	170	310	70
Nov.	330	110	210	310	80
Dec.	350	120	200	360	90

- This can also be represented by using 3 straight lines where two for business dimensions and one for metrics as shown in Fig. 2.14.1. This type of representation is called as Multidimensional Domain Structure (MDS)

Scanned by CamScanner

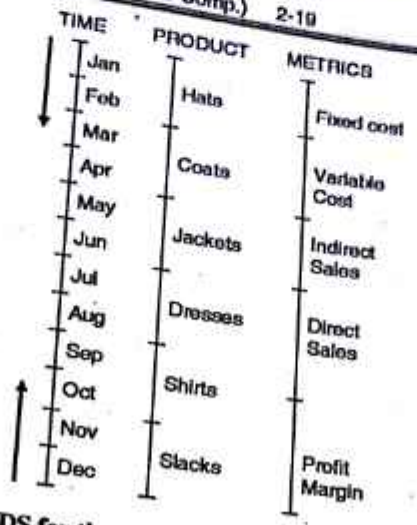


Fig. 2.14.1 : MDS for three Dimensions Time, Product and Metrics

- But in general business have more than two dimensions. Representation of more than three dimensions is called as hypercube which represents multidimensional data. So add one more dimension Store with Time and Product. MDS for 4 dimensions is given in Fig. 2.14.2.

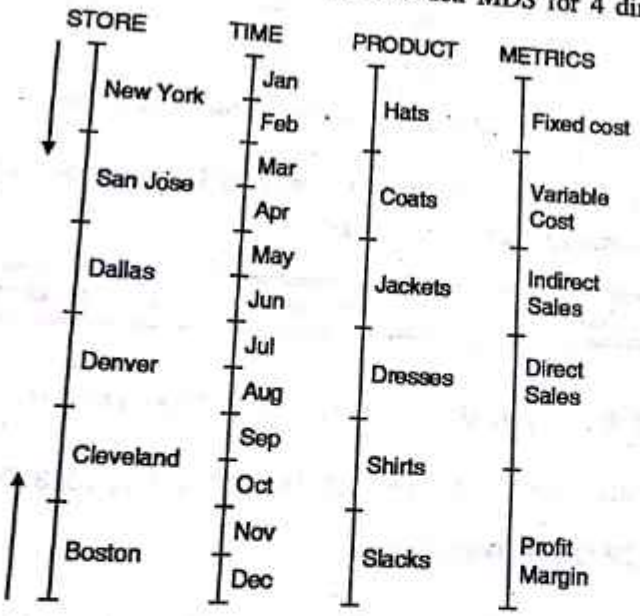
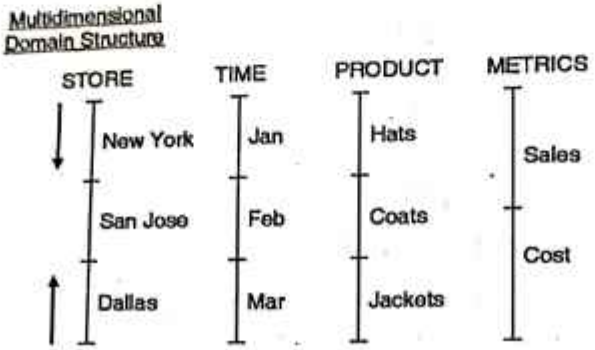


Fig. 2.14.2 : MDS for 4 dimensions

Scanned by CamScanner

- Representation of these 4 dimensions on the spreadsheet is also possible by combining multiple logical dimensions within the same display group. In the Fig. 2.14.3, product and metrics (Sales and Cost) are combined to display as columns. One page represents the sales for store : New York.



New York Store						
	Hats : Sales	Hats : Cost	Coats : Sales	Costs : Cost	Jackets : Sales	Jackets : Cost
Jan.	450	350	550	450	500	400
Feb.	380	280	460	360	400	320
Mar.	400	310	480	410	450	400

Fig. 2.14.3 : MDS and page display for 4 dimensions

- So hypercube is used for representation of more than 3 dimensions. Similarly one can add more dimensions and represent it using MDS.

Syllabus Topic : OLAP Operations - Drill down, Roll up, Slice, Dice and Rotation

2.15 OLAP Operations in Multidimensional Data Model

→ (MU - Dec. 2010, May 2012, Dec. 2012, May 2013, Dec. 2014, May 2016)

OLAP Operations OR OLAP Techniques

- OLAP techniques are implemented to retrieve the information from data warehouse into OLAP multi-dimensional databases. So information can be retrieved using front end system.

Scanned by CamScanner

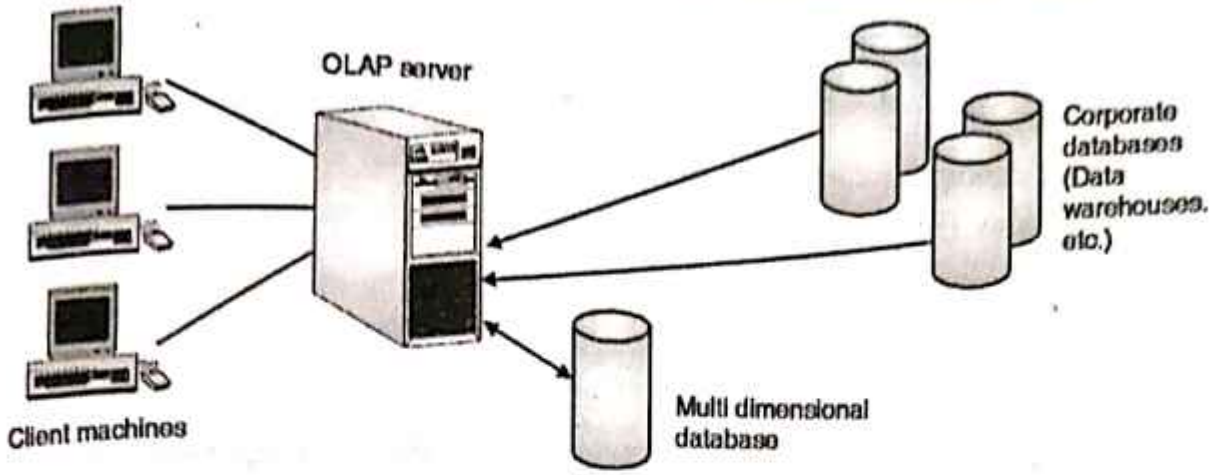


Fig. 2.15.1 : OLAP Implementation

- Multidimensional models is used to inhabit data in multi-dimensional matrices like Data Cubes or Hypercubes. A standard spreadsheet, signifying a conventional database, is a two-dimensional matrix. One example would be a spreadsheet of regional sales by product for a particular time period. Products sold with respect to region can be shown in 2 dimensional matrix but as one more dimension like time is added then it produces 3 dimensional matrix as shown in Fig. 2.15.2.

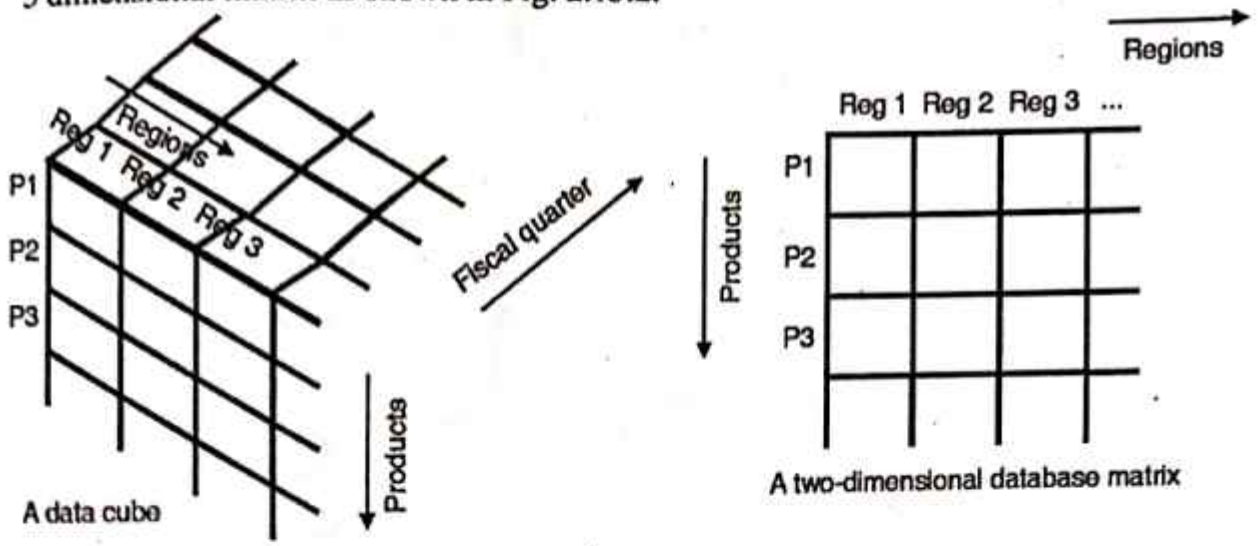


Fig. 2.15.2 : Pictorial view of data cube and 2D database

- A multidimensional model has two types of tables :

1. **Dimension tables :** contains attributes of dimensions
2. **Fact tables :** contains facts or measures

The following techniques are used for OLAP implementation :

Example : Let us consider a company of Electronic Products. Data cube of company consists of 3 dimensions Location (aggregated with respect to city), Time (is aggregated with respect to quarters) and item (aggregated with respect to item types).

Scanned by CamScanner

1. Consolidation or Roll Up

- Multi-dimensional databases generally have hierarchies with respect to dimensions.
- Consolidation is rolling up or adding data relationship with respect to one or more dimensions. For example, adding up all product sales to get total City data.
- For example, Fig. 2.15.3 shows the result of roll up operation performed on the central cube by climbing up the concept hierarchy for location. This hierarchy was defined as the total order street <city <province_or_state <country.
- The roll up operation shown aggregates the data by city to the country by location hierarchy.

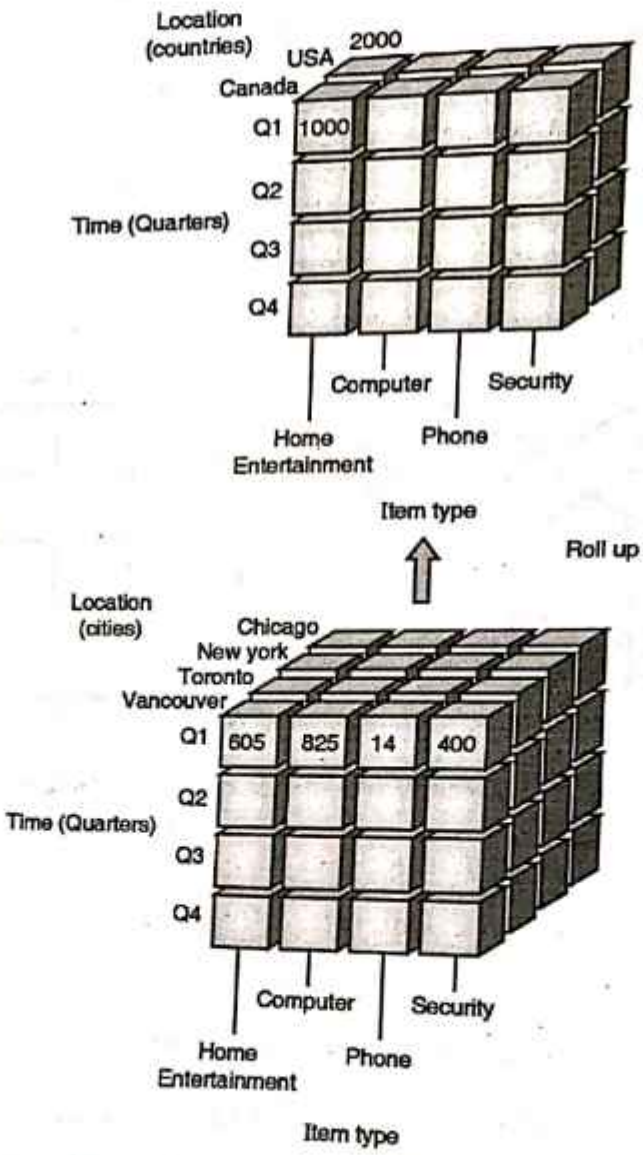


Fig. 2.15.3 : Roll -up or drill up

Scanned by CamScanner

2. Drill-down

- Drill Down is defined as changing the view of data to a greater level of detail.
- For example, the Fig. 2.15.4 shows the result of drill down operations performed on the upper cube by stepping down a concept hierarchy for time defined as day<month<quarter<year.

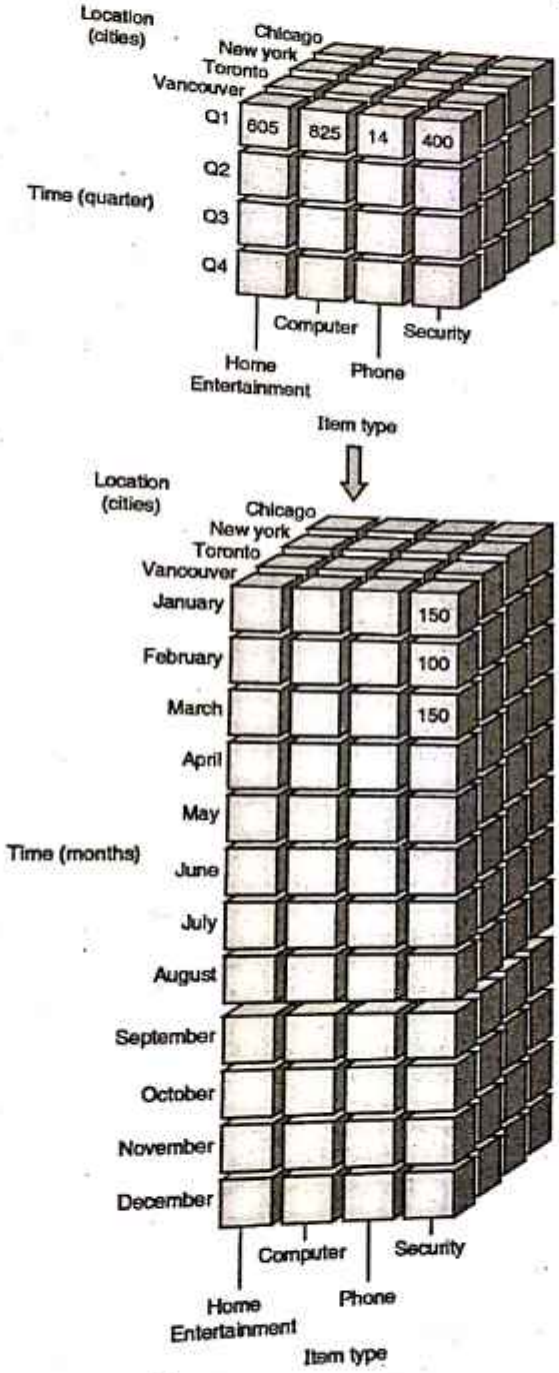


Fig. 2.15.4 : Drill Down

Scanned by CamScanner

y as per the New Revised Syllabus (Rev - 2016) of
nbai University
academic year 2018-2019
r Choice Based Credit and Grading System)

3. Slicing and dicing

- Slicing and dicing refers to the ability to look at a database from various viewpoints.
- Slice operation carry out selection with respect to one dimension of the given cube and produces a sub cube.
- For example, Fig. 2.15.5 shows the slice operation where the sales data are selected from the left cube for the dimension time using the criterion time = "Q1".

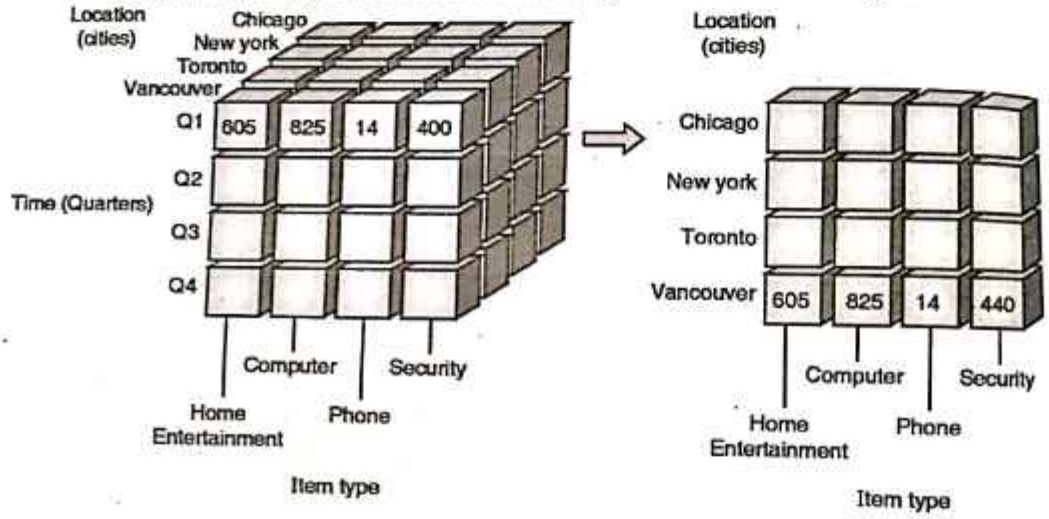


Fig. 2.15.5 : Slice operation

4. Dice

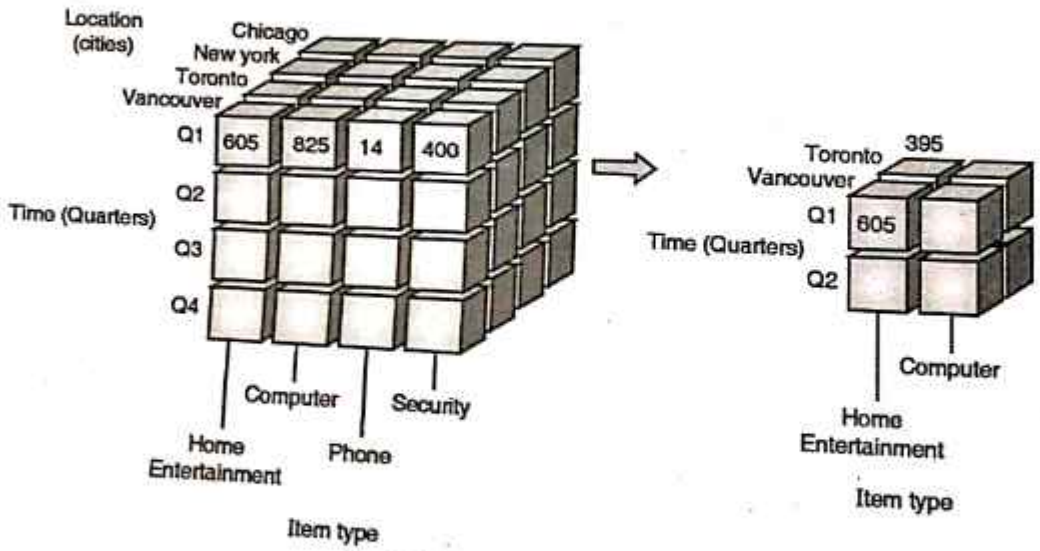


Fig. 2.15.6 : Dice operation

Scanned by CamScanner

- Dice operation carry out selection with respect to two or more dimensions of the given cube and produces a sub cube.
- For example, the dice operation is performed on the left cube based on three dimension as Location, Time and Item as shown in Fig. 2.15.6 where the criteria is (location= "Torrento" or "Vancouver") and (time = "Q1" or "Q2") and (item= "home entertainment" or "computer").

5. **Pivot / Rotate**

- Pivot technique is used for visualization of data. This operation rotates the data axis to give another presentation of the data.
- For example Fig. 2.15.7 shows the pivot operation where the item and location axis in a 2-D slice are rotated.

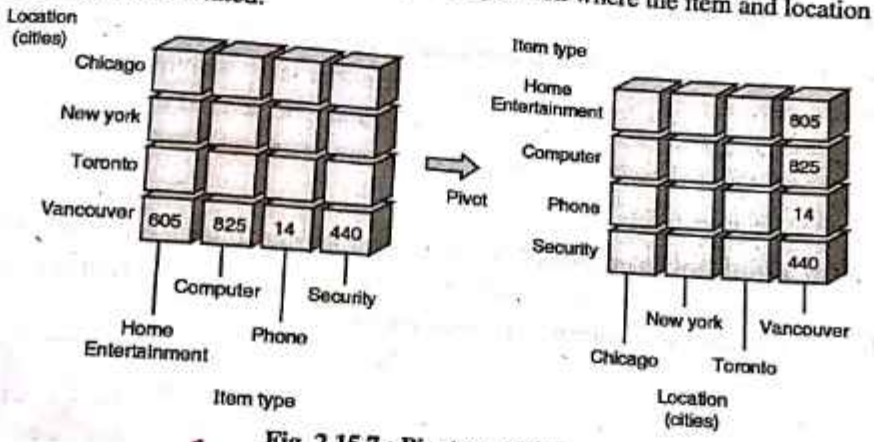


Fig. 2.15.7 : Pivot operation

6. **Other OLAP operations**

- **Drill across :** This technique is used when there is need to execute a query involving more than one fact table.
- **Drill through :** This technique uses relational SQL facilities to drill through the bottom level of the data cube.

Syllabus Topic : OLAP Models - MOLAP, ROLAP

2.16 OLAP Models : MOLAP, ROLAP, HOLAP, DOLAP

→ (MU - May 2016)

Approaches to OLAP Servers

In the OLAP world, there are mainly two different types of OLAP servers: Multidimensional OLAP (MOLAP) and Relational OLAP (ROLAP). Hybrid OLAP (HOLAP) refers to technologies that combine MOLAP and ROLAP.

Scanned by CamScanner

2.16.1 MOLAP

This is the more traditional way of OLAP analysis. In MOLAP, data is stored in a multidimensional cube. The storage is not in the relational database, but in proprietary formats.

Advantages of MOLAP

1. **Excellent performance :** MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.
2. **Can perform complex calculations :** All calculations have been pre-generated when the cube is created.

Disadvantages of MOLAP

1. **Limited in the amount of data it can handle :** Because all calculations are performed when the cube is built, it is not possible to include a large amount of data in the cube itself. This is not to say that the data in the cube cannot be derived from a large amount of data. Indeed, this is possible. But in this case, only summary-level information will be included in the cube itself.
2. **Requires additional investment :** Cube technology are often proprietary and do not already exists in the organization. Therefore, to adopt MOLAP technology, chances of additional investments in human and capital resources are needed.

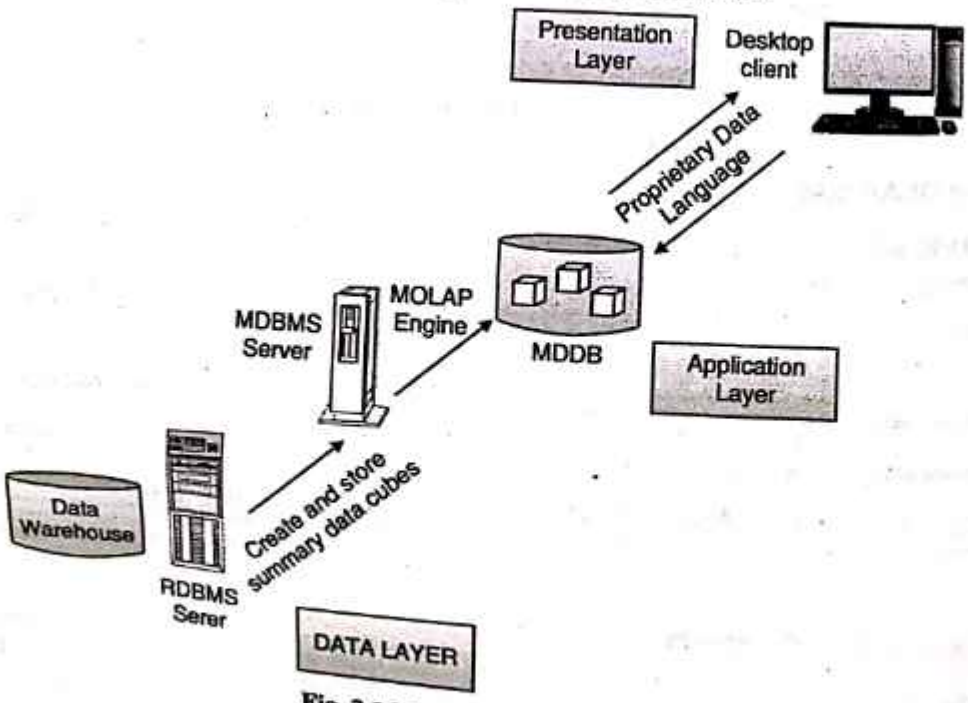


Fig. 2.16.1 : MOLAP Process

Scanned by CamScanner

2.16.2 ROLAP

This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality. In essence, each action of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement.

Advantages of ROLAP

1. **Can handle large amounts of data :** The data size limitation of ROLAP technology is the limitation on data size of the underlying relational database. In other words, ROLAP itself places no limitation on amount of data.
2. **Can leverage functionalities inherent in the relational database :** Often, relational database already comes with a host of functionalities. ROLAP technologies, since they sit on top of the relational database, can therefore leverage these functionalities.

Disadvantages of ROLAP

1. **Performance can be slow :** Because each ROLAP report is essentially a SQL query (or multiple SQL queries) in the relational database, the query time can be long if the underlying data size is large.
2. **Limited by SQL functionalities :** Because ROLAP technology mainly relies on generating SQL statements to query the relational database, and SQL statements do not fit all needs (for example, it is difficult to perform complex calculations using SQL), ROLAP technologies are therefore traditionally limited by what SQL can do.

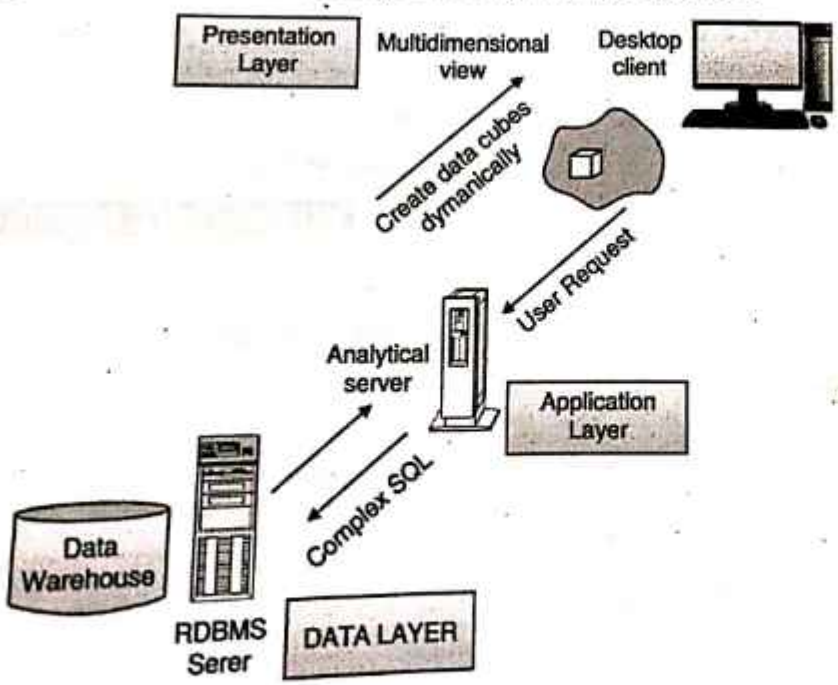


Fig. 2.16.2 : ROLAP Process

Scanned by CamScanner

3. ROLAP vendors have mitigated this risk by building into the tool out-of-the-box comp functions as well as the ability to allow users to define their own functions.

2.16.3 HOLAP

- HOLAP technologies attempt to combine the advantages of MOLAP and ROLAP. F summary-type information, HOLAP leverages cube technology for faster performance.
- When detail information is needed, HOLAP can "drill through" from the cube into t underlying relational data.
- For example, a HOLAP server may allow large volumes of detail data to be stored in relational database, while aggregations are kept in a separate MOLAP store. Th Microsoft SQL Server 7.0 OLAP Services supports a hybrid OLAP server.

2.16.4 DOLAP

It is Desktop Online Analytical Processing and variation of ROLAP. It offers portability to users of OLAP. For DOLAP, it needs only DOLAP software to be present on machine. Through this software, multidimensional datasets are formed and transferred to desktop machine.

2.17 Examples of OLAP

Ex. 2.17.1 : Consider a data warehouse for a hospital where there are three dimension (a) Doctor (b) Patient (c) Time

And two measures i) count ii) charge where charge is the fee that the doctor charges a patient for a visit.

Using the above example describe the following OLAP operations

1) Slice 2) Dice 3) Rollup 4) Drill down 5) Pivot

MU - May 2013, May 2014, 10 Marks

Soln. :

There are four tables, out of 3 dimension tables and 1 fact table.

Dimension tables

1. Doctor (DID, name, phone, location,pin,specialisation)
2. Patient (PID,name, phone , state, city,location, pin)
3. Time (TID,day, month, quarter , year)

Fact Table : Fact_table (D

D01

D02

D0

Operations

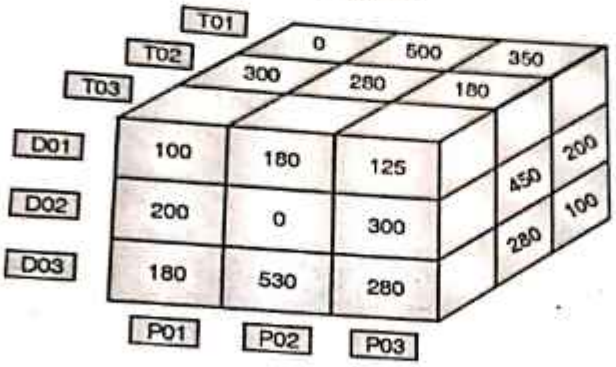
1. **Slice :** Slice on fac patient axis thus it

2. **Dice :** It is a s like dice on TID = 02, 03

3. **Roll up :** hierarchy or count

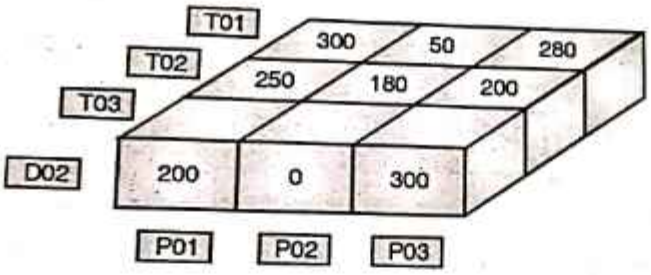
Scanned by CamScanner

Fact Table : Fact_table (DID,PID,TID, count, charge)

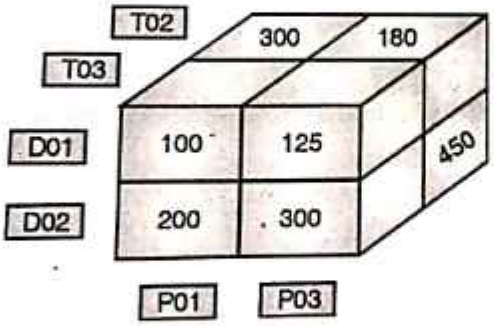


Operations

1. **Slice :** Slice on fact table with DID = 2 , this cuts the cube at DID = 2 along the time and patient axis thus it will display a slice of cube, in which time on x and patient on y axis.

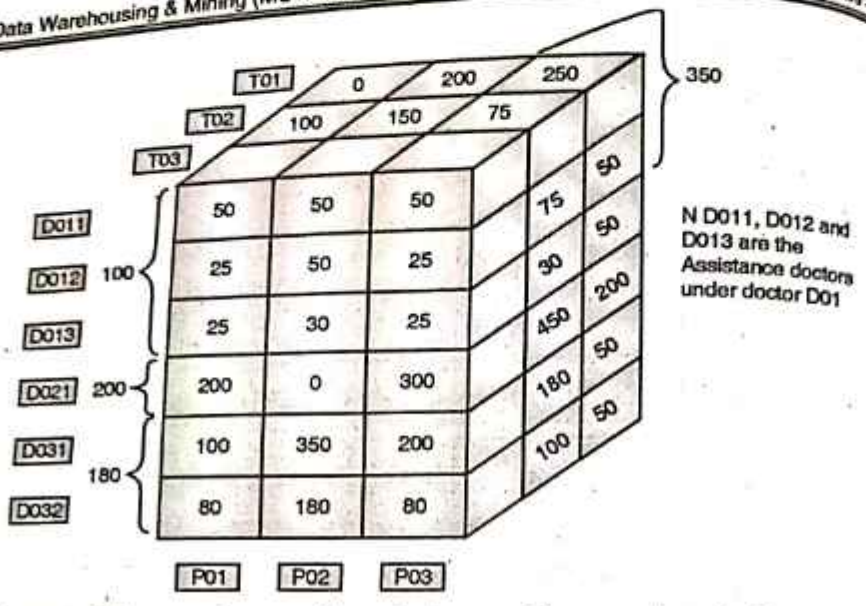


2. **Dice :** It is a sub cube of main cube. Thus it cuts the cube with more than one predicate like dice on cube with DID = 2, and DID = 01 and PID = 01 and PID = 03 and TID = 02, 03



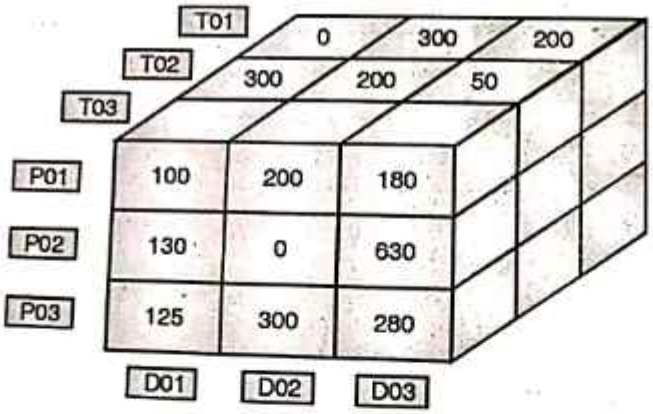
3. **Roll up :** It gives summary based on concept hierarchies. Assuming there exists concept hierarchy in patient table as state->city->location. Then roll up will summarise the charges or count in terms of city or further roll up will give charges for a particular state etc.

Scanned by CamScanner



N D011, D012 and D013 are the Assistance doctors under doctor D01

4. **Drill down :** It is opposite to roll up that means if currently cube is summarised w respect to city then drill down will also show summarisation with respect to location.



5. **Pivot :** It rotates the cube, sub cube or rolled –up or drilled –down cube, thus changing th view of the cube.

Ex. 2.17.2 : All Electronics Company have sales department consider three dimensions namely

(i) Time (ii) Product (iii) store

The Schema Contains a central fact table sales with two measures

(i) dollars-cost and (ii) units-sold

Using the above example describe the following OLAP operations :

(i) Dice (ii) Slice

(iii) Roll-up (iv) drill Down.

MU - Dec. 2013, 10 Marks

Scanned by CamScanner

Soln. :

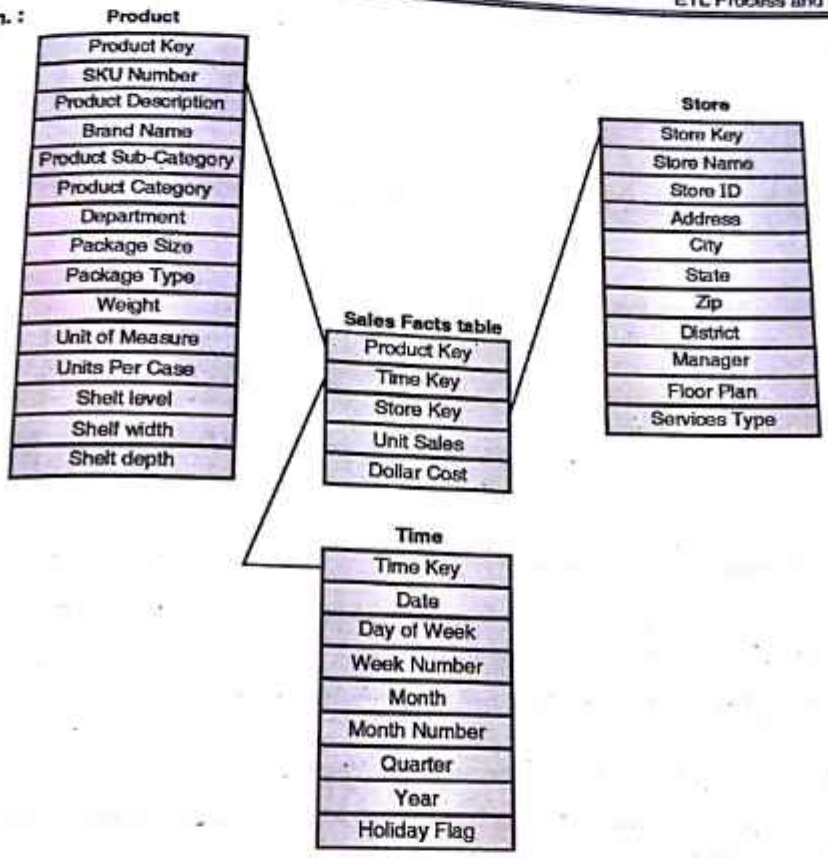


Fig. P. 2.17.2 : Star Schema for Electronics Company sales department

There are four tables, out of these 3 dimension tables and 1 fact table.

For OLAP operations refer Example 2.17.1.

Ex. 2.17.3 : The College wants to record the grades for the courses completed by students. There are four dimensions :

(i) Course (ii) Professor

(iii) Student (iv) Period

The only fact that is to be recorded in the table is course-grade.

(i) Design star schema.

(ii) Write DMQL for the above star schema

(iii) Using the above example describe the following OLAP operations

Slice, dice, roll-up, drill-down, Pivot. **MU - May 2011, 10 Marks**

Scanned by CamScanner

Soln. :

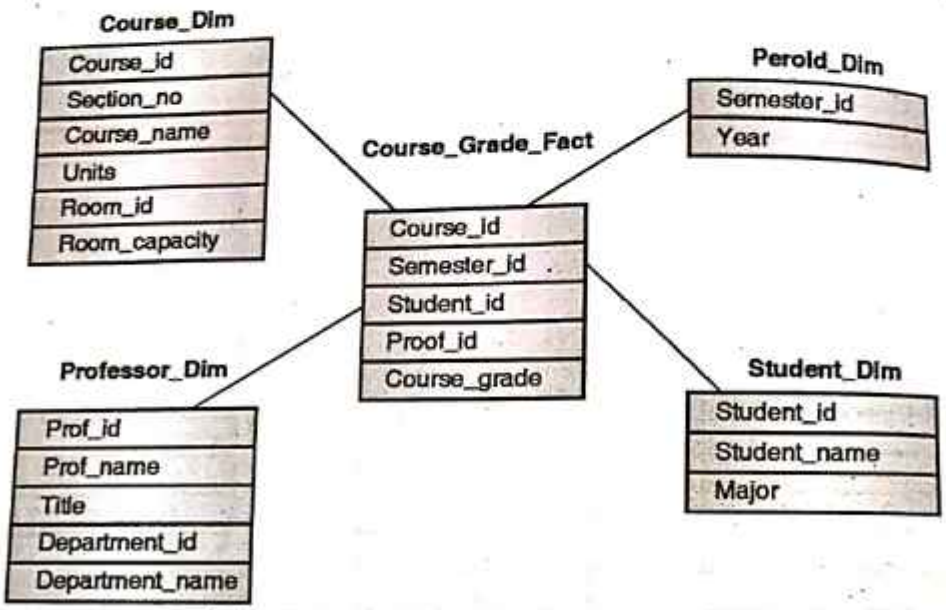


Fig. P. 2.17.3 : Star Schema for college

DMQL for the above star schema

- Define cube Course_Grade_Fact (Course_Dim,Period_D Proffesor_Dim,Student_Dim) : course-grade
- Define dimension Course_Dim as (Course_id, Section_id, Course_name, Un Room_id, Room_capacity)
- Define dimension Period_Dim as (Semister_id, Year)
- Define dimension Professor_Dim as (Prof_id, Prof_name, Title, Department_i Department_name)
- Define dimension Student_Dim as (Student_id, Student_name, Major)

For OLAP operations refer Example 2.17.1.

2.18 University Questions and Answers

May 2010

Q. 1 Explain ETL of data warehousing in detail.
(Ans. : Refer sections 2.5, 2.6, 2.7 and 2.8)

(10 Marks)

Dec. 2010

Q. 2 Explain ETL of data warehousing in detail.
(Ans. : Refer sections 2.5, 2.6, 2.7 and 2.8)

(10 Marks)

Scanned by CamScanner

Q. 3 Distinguish between OLAP and OLTP. *(Ans. : Refer section 2.12)* **(10 Marks)**

Q. 4 Explain different OLAP operations. *(Ans. : Refer section 2.15)* **(10 Marks)**

May 2011

Q. 5 Explain the major steps in the ETL process with a suitable diagram and an example. *(Ans. : Refer sections 2.5, 2.6, 2.7 and 2.8)* **(10 Marks)**

Q. 6 Compare between OLTP and OLAP. *(Ans. : Refer section 2.12)* **(5 Marks)**

Q. 7 The college wants to record the grades for the courses completed by students. There are four dimensions :

(i) Course (iii) Student

(ii) Professor (iv) Period.

(a) The only fact that is to be recorded in the table is course-grade :

(i) Design star schema. **(5 Marks)**

(ii) Write DMQL for the above star schema **(5 Marks)**

(b) Using the above example describe the following OLAP operations :

Slice, Dice, Roll-up, Drill-down, Pivot.

(Ans. : Refer Ex.2.17.3) **(10 Marks)**

Dec. 2011

Q. 8 Explain ETL of data warehousing in detail. *(Ans. : Refer sections 2.5, 2.6, 2.7 and 2.8)* **(10 Marks)**

May 2012

Q. 9 What is meant by ETL? Explain the ETL process in detail. *(Ans. : Refer sections 2.1, 2.2,2.5, 2.6, 2.7 and 2.8)* **(10 Marks)**

Q. 10 Compare OLTP and OLAP systems. *(Ans. : Refer section 2.12)* **(5 Marks)**

Q. 11 Write short notes on OLAP operations. *(Ans. : Refer section 2.15)* **(10 Marks)**

Dec. 2012

Q. 12 Explain the ETL cycle for a data warehouse in detail. *(Ans. : Refer sections 2.5, 2.6, 2.7 and 2.8)* **(10 Marks)**

Scanned by CamScanner

Q. 13 Consider a data warehouse storing sales details of various goods sold, and the time the sale, Using this example describe the following OLAP operations **(10 Marks)**

(1) Slice (2) Dice (3) Rollup (4) Drill down

(Ans. : Refer section 2.15)

May 2013

Q. 14 Describe the steps of the ETL (Extract - Transform - Load) cycle. **(10 Marks)**

(Ans. : Refer sections 2.5, 2.6, 2.7 and 2.8)

Q. 15 Consider a data warehouse for a hospital, where there are three dimensions:

(1) Doctor (2) Patient (3) Time; and two measures: (1) Count and (2) Fees

For this example create a OLAP cube and describe the following OLAP operations: **(10 Marks)**

(1) Slice (2) Dice (3) Rollup (4) Drill Down (5) Pivot *(Ans. : Refer Ex. 2.17.1)*

Dec. 2013

Q. 16 Explain ETL (Extract Transform Load) cycle in a Data Warehouse in detail. **(10 Marks)**

(Ans. : Refer sections 2.5, 2.6, 2.7 and 2.8)

Q. 17 All Electronic company have sales department sales, consider three dimensions namely.

(i) Time (ii) Product (iii) Store

The schema contains central fact table sales with two measures

(i) Dollars-cost and (ii) Units-Sold

Using the above example, describe the following OLAP operations

(i) Dice (ii) Slice

(iii) Roll-Up (iv) Drill-Down *(Ans. : Refer Ex. 2.17.2)* **(10 Marks)**

Q. 18 Compare between OLAP and OLTP. *(Ans. : Refer section 2.12)* **(10 Marks)**

May 2014

Q. 19 Describe clearly the different steps of the ETL (Extract - Transform - Load) cycle in data warehousing. *(Ans. : Refer sections 2.5, 2.6, 2.7 and 2.8)* **(10 Marks)**

Q. 20 Consider a data warehouse for a hospital, where there are three dimensions : (1) Doctor (2) Patient (3) Time ; and two measures : (1) Count and (2) Fees ; For this example create a OLAP cube and describe the following OLAP operations : (1) Slice (2) Dice (3) Rollup (4) Drill Down (5) Pivot. *(Ans. : Refer Ex. 2.17.1)* **(10 Marks)**

Scanned by CamScanner

Dec. 2014

Q. 21 Explain the ETL (Extract, Transform Load) cycle.
(Ans. : Refer sections 2.5, 2.6, 2.7 and 2.8) **(10 Marks)**

Q. 22 Describe the following OLAP operations using an example :

(1) Slice (2) Dice (3) Rollup

(4) Drill down (5) Pivot *(Ans. : Refer section 2.15)* **(10 Marks)**

May 2016

Q. 23 Describe the steps of ETL process.
(Ans. : Refer sections 2.5, 2.6, 2.7 and 2.8) **(10 Marks)**

Q. 24 We would like to view sales data of a company with respects three dimensions namely location, item and time. represent the sales data in the form of a 3-d data cube for the above and perform roll up, drill down. slice and dice OLAP operations on the above data cube and Illustrate. *(Ans. : Refer section 2.15)* **(10 Marks)**

Q. 25 Discuss various OLAP models. *(Ans. : Refer section 2.16)* **(10 Marks)**

❑❑❑

Chapter Ends

Scanned by CamScanner

Strictly as per the New Revised Syllabus (Rev - 2016) of
Mumbai University
w.e.f. academic year 2018-2019
(As per Choice Based Credit and Grading System)

SOFTW
ENGI

: CSC601)
er VI - C
, Brahm

ublications, Pune
roughout
Division

Module 3

CHAPTER 3

Introduction to Data Mining, Data Exploration and Preprocessing

Syllabus :

Introduction to Data Mining, Data Exploration and Preprocessing : Data Mining Task Primitives, Architecture, Techniques, KDD process, Issues in Data Mining, Applications of Data Mining, Data Exploration : Types of Attributes, Statistical Description of Data, Data Visualization, Data Preprocessing : Cleaning, Integration, Reduction : Attribute subset selection, Histograms, Clustering and Sampling, Data Transformation and Data Discretization : Normalization, Binning, Concept hierarchy generation, Concept Description: Attribute oriented Induction for Data Characterization.

3.1 What is Data Mining ?

→ (MU - Dec. 2011)

- Data Mining is a new technology, which helps organizations to process data through algorithms to uncover meaningful patterns and correlations from large databases that otherwise may not be possible with standard analysis and reporting.
- Data mining tools can help to understand the business better and also improve future performance through predictive analytics and make them proactive and allow knowledge driven decisions.
- Issues related to information extraction from large databases, data mining field brings together methods from several domains like Machine Learning, Statistics, Pattern Recognition, Databases and Visualization.
- Data mining field finds its application in market analysis and management like, for e.g. customer relationship management, cross selling, market segmentation. It can also be used in risk analysis and management for forecasting, customer retention, improved underwriting, quality control, competitive analysis and credit scoring.

Scanned by CamScanner

Definition

- Data mining is processing data to identify patterns and establish relationships.
- Data mining is the process of analysing large amounts of data stored in a data warehouse for useful information which makes use of artificial intelligence techniques, neural networks, and advanced statistical tools (such as cluster analysis) to reveal trends, patterns and relationships, which otherwise may be undetected.
- Data Mining is a non-trivial process of identifying :
  - Valid,
  - Novel,
  - Potentially useful, understandable patterns in data.

Syllabus Topic : Data Mining Task Primitives

3.2 Data Mining Task Primitives

Data mining primitives define a data mining task, which can be specified in the form of a data mining query.

1. Task Relevant Data
2. Kinds of knowledge to be mined
3. Background knowledge
4. Interestingness measure
5. Presentation and visualization of discovered patterns

1. Task relevant data

- Specify the data on which the data mining function to be performed.
- Using relational query, a set of task relevant data can be collected.
- Before data mining analysis, data can be cleaned or transformed.
- Minable view is created i.e. the set of task relevant data for data mining.

Scanned by CamScanner

2. The kind of knowledge to be mined

- Specify the knowledge to be mined.
- Kinds of knowledge include concept description, association, classification, prediction and clustering.
- User can also provide pattern templates. Also called metapatterns or metarules or metaqueries.

3. Background knowledge

- It is the information about the domain to be mined.
- Concept hierarchies is the form of background knowledge which helps to discover the knowledge at multiple levels of abstraction.

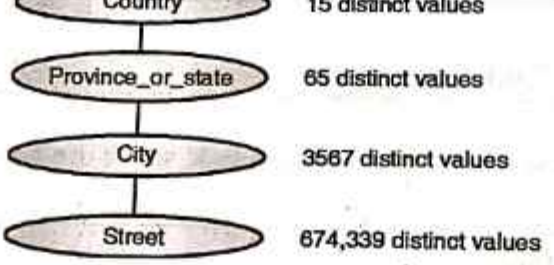


Fig. 3.2.1 : Concept hierarchy for the dimension location

Four major types of concept hierarchies

a) **Schema hierarchies :** It is the total or partial order among attributes in the database schema.

Example : Location hierarchy as street < city < province/state < country

b) **Set-grouping hierarchies :** It organizes values into sets or groups of constants.

Example : For attribute salary, a set-grouping hierarchy can be specified in terms of ranges as in the following :

{low, avg, high} € all (salary)

{1000....5000} € low

{5001....10000} € avg

{10001....15000} € high

Scanned by CamScanner

c) **Operation-derived hierarchies :** It is based on operation specified which may include decoding of information-encoded strings, information extraction from complex data objects, data clustering.

Example : URL or email address

xyz@cs.mu.in gives login name < dept. <univ.< country

d) **Rule-based hierarchies :** It occurs when either whole or portion of a concept hierarchy is defined as a set of rules and is evaluated dynamically based on current database data and rule definition.

Example :

Following rules are used to categorize items as *low_profit, medium_profit*and*high_profit_margin.*

low_profit_margin (Z) <= price (Z, A1)^cost (Z, A2)^((A1-A2)<80)

medium_profit_margin (Z) <= price (Z, A1)^cost (Z, A2)^((A1-A2) ≥80)^((A1-A2)≤350)

high_profit_margin (Z) <= price (Z, A1)^cost (Z, A2)^(A1-A2)>350)

4. Interestingness measures

- It is used to confine the number of uninteresting patterns returned by the process.
- Based on the structure of patterns and statistics underlying them.
- Each measure is associated a threshold which can be controlled by the user.
- Patterns not meeting the threshold are not presented to the user.
- Objective measures of pattern interestingness :
  - **Simplicity :** A patterns interestingness is based on its overall simplicity for human comprehension.
  - **Example :** Rule length is a simplicity measure.
  - **Certainty (confidence) :** Assesses the validity or trustworthiness of a pattern. Confidence is a certainty measure

$$\text{Confidence } (A \Rightarrow B) = \left(\frac{\text{(Number of tuples containing both A and B)}}{\text{(Number of tuples containing A)}} \right)$$

  - **Utility (support) :** It is the usefulness of a pattern support

$$(A \Rightarrow B) = \frac{\text{(Number of tuples containing both A and B)}}{\text{(total number of tuples)}}$$

  - **Novelty :** Patterns contributing new information to the given pattern set are called novel patterns (**Example :** Data exception).

Scanned by CamScanner

5. **Presentation and visualization of discovered patterns**

- Data mining systems should be able to display the discovered patterns in multiple forms, such as rules, tables, crosstabs (cross-tabulations), pie or bar charts, decision trees, cubes, or other visual representations.
- User must be able to specify the forms of presentation to be used for displaying the discovered patterns.

Syllabus Topic : Architecture

3.3 Architecture of a Typical Data Mining System

→ (MU - May 2010, Dec. 2010, Dec. 2011, Dec. 2013, May 2016)

Architecture of a typical data mining system may have the following major components as shown in Fig. 3.3.1 :

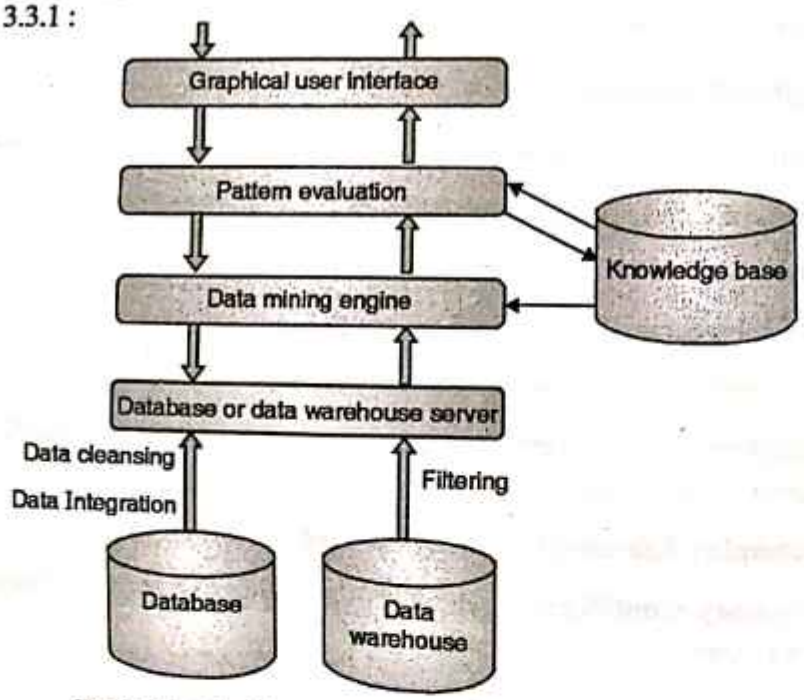


Fig. 3.3.1 : Architecture of typical data mining system

1. **Database, data warehouse, or other information repository**

These are information repositories. Data cleaning and data integration techniques may be performed on the data.

Scanned by CamScanner

2. **Databases or data warehouse server**

   It fetches the data as per the user's requirement which is need for data mining task.

3. **Knowledge base**

   This is used to guide the search, and gives interesting and hidden patterns from data.

4. **Data mining engine**

   It performs the data mining task such as characterization, association, classification, cluster analysis etc.

5. **Pattern evaluation module**

   It is integrated with the mining module and it helps in searching only the interesting patterns.

6. **Graphical user interface**

   This module is used to communicate between user and the data mining system and allow users to browse database or data warehouse schemas.

Syllabus Topic : Techniques

3.4 Data Mining Technique

→ **(MU - Dec. 2011)**

- Many techniques that strongly influence the development of data mining methods.
- Some of such technologies are given below :

3.4.1 Statistics

- Statistics is a discipline of science, which uses mathematical analysis to quantify representations, model and summarize empirical data or real world observations.
- Statistical analysis involves collection of methods and applying them on large amounts of data to draw conclusions and report the trend.

3.4.2 Machine Learning

- Machine learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. When new data is exposed, computer programs can teach themselves to grow or change due to machine learning.

Scanned by CamScanner

- For example, Facebook's News Feed changes according to the user's personal interactions with other users.

1. **Supervised learning :** One standard formulation of the supervised learning task is the classification problem. In classification, the training dataset is having labelled examples. Based on training dataset, classification model is constructed which is used to give label to unseen data. Neural networks and decision trees techniques are highly dependent on the previous information of classification where the classes are pre-determined by classifications techniques.

2. **Unsupervised learning :** One standard formulation of the unsupervised learning task is the Clustering problem where the examples of training dataset are not labelled. Based on similarity measures, clusters are formed.

3. **Semi-supervised learning :** This learning technique combines both labelled and unlabeled examples to generate an appropriate function or classifier. This method generally avoids the large number of labelled examples.

4. **Active learning :** The user play an important role in active learning. The algorithm itself decides which thing you should label. Active learning is a powerful approach in analysing data effectively.

3.4.3 Information Retrieval (IR)

Information Retrieval deals with uncertainty and vagueness in information systems :

- Uncertain representations of the semantics of objects (text, images,...).
- Vague specifications of information needs (iterative querying).
- **Example :** Find documents *relevant* to an information need from a large document set.

3.4.4 Database Systems and Data Warehouses

Database

1. Databases are used to record the data and also be used for data warehousing. Online Transactional Processing (OLTP) uses databases for day to day transaction purpose.
2. To remove the redundant data and save the storage space, data is normalized and stored in the form of tables.
3. Entity-Relational modeling techniques are used for Relational Database Management System design.

Scanned by CamScanner

4. Write operation on databases are optimized in databases.
5. Query writing is complex so performance is low for query analysis.

Data warehouse

1. Data Warehouses are used to store historical data which helps to take strategic decision for business. It is used for Online Analytical Processing (OLAP) which helps to analyze the data.
2. Data is de-normalized so tables are not complex and reduces the response time for analytical queries.
3. Data-modeling techniques like star schema are used for the Data Warehouse design.
4. Read operation on data warehouse are optimized as it has been used frequently for analysis purpose so performance is high for queries.

3.4.5 Decision Support System

- Decision Support system is a category of information systems, which helps in decision making for business and organisations.
- It is an interactive software based systems which helps decision makers to extract useful information from raw data, documents and take appropriate decisions, modelling business by identifying problems and solving them.
- Information that is gathered and presented by a DSS are :
  - A list of Information assets like legacy, relational data sources, cubes, data warehouse, data marts.
  - Comparative statements of sales.
  - Projected revenue based on assumptions of new product sales.

Syllabus Topic : KDD Process

3.5 Knowledge Discovery in Database (KDD)

→ (MU - May 2010, Dec. 2010, May 2011, May 2012, Dec. 2012, Dec. 2013, May 2016)

- The process of discovering knowledge in data and application of data mining methods refers to the term *Knowledge Discovery in Databases* (KDD).
- It includes a wide variety of application domains, which include Artificial Intelligence, Pattern Recognition, Machine Learning Statistics and Data Visualization.

Scanned by CamScanner

- The main goal includes extracting knowledge from large databases, the goal is achieve by using various data mining algorithms to identify useful patterns according to some predefined measures and thresholds.

Outline steps of the KDD process

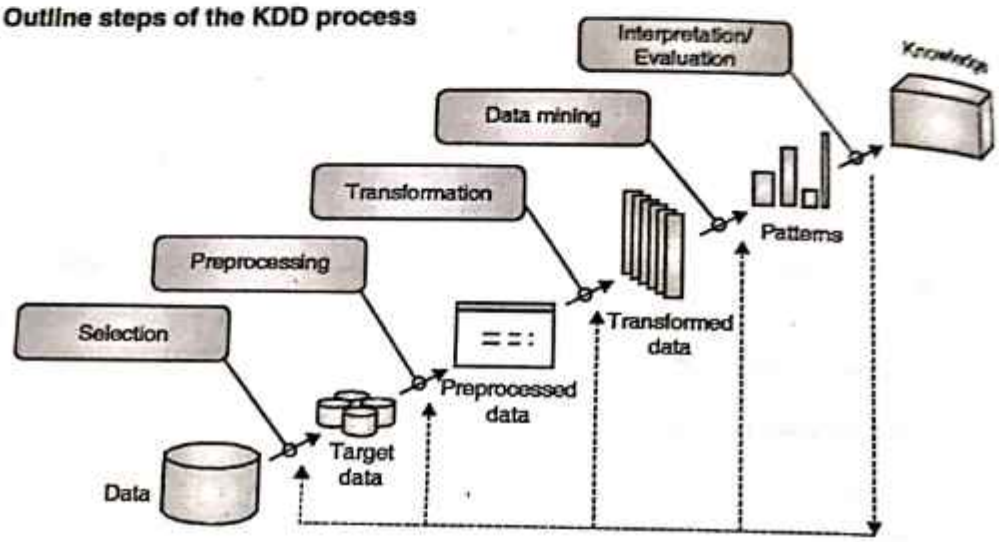


Fig. 3.5.1 : KDD Process

The overall process of finding and interpreting patterns from data involves the repeated application of the following steps :

1. **Developing an understanding of :**

   (a) The application domain

   (b) The relevant prior knowledge

   (c) The goals of the end-user.

2. **Creating a target data set :**

   Selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.

3. **Data cleaning and pre-processing :**

   (a) Noise or outliers are removed.

   (b) Essential information is collected for modelling or accounting for noise.

   (c) Missing data fields are handled by using appropriate strategies.

   (d) Time sequence information and changes are maintained.

Scanned by CamScanner

4. **Data reduction and projection :**

   (a) Based on the goal of the task, useful features are found to represent the data.

   (b) The number of variables may be effectively reduced using methods like dimensionality reduction or transformation. Invariant representations for the data may also be found out.

5. **Choosing the data mining task :**

   Selecting the appropriate Data mining tasks like classification, clustering, regression based on the goal of the KDD process.

6. **Choosing the data mining algorithm(s) :**

   (a) Pattern search is done using the appropriate Data Mining method(s).

   (b) A decision is taken on which models and parameters may be appropriate.

   (c) Considering the overall criteria of the KDD process a match for the particular data mining method is done.

7. **Data mining :**

   Using a representational form or other representations like classification, rules or trees, regression clustering for searching patterns of interest.

8. **Interpreting mined patterns**

9. **Consolidating discovered knowledge**

The terms *knowledge discovery* and *data mining* are distinct.

KDD	Data Mining
KDD is a field of computer science, which helps humans in extracting useful, previously undiscovered knowledge from data. It makes use of tools and theories for the same.	Data Mining is one of the step in the KDD process, it applies the appropriate algorithm based on the goal of the KDD process for identifying patterns from data.

Scanned by CamScanner

Syllabus Topic : Issues in Data Mining

3.6 Major Issues in Data Mining

Mining methodology and user interaction issues

- Mining different kinds of knowledge in database.
- Interactive mining of knowledge at multiple levels of abstraction.
- Incorporation of background knowledge.
- Data mining query language and ad hoc data mining.
- Presentation and visualization of data mining results.
- Handling noisy or incomplete data.
- Pattern Evaluation.

Performance issues

- Efficiency and scalability of data mining algorithms.
- Parallel, distributed and incremental mining algorithm.

Issues relating to the diversity of database types

- Handling of relational and complex types of data.
- Mining information from heterogeneous databases and global information system.

Syllabus Topic : Applications of Data Mining

3.7 Applications of Data Mining

→ **(MU - Dec. 2011)**

- Data Mining has been used in numerous areas, which include both private as well as public sectors.
- The use of Data mining in major industry areas like Banking, Retail, Medicine, insurance can help reduce costs, increase their sales and enhance research and development.
- For example in banking sector data mining can be used for customer retention, fraud prevention by credit card approval and fraud detection.
- Prediction models can be developed to help analyse data collected over years. For e.g. customer data can used to find out whether the customer can avail loan from the bank, or an accident claim is fraudulent and needs further investigation.

- Effectiveness of a medicin using data mining.
- Data mining can be used i for diseases, by analysing
- A large amount of data i may be collected for anal campaign effectiveness, more.
- Telecommunication indu the customer data which shift to competitors.

Syllabus

3.8 Types of Attrib

Data Objects

- A data object is a log the same entity. It als
- **Example :** In a prod store, employee, cust
- Every data object is database in the form attributes.

Attributes Types

- An attribute is a characteristic of a
- The attributes may

(i) Nominal attri
(ii) Binary attrib
(iii) Ordinal attri
(iv) Numeric att
(v) Discrete ve

Scanned by CamScanner

- Effectiveness of a medicine or certain procedure may be predicted in medical domain by using data mining.
- Data mining can be used in Pharmaceutical firms as a guide to research on new treatments for diseases, by analysing chemical compounds and genetic materials.
- A large amount of data in retail industry like purchasing history, transportation services may be collected for analysis purpose. This data can help multidimensional analysis, sales campaign effectiveness, customer retention and recommendation of products and much more.
- Telecommunication industry also uses data mining, for e.g. they may do analysis based on the customer data which of them are likely to remain as subscribers and which one will shift to competitors.

Syllabus Topic : Data Exploration - Types of Attributes

3.8 Types of Attributes

Data Objects

- A data object is a logical cluster of all tables in the data set which contains data related to the same entity. It also represents an object view of the same.
- **Example :** In a product manufacturing company, product, customer are objects. In a retail store, employee, customer, items and sales are objects.
- Every data object is described by its properties called as attributes and it is stored in the database in the form of a row or tuple. The columns of this data tuple are known to be attributes.

Attributes Types

- An attribute is a property or characteristic of a data object. For e.g. Gender is a characteristic of a data object person.
- The attributes may have values like :

(i) Nominal attributes

(ii) Binary attributes

(iii) Ordinal attributes

(iv) Numeric attributes

(v) Discrete versus continuous attributes

Scanned by CamScanner

(I) Nominal attributes

- Nominal attributes are also called as Categorical attributes and allow for only qualitative classification.
- Every individual item has a certain distinct categories, but quantification or ranking the order of the categories is not possible.
- The nominal attribute categories can be numbered arbitrarily.
- Arithmetic and logical operations on the nominal data cannot be performed.
- Typical examples of such attributes are :

Car owner :	1. Yes 2. No
Employment status :	1. Unemployed 2. Employed

(II) Binary attributes

- A nominal attribute which has either of the two states 0 or 1 is called Binary attribute, where 0 means that the attribute is absent and 1 means that it is present.
- **Symmetric binary variable :** If both of its states i.e. 0 and 1 are equally valuable. Here we cannot decide which outcome should be 0 and which outcome should be 1. For example : Marital status of a person is "Married or Unmarried". In this case both are equally valuable and difficult to represent in terms of 0(absent) and 1(present).
- **Asymmetric binary variable :** If the outcome of the states are not equally important. An example of such a variable is the presence or absence of a relatively rare attribute. For example : Person is "handicapped or not handicapped". The most important outcome is usually coded as 1 (present) and the other is coded as 0 (absent).

(III) Ordinal attributes

- A discrete ordinal attribute is a nominal attribute, which have meaningful order or rank for its different states.
- The interval between different states is uneven due to which arithmetic operations are not possible, however logical operations may be applied.
- For example, Considering Age as an ordinal attribute, it can have three different states based on an uneven range of age value. Similarly income can also be considered as an ordinal attribute, which is categorised as low, medium, high based on the income value.

(IV) Numeric attribute

Numeric Attributes
either have an intege

(a) Interval scaled
(b) Ratio scaled att

(a) Interval-scaled at

- Interval-scaled
- **Example :** v
ordering, com
scaled attribu

(b) Ratio-scaled attr

- Ratio scaled
They are also
- Operations l
are not possi
- **For exampl**
be 50 degre
a liquid at 2
- There are th
  - As inte
that it
  - As co
  - Trans
treatin

Scanned by CamScanner

Age :	1. Teenage 2. Young 3. Old
Income :	1. Low 2. Medium 3. High

(iv) Numeric attributes

Numeric Attributes are quantifiable. It can be measured in terms of a quantity, which can either have an integer or real value. They can be of two types :

(a) Interval scaled attributes

(b) Ratio scaled attributes

(a) Interval-scaled attributes

- Interval-scaled attributes are continuous measurement on a linear scale.
- **Example :** weight, height and weather temperature. These attributes allow for ordering, comparing and quantifying the difference between the values. An interval-scaled attributes has values whose differences are interpretable.

(b) Ratio-scaled attributes

- Ratio scaled attributes are continuous positive measurements on a non linear scale. They are also interval scaled data but are not measured on a linear scale.
- Operations like addition, subtraction can be performed but multiplication and division are not possible.
- **For example :** For instance, if a liquid is at 40 degrees and we add 10 degrees, it will be 50 degrees. However, a liquid at 40 degrees does not have twice the temperature of a liquid at 20 degrees because 0 degrees does not represent "no temperature"
- There are three different ways to handle the ratio-scaled variables :
  - As interval scaled variables. The drawback of handling them as interval scaled is that it can distort the results.
  - As continuous ordinal scale.
  - Transforming the data (for example, logarithmic transformation) and then treating the results as interval scaled variables.

Scanned by CamScanner

(v) Discrete versus continuous attributes

- If an attribute can take any value between two specified values then it is called a continuous else it is discrete. An attribute will be continuous on one scale and discrete on another.
- **For example :** If we try to measure the amount of water consumed by counting the individual water molecules then it will be discrete else it will be continuous.
- **Examples of continuous attributes** includes time spent waiting, direction of travel, water consumed etc.
- **Examples of discrete attributes** includes voltage output of a digital device, a person's age in years.

Syllabus Topic : Statistical Description of Data

3.9 Statistical Description of Data

- Statistical description of data is useful in finding the properties of data which can help identify in finding whether a particular data is noise or an outlier.
- Following are some of the measures which can be used for measuring the statistical properties of the data.

3.9.1 Central Tendency

- Central tendency is also known as measure of central location that describes the central position within the set of data. The various measures of central tendency are :

(i)	Mean	(ii)	Median
(iii)	Mode	(iv)	Midrange

- Based on conditions, a suitable measure is applied to calculate central tendency.

i) Mean

- It is one of the most common measure of central tendency. It can be used with both continuous and discrete attributes. It is mostly used with continuous attributes. Mean is equal to the sum of the all the values in a data set divided by the total number of values in the data set.

- Mean is given by $\bar{x}$ (

Or

Where, μ represents

- For example the me

85	55	8

- Mean is the only r
each value from m
- If the weights are
weighted average

- Mean has one di
such types of sit

ii) Median

- This measure is
skewed). As the
under such a s
values.
- Median is the
magnitude (sm
- For example,

85	55

Scanned by CamScanner

- Mean is given by $\bar{x}$ (pronounced as x bar)

$$\bar{x} = \frac{(x_1 + x_2 + \ldots + x_n)}{n}$$

Or

$$\mu = \frac{\Sigma x}{n}$$

Where, μ represents the mean.

- For example the mean for the following data is,

85	55	89	66	25	14	96	78	87	45	92

$$\mu = 64.7$$

- Mean is the only measure of central tendency in which the sum of the deviations of each value from mean is always zero.
- If the weights are associated with the value x_i then weighted arithmetic mean or the weighted average is,

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Mean has one disadvantage; it is highly susceptible to the influence of outliers. Under such types of situations median would be a better measure of central tendency.

ii) Median

- This measure is suitable when the data is skewed (frequency distribution of data is skewed). As the data becomes skewed, it loses its ability to provide a central position under such a situation median can be used as it is not influenced by the skewed values.
- Median is the middle score for a data set, which has been arranged in the order of magnitude (smallest first).
- For example, if we have the following data set :

85	55	89	66	25	14	96	78	87	45	92

Scanned by CamScanner

- Rearranging the above data in the order of magnitude :

14	25	45	55	66	78	85	87	89	92	96

- The median for the above data is 78. It is the middle position element as there are five scores before it and five scores after it.
- The above example is for odd number of elements in the data set, but in case if we have even number of elements for example :

14	25	45	55	66	78	85	87	89	92	96	100

- In the above case an average value of two values i.e. 78 and 85 is considered as median of the above data set. i.e. 81.5
- The median is a useful number in cases where the distribution has very large extreme values, which would otherwise skew the data.

$$\text{Median} = L_1 + \left(\frac{n/2 - (\Sigma f)l}{f_{median}}\right)C$$

Where, L_1 = Lower class boundary

n = Number of values in the data

$(\Sigma f)l$ = The sum of the frequencies of all of the classes that are lower than the median class

f_{median} = Frequency of median class

C = Size of the median class interval

iii) Mode

- The mode is the most frequently occurring value in the data set.
- If we consider a histogram then it is the highest bar in the chart. Mostly mode is used for categorical data, where the most common category is to be known.
- Data set with one mode is called unimodal.
- Data sets with two modes are called bimodal.
- Data sets with three modes are called trimodal.
- The empirical relation is :

$$\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})$$

Scanned by CamScanner

iv) Midrange

The midrange is the average of the largest and smallest value in a data set.

3.9.2 Dispersion of Data

The degree to which numeric data tend to spread is called the dispersion or variance of the data.

(i) Quartiles : Q_1 (25th percentile), Q_3 (75th percentile)

(ii) Inter-Quartile Range (IQR) : The distance between the first and third quartile is a simple measure of spread that gives the range covered by the middle half of the data.

$$IQR = Q_3 - Q_1$$

(iii) Five number summary : A fuller summary of the shape of a distribution can be obtained by providing highest and lowest data values. This is known as five number summary and written in order as,

$$\text{min}, Q_1, M, Q_3, \text{max}$$

Where, M = Median

Q_1, Q_3 = Quartile

Min = Smallest individual observation

Max = Largest individual observation

(iv) Box-plot

- The visual representation of a distribution is the box-plot.
- Box plots used for assigning location and variation information in data sets. It also detects and shows the location and variation changes between dissimilar groups of data.
- Box plots have Vertical axis which gives response variable and Horizontal axis which gives the factor of interest.
- To create box plot, order the data in ascending order.
- **Example :** 7,8,3,1,5,6,10,13,5

1	3	5	5	6	7	8	10	13

Scanned by CamScanner

- Calculate the median of that data which divides the data into two halves.

1	3	5	5	6	7	8	10	13

The median is $Q_2 = 6$

- Again find the median of those two halves which divides data into quarters.

1	3	5	5

7	8	10	13

- Calculate the median of both the halves as there are 4 values, the median is average of two middle values :

$$Q_1 = (3 + 5)/2 = 4$$

Similarly for second half,

$$Q_3 = (8 + 10)/2 = 9$$

- Now there are three points, Q_2 (the first middle point), Q_1, Q_3 (the middle point of two halves).
- These three points divides the whole data set into quarters which are called as "quartiles".
- The minimum value is 1 and the maximum value is 13, so we have :

  Min : 1, Q_1: 4, Q_2: 6, Q_3: 9, max : 13

- Then the box plot looks like this :

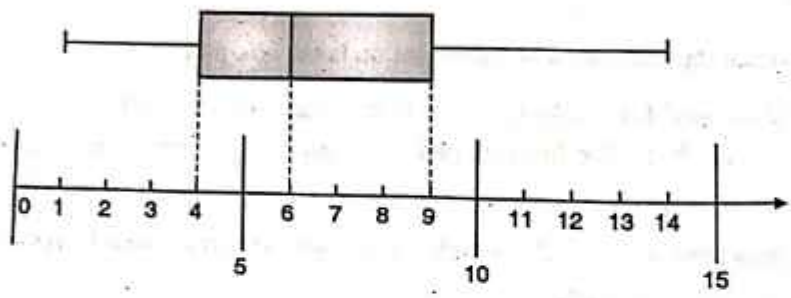


Fig. 3.9.1 : Box plot for dataset

(v) Outlier : Usually, a value higher/lower than 1.5 x IQR **(Inter-Quartile Range)**

(vi) Variance : The variance of n observations $x_1, x_2, \ldots\ldots, x_n$ is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

(vii) Standard deviation s is the

Measures spread about the me

3.9.3 Graphic Displays of B

(i) Bar charts, pie charts, line

(ii) Box-plots

(iii) Histograms

(iv) Quantile plots

(v) Quantile-Quantile plots (

(vi) Scatter plots

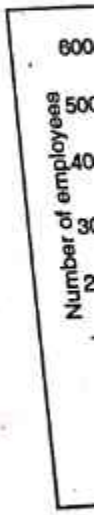
(vii) Loess curves

(I). Box-plots

The information about b

(II) Histograms

- Provide a first loo single attribute.
- Consists of a set present in the giv
- **Example :** Distr



Scanned by CamScanner

(vii) Standard deviation s is the square root of variance s^2

Measures spread about the mean. It is zero if and only if all the values are equal.

3.9.3 Graphic Displays of Basic Statistical Descriptions of Data

(i) Bar charts, pie charts, line graphs

(ii) Box-plots

(iii) Histograms

(iv) Quantile plots

(v) Quantile-Quantile plots (Q-Q plots)

(vi) Scatter plots

(vii) Loess curves

(i). Box-plots

The information about box-plots is already given in Section 3.9.2.

(ii) Histograms

- Provide a first look at a univariate data distribution, i.e., a distribution of values of a single attribute.
- Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data.
- **Example :** Distribution of salaries of the Acme Corporation

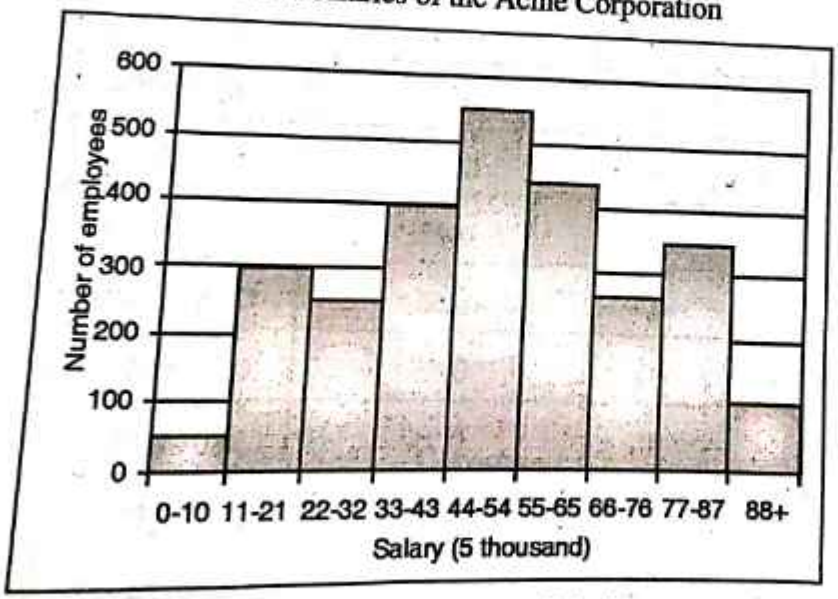


Fig. 3.9.2 : Histogram for Salaries

Scanned by CamScanner

- A vertical bar graph and a histogram differ in these ways :
  - In a histogram, frequency is measured by the area of the column.
  - In a vertical bar graph, frequency is measured by the height of the bar.

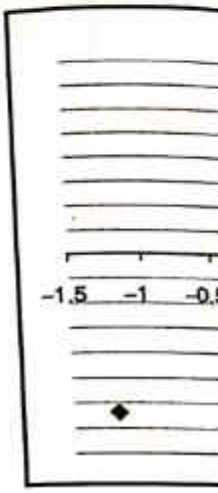
(iii) Quantile plots

- The normal quantile plot is used to compare the data values with the values that can be predicted for a standard normal distribution.
- To draw normal quantile plot, compute two additional numbers for each value of variable.

1. Sort the data value from lowest to highest and in the column labeled x(i) place the data values is ascending order.
2. Next column shows the quartile for each data value representation.
3. Now calculate the value of the standard normal distribution that lies at the quantile just computed in previous column.
4. A normal quantile plot is formed by plotting each data value against the value of the standard normal distribution.

Example :

i	x(i)	Quantile = i – 0.5/n	z = (x(i) – u)/σ
1	– 1.2	0.05	– 1.272108844
2	– 0.4	0.15	– 0.727891156
3	– 0.3	0.25	– 0.659863946
4	– 0.2	0.35	– 0.591836735
5	0	0.45	– 0.455782313
6	0.1	0.55	– 0.387755102
7	0.6	0.65	– 0.047619048
8	2.7	0.75	1.380952381
9	2.6	0.85	1.31292517
10	2.8	0.95	1.448979592



(iv) Quantile - Quantil

- Quantile-quanti
- As compared t
  but it needs mo
- One of the qua
- For example, (
- Order the ma
  y-axis values.
- If the marks a
  standard Norm
- As we have n
- Finally you p
- So a Q-Q pl
  measuremen

Scanned by CamScanner

mean μ = 0.67

StdDev σ = 1.471998

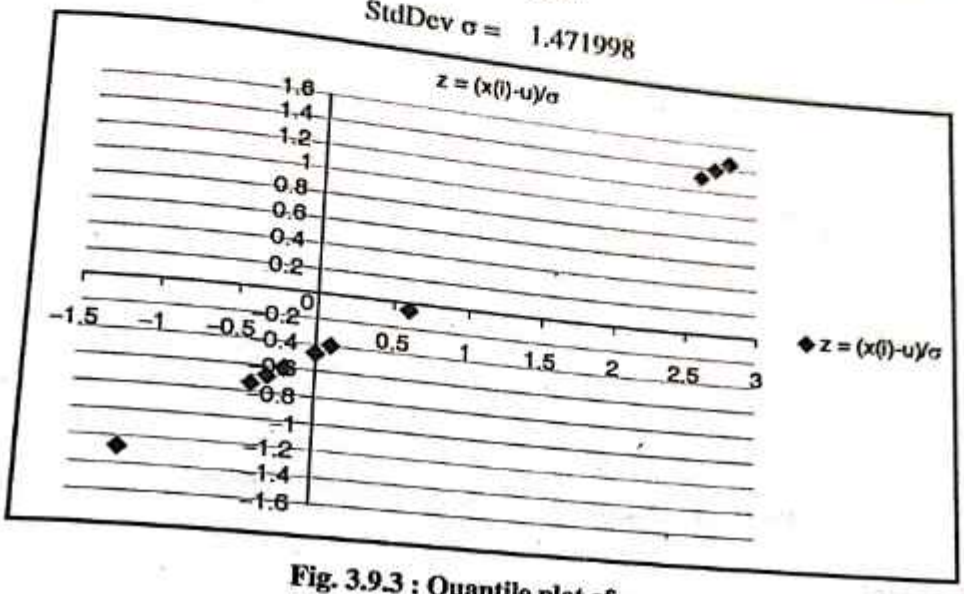


Fig. 3.9.3 : Quantile plot of z-score

(iv) Quantile - Quantile plots (Q-Q plots)

- Quantile-quantile plots are useful to compare the quantiles of two sets of numbers.
- As compared to mean and medians, the comparison using Q-Q plots is more detailed but it needs more samples for comparison.
- One of the quantiles is your sample observations placed in ascending order.
- For example, Consider marks of 30 students.
- Order the marks in ascending order from smallest to largest. That will be your y-axis values.
- If the marks are normally distributed, then x-axis values would be the quantiles of a standard Normal distribution.
- As we have marks of 30 students, we will need 30 Normal quantiles.
- Finally you plot marks versus the z-values.
- So a Q-Q plot is used to determine how well a theoretical distribution models a set of measurements.

Scanned by CamScanner

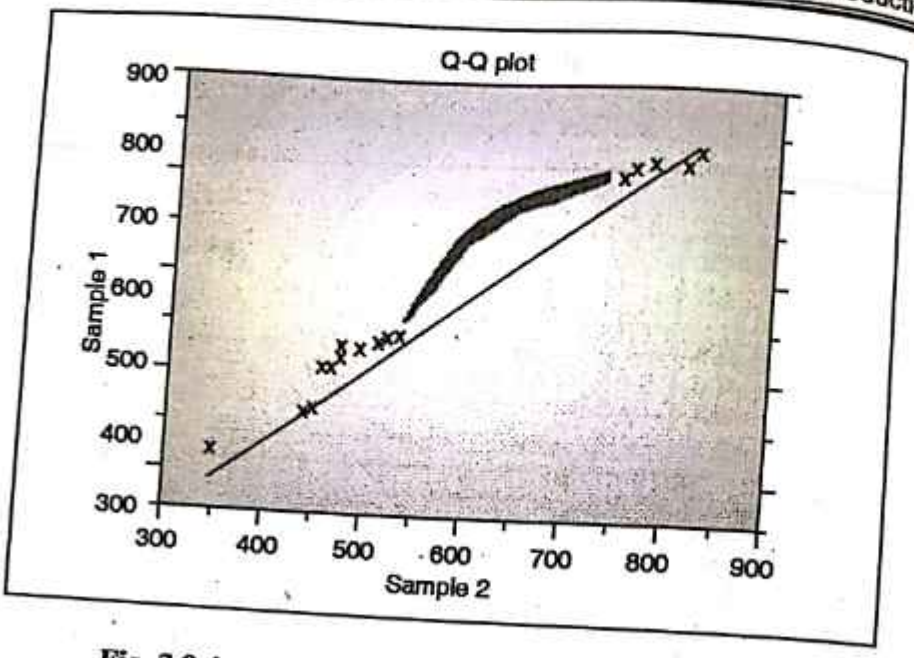


Fig. 3.9.4 : Quantile-Quantile plot for two samples

(v) Scatter plot

- Provide a first look at bivariate data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane.
- Scatter plots tell the relationships between two variables : Response variable is Y and it is usually on vertical axis.
- Variable which is related to response is X and it is on horizontal axis.
- **Example :** Assume that during a three-hour period spent outside, a person recorded the temperature and their water consumption. The experiment was conducted on 7 randomly selected days during the summer. The data is shown in the Table 3.9.1.

Table 3.9.1 : Temperature and Water consumption per day

Day	Temperature (F)	Water Consumption (oz)
1	99	48
2	85	27
3	97	48
4	75	16
5	92	32
6	85	25
7	83	[illegible]

Scanned by CamScanner

- Corresponding Scatter plot of Water Consumption based on Temperature is given in the Fig. 3.9.5 :

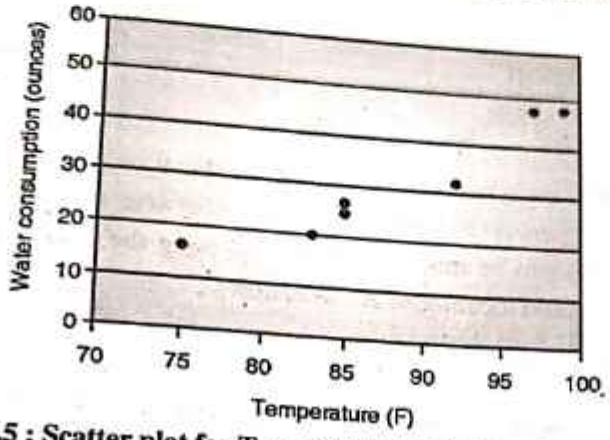


Fig. 3.9.5 : Scatter plot for Temperature and Water Consumption

(vi) Loess curve

- Loess curve is used for fitting a smooth curve between two variables.
- The procedure originated as LOWESS (Locally Weighted Scatter-plot Smoother).
- It is also called local regression.
- To get the better perception of pattern, a smooth curve is added in scatter plot as shown in Fig. 3.9.6.

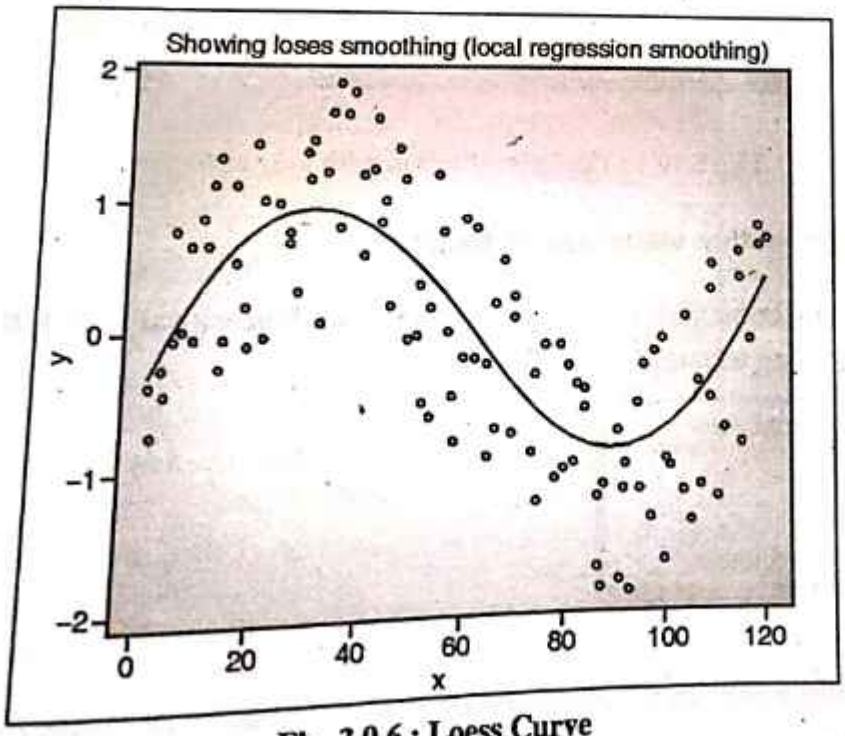


Fig. 3.9.6 : Loess Curve

Scanned by CamScanner

Syllabus Topic : Data Visualization

3.10 Data Visualization

→ (MU - Dec.

Data visualisation is presenting the data in a graphical or pictorial format. Visualisation techniques help people to analyse things which are otherwise not possible when the data is large. Patterns in the data can be marked very easily using the data visualisation techniques. Some of the data visualisation techniques are as follows :

1. Pixel-oriented visualization techniques

- In pixel based visualisation techniques, there is a separate sub windows for the values of each attribute and is represented by one colored pixel.
- It maximises the amount of information represented at one time without any overlap.
- A tuple with m variables has different m colored pixel to represent each variable and each variable has sub window.
- Based on data characteristics and visualization task, the color mapping of the pixels is decided.

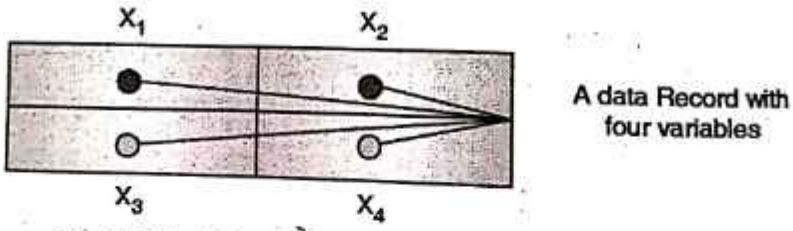


Fig. 3.10.1 : Pixel visualisation with four variables

2. Geometric projection visualization techniques

Geometric transformations and projections of multidimensional data sets can be found using the following techniques :

(i) Scatterplot matrices

(ii) Hyper slice

(iii) Parallel coordinates

(i) Scatterplot matrices : It is composed of scatter plots of all possible pairs of variables in a dataset.

Scanned by CamScanner

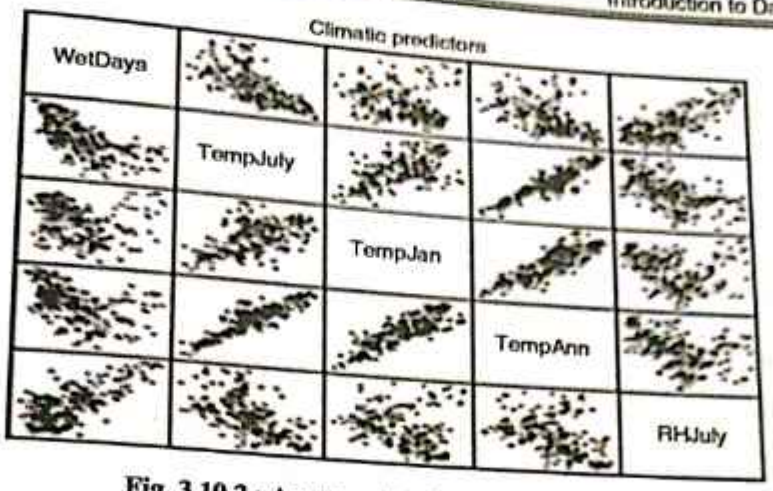


Fig. 3.10.2 : An example of scatter plot

(ii) Hyperslice : It is an extension to scatter plot matrices. They represent a multidimensional function as a matrix of orthogonal two dimensional slices.

(iii) Parallel co-ordinates : The parallel vertical lines separated define the axes. A point in the Cartesian coordinates corresponds to a polyline in parallel coordinates.

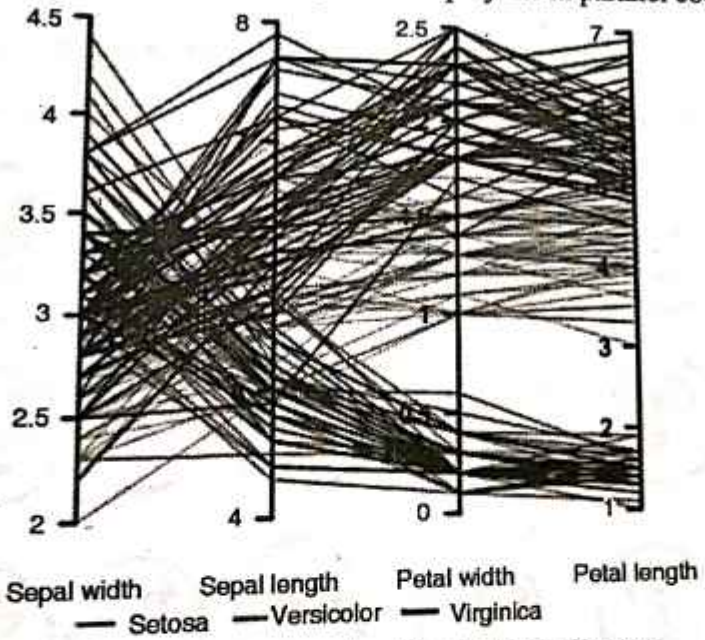


Fig. 3.10.3 : An example of Parallel coordinates

3. Icon-based visualization techniques

- Icon based visualisation techniques also known as iconic display techniques.

Scanned by CamScanner

- Each multidimensional data item is mapped to an icon.
- This technique allows visualisation of large amounts of data.
- Two most commonly used icon based techniques are :

(i) Chernoff faces

(ii) Stick figures

(i) Chernoff faces

- Illustration of trends in multidimensional data can be done by using Chernoff fa This concept was introduced by Herman Chernoff in the year 1973.
- The faces in Chernoff faces are related to facial expressions or features of hu being. So to distinguish between them is easy.
- Different data dimensions were mapped to different facial features, for example face width, the length or curvature of the mouth, the length of the nose etc.
- An example of Chernoff faces is shown below; they use facial features to repres trends in the values of the data, not the specific values themselves.
- They display multidimensional data of upto 18 variables or dimensions.
- In Fig. 3.10.4, each face represents an n-dimensional data points (n<=18).

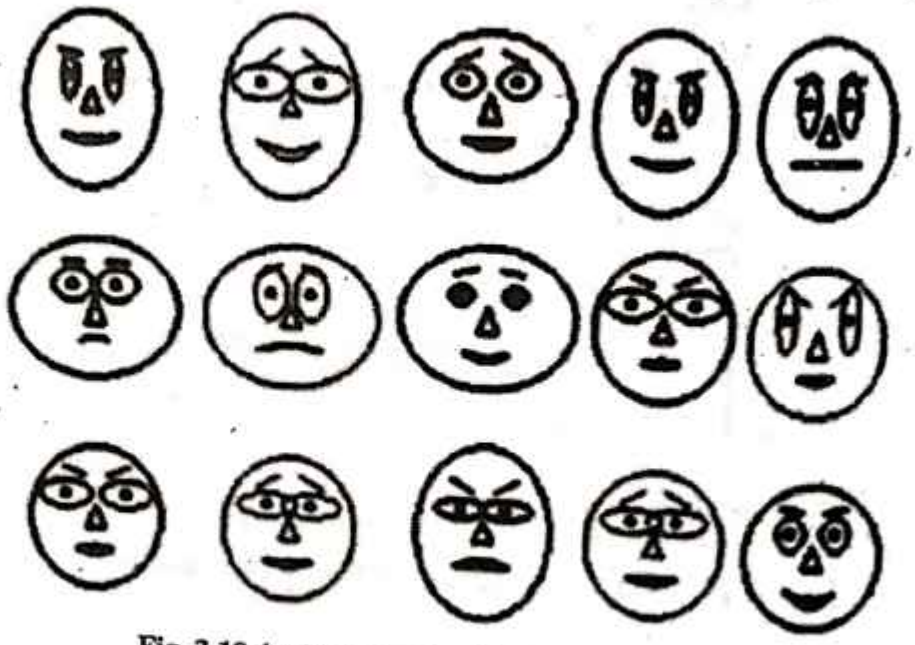
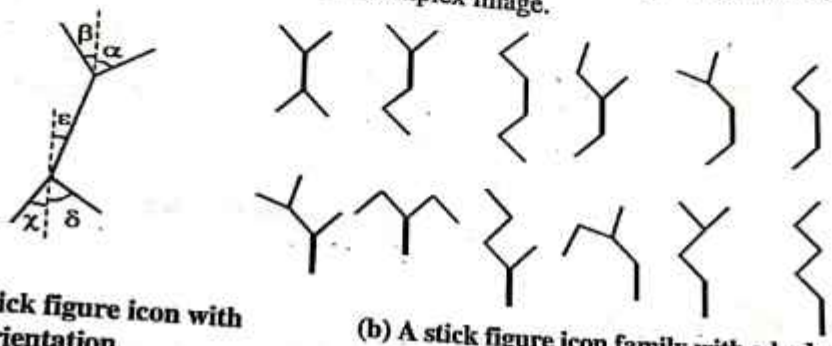


Fig. 3.10.4 : An example of Chernoff faces

Scanned by CamScanner

(ii) Stick figures

- Pickett and Grinstein introduced stick figure icon visualisation technique.
- The Fig. 3.10.5 represents the original stick figure with five stick and family of twelve of them.
- This icon family is designed to display data with up to five variates.
- Stick icon can be used to display bivariate MRI data by using a two stick icon which helps to differentiate the texture in a complex image.



(a) A five stick figure icon with orientation

(b) A stick figure icon family with a body and four limb

Fig. 3.10.5 : Example of stick figure

4. Hierarchical visualization techniques

- The visualisation techniques discussed above display multiple dimensions simultaneously. However for a large data set having large number of dimensions the above techniques may not be useful.
- Hierarchical visualisation techniques partition all dimensions in to subset (subspaces).
- These subspaces are visualised in a hierarchical manner.
- Some of the visualisation techniques are :

(i) Dimensional stacking

(ii) Mosaic Plot

(iii) Worlds-within-worlds

(iv) Tree-map

(v) Visualizing complex data and relations

Scanned by CamScanner

(i) Dimensional stacking

- In dimension stacking, partition the n-dimensional attribute space in 2-dimensio subspaces.
- Attribute values are partitioned into various classes.
- Each element is a two dimensional space is a xy plot.
- Mark the important attributes and are used on the outer levels.

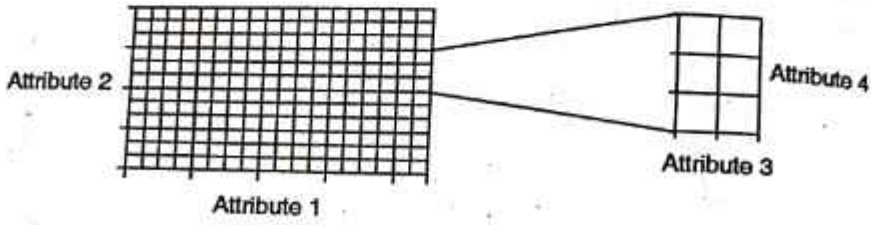


Fig. 3.10.6 : Data in dimension stacking

(ii) Mosaic plot

- Mosaic plots give a graphical illustration of the successive decompositions.
- Rectangles are used to represent the count of categorical data and at every stage rectangles are split parallel.
- To draw a mosaic plot, a contingency table of data and chosen ordering of variable with the response variable is required.
- **Example :** In titanic example , out of all women , 67% survived which is coded as 1 and 33% died which is coded as 0. So the women bar shows as 67/33 split. Among men, only 17% survived, so this bar shows a 17/83 split.

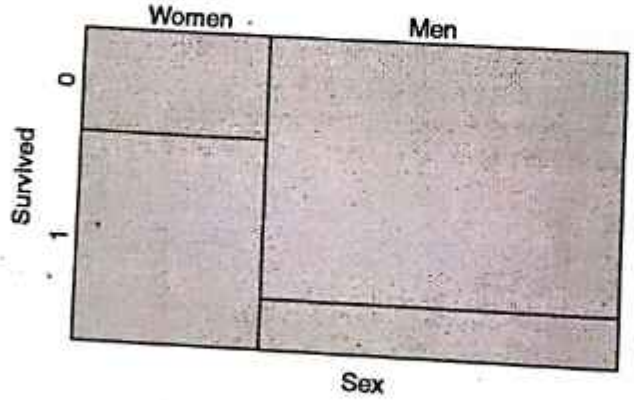


Fig. 3.10.7 : Mosaic Plot for Titanic

Scanned by CamScanner

(iii) Worlds within worlds

- Worlds within worlds are useful to generate an interactive hierarchy of display.
- Innermost world must have a function and two most important parameters.
- Remaining parameters fix with constant value.
- Through this N-vision of data are possible like data glove and stereo displays, including rotation, scaling (inner) and translation (inner/outer)
- Using queries static interaction is also possible.

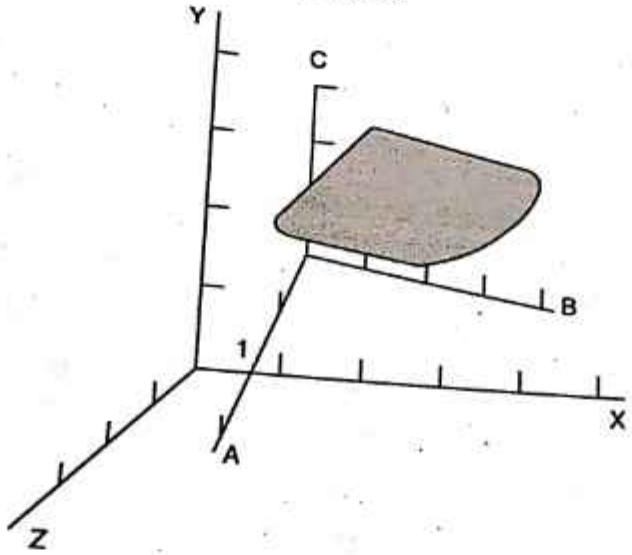


Fig. 3.10.8 : Worlds within worlds visualization

(iv) Tree-maps

- **Tree** maps visualization techniques are well suited for displaying large amounts of hierarchical structured data.
- The visualization space is divided into multiple rectangles that are sized and ordered according to a quantitative variable.
- The levels in the hierarchy are seen rectangles containing other rectangles.
- Each set of rectangles on the same level in the hierarchy represents a column or an expression in a data set.

Scanned by CamScanner

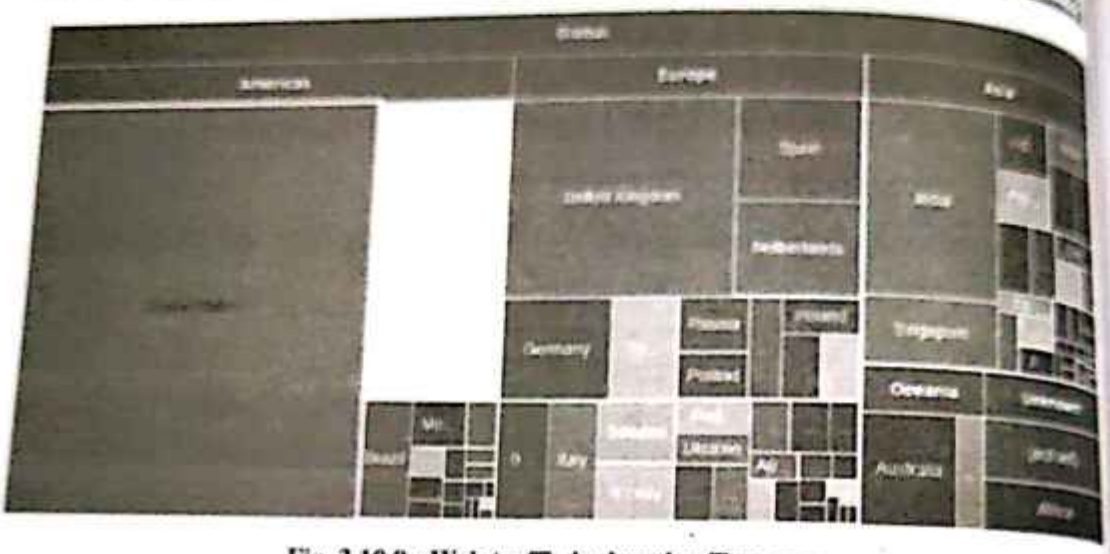


Fig. 3.10.9 : Web traffic by location Tree-map

- Each individual rectangle on a level in the hierarchy represents a category in a column.
- For example, in the Fig. 3.10.9, a rectangle representing global below which there are rectangles representing continents which contain several rectangles representing countries in that continent.
- Each rectangle representing a country may in turn contain rectangles representing states in these countries.

(v) Visualizing complex data and relations

- This technique is useful to visualize non-numeric data such as text, pictures, blog entries and product reviews.
- A tag cloud is a visualization method which helps to understand the information of user generated tags.
- Arrange the tags alphabetically or with the user preferences with different font sizes and colors.
- Tag clouds are used in two ways that with the size of tag, we can find out that how many times that tag is applied on that item by different users or that tag has been applied to how many items.

Scanned by CamScanner

Fig. 3.10.10 : Social data visualisation

Syllabus Topic : Data Preprocessing

3.11 Data Preprocessing

- Process that involves transformation of data into information through classifying, sorting, merging, recording, retrieving, transmitting, or reporting is called data processing. Data processing can be manual or computer based.
- In Business related world, data processing refers to data processing so as to enable effective functioning of the organisations and businesses.
- Computer data processing refers to a process that takes the data input via a program and summarizes, analyse the same or convert it to useful information.
- The processing of data may also be automated.
- Data processing systems are also known as information systems.
- When data processing does not involve any data manipulation and only converts the data type it may be called as data conversion.

3.11.1 Form of Data Pre-processing

Why Pre-processing is Required ?

1. **Real world data are generally**

   - **Incomplete :** The data is said to be incomplete when certain attributes or attributes values are missing or only aggregate data is available.

Scanned by CamScanner

- **Noisy :** When the data contains errors or some outliers it is considered to be noisy data.
- **Inconsistent :** When the data contains differences in codes or names it is inconsistent data.

2. **Tasks in data pre-processing**

- **Data cleaning :** This process consists of filling of missing values, smoothening noisy data, identifying and removing any outliers present and resolving inconsistencies.
- **Data integration :** This refers to integrating data from multiple sources like databases, data cubes, or files.
- **Data transformation :** Normalization and aggregation.
- **Data reduction :** In data reduction the amount of data is reduced but same analytical results are produced.
- **Data discretization :** Part of data reduction, replacing numerical attributes with nominal ones.

Different Forms of Data Pre-processing

1. Data cleaning
2. Data integration and transformation
3. Data reduction
4. Data discretization and Concept hierarchy generation

Syllabus Topic : Cleaning

3.12 Data Cleaning

→ **(MU - May 2016)**

Data cleaning is also known as scrubbing. The data cleaning process detects and removes the errors and inconsistencies and improves the quality of the data. Data quality problems arise due to misspellings during data entry, missing values or any other invalid data.

3.12.1 Reasons for "D

- Dummy values
- Absence of data
- Multipurpose fields
- Cryptic data
- Contradicting data
- Inappropriate use
- Violation of busi
- Reused primary
- Non-unique ider
- Data integration

Why data cleaning

- Source System
- Specialised too
- Some of the (Trillium) and

3.12.2 Steps in

1. **Parsing**
   - Parsing source
   - For exa the add
2. **Correctin**
   - This using
   - For e

Scanned by CamScanner

3.12.1 Reasons for "Dirty" Data

- Dummy values
- Absence of data
- Multipurpose fields
- Cryptic data
- Contradicting data
- Inappropriate use of address lines
- Violation of business rules
- Reused primary keys
- Non-unique identifiers
- Data integration problems

Why data cleaning or cleansing is required ?

- Source Systems data is not clean; it contains certain errors and inconsistencies.
- Specialised tools are available which can be used for cleaning the data.
- Some of the Leading data cleansing vendors include Vality (Integrity), Harte-Hanks (Trillium) and Firstlogic.

3.12.2 Steps in Data Cleansing

1. Parsing

- Parsing is a process in which individual data elements are located and identified in the source systems and then these elements are isolated in the target files.
- For example, parsing of name into First name, Middle name and Last name or parsing the address into street name, city, state and country.

2. Correcting

- This is the next phase after parsing, in which individual data elements are corrected using data algorithm and secondary data sources.
- For example, in the address attribute replacing a vanity address and adding a zip code.

Scanned by CamScanner

3. Standardizing

- In standardizing process conversion routines are used to transform data into a consistent format using both standard and custom business rules.
- For example, addition of a prename, replacing a nickname and using a preferred street name.

4. Matching

- Matching process involves eliminating duplications by searching and matching records with parsed, corrected and standardised data using some standard business rules.
- For example, identification of similar names and addresses.

5. Consolidating

Consolidation involves merging the records into one representation by analysing and identifying relationship between matched records.

6. Data cleansing must deal with many types of possible errors

- Data can have many errors like missing data, or incorrect data at one source.
- When more than one source is involved there is a possibility of inconsistency and conflicting data.

7. Data staging

- Data staging is an interim step between data extraction and remaining steps.
- Using different processes like native interfaces, flat files, FTP sessions, data is accumulated from asynchronous sources.
- After a certain predefined interval, data is loaded into the warehouse after the transformation process.
- No end user access is available to the staging file.
- For data staging, operational data store may be used.

.12.3 Missing Values

Issing data values

This involves searching for empty fields where values should occur.

Scanned by CamScanner

- Data preprocessing is one of the most important stages in data mining. Real world data is incomplete, noisy or inconsistent, this data is corrected in data preprocessing process by filling out the missing values, smoothening out the noise and correcting inconsistencies.
- There are several techniques for dealing with missing data, choosing one of them would be dependent on problems domain and the goal for data mining process.
- Following are the different ways for handle missing values in databases :

1. **Ignore the data row**

   - In case of classification suppose a class label is missing for a row, such a data row could be ignored, or many attributes within a row are missing even in this case data row could be ignored. If the percentage of such rows is high it will result in poor performance.
   - For example, suppose we have to build a model for predicting student success in college. For this purpose a student's database having information about age, score, address, etc. and column classifying their success in college to "LOW", "MEDIUM" and "HIGH". In this the data rows in which the success column is missing. These types of rows are of no use in the model therefore they can be ignored.

2. **Fill the missing values manually**

   This is not feasible for large data set and also time consuming.

3. **Use a global constant to fill in for missing values**

   - When missing values are difficult to be predicted, a global constant value like "unknown", "N/A" or "minus infinity" can be used to fill all the missing values.
   - For example, consider the students database, if the address attribute is missing for some students it does not makes sense in filling up these values rather a global constant can be used.

4. **Use attribute mean**

   - For missing values, mean or median of its discrete values may be used as a replacement.
   - For example, in a database of family incomes, missing values may be replaced with the average income.

Scanned by CamScanner

5. **Use attribute mean for all samples belonging to the same class**

- Instead of replacing the missing values by mean or median of all the rows in the database, rather we could consider class wise data for missing values to be replaced by its mean or median to make it more relevant.
- For example, consider a car pricing database with classes like "luxury" and "low budget" and missing values need to filled in, replacing missing cost of a luxury car with average cost of all luxury car makes the data more accurate.

6. **Use a data-mining algorithm to predict the most probable value**

- Missing values may also be filled up by using techniques like regression, inference based tools using Bayesian formalism, decision trees, clustering algorithms.
- For example, clustering method may be used to form clusters and then the mean or median of that cluster may be used for missing value. Decision tree may be used to predict the most probable value based on the other attributes.

3.12.4 Noisy Data

- A random error or variance in a measure variable is known as noise.
- Noise in the data may be introduced due to :
  - Fault in data collection instruments.
  - Error introduced at data entry by a human or a computer.
  - Data transmission errors.
- Different types of noise in data :
  - Unknown encoding : Gender : E
  - Out of range values : Temperature : 1004, Age : 125
  - Inconsistent entries : DoB : 10-Feb-2003; Age : 30
  - Inconsistent formats : DoB : 11-Feb-1984; DoJ : 2/11/2007

How to handle noisy data ?

Different data smoothing techniques are given below :

1. **Binning**

- Considering the neighbourhood of the sorted data smoothening can be applied.

Scanned by CamScanner

- The sorted data is placed into bins or buckets.
- Smoothing by bin means.
- Smoothing by bin medians.
- Smoothing by bin boundaries.

Different approaches of binning

(a) Equal-width (distance) partitioning

- Divides the range into *N* intervals of equal size : uniform grid.

$$\text{bin width} = (\text{max value} - \text{min value}) / N$$

- **Example :** Consider a set of observed values in the range from 0 to 100.
- The data could be placed into 5 bins as follows :

$$\text{width} = (100 - 0)/5 = 20$$

Bins formed are : [0-20], (20-40], (40-60], (60-80], (80-100]

The first and the last bin is extended to allow values outside the range :

(-infinity-20], (20-40], (40-60], (60-80], (80-infinity)

Disadvantages

1. Outliers in the data may be a problem.
2. Skewed data cannot be held with this method.

(b) Equal-depth (frequency) partitioning or Equal-height binning

- The entire range is divided into N intervals, each containing approximately the same number of samples.
- This results in good data scaling.
- Handling categorical attributes may be a problem.

Example :

- Let us consider sorted data for e.g. Price in INR

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- Partition into (equal-depth) bins : (N=3)

Bin 1 : 4, 8, 9, 15

Scanned by CamScanner

Bin 2 : 21, 21, 24, 25

Bin 3 : 26, 28, 29, 34

- Smoothing by bin means :

Replace each value of bin with its mean value.

Bin 1 : 9, 9, 9, 9

Bin 2 : 23, 23, 23, 23

Bin 3 : 29, 29, 29, 29

- Smoothing by bin boundaries :

In this method the minimum and maximum values of the bean boundaries is found and each value is replaced with its nearest value either minimum or maximum.

Bin 1 : 4, 4, 4, 15

Bin 2 : 21, 21, 25, 25

Bin 3 : 26, 26, 26, 34

2. Outlier analysis by clustering

- Partition data set into clusters and one can store cluster representation only, i.e. replace all values of the cluster by that one value representing the cluster.
- Outliers can be detected by using clustering techniques, where related values are organized into groups or clusters.

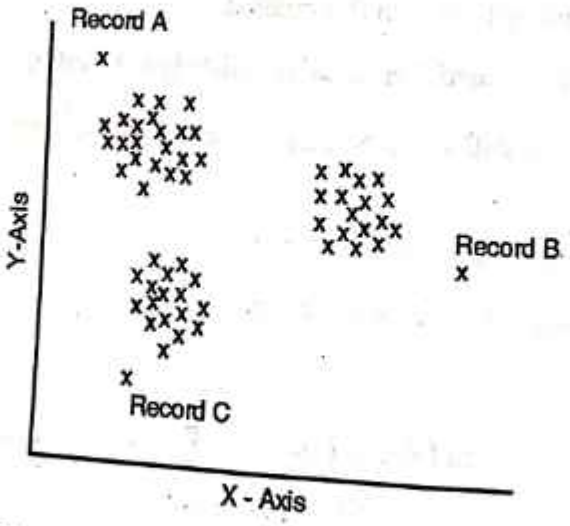


Fig. 3.12.1 : Graphical Example of Clustering

Scanned by CamScanner

- Perform clustering on attributes values and replace all values in the cluster by a cluster representative.

3. Regression

- Regression is a statistical measure used to determine the strength of the relationship between one dependent variable denoted by Y and a series of independent changing variables.
- Smooth by fitting the data into regression functions.
- Use regression analysis on values of attributes to fill missing values.
- The two basic types of regression are linear regression and multiple regressions.
- The difference between Linear and multiple regressions is that former uses one independent variable to predict the outcome, while the later uses two or more independent variables to predict the outcome.
- The general form of each type of regression is :

Linear Regression : $Y = a + bX + u$

Multiple Regression : $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + u$

Where,

Y = The variable that we are trying to predict

X = The variable that we are using to predict Y

a = The intercept

b = The slope

u = The regression residual.

- In multiple regressions each variable is differentiated with subscripted numbers.
- Regression uses a group of random variables for prediction and finds a mathematical relationship between them. This relationship is depicted in the form of a straight line (Linear regression) that approximates all the points in the best way.
- Regression may be used to determine for e.g. price of a commodity, interest rates, the price movement of an asset influenced by industries or sectors.

Scanned by CamScanner

Log linear model

- In Log linear regression a best fit between the data and a log linear model is found.
- Major assumption : A linear relationship exists between the log of the dependent and independent variables.
- Log linear models are models that postulate a linear relationship between the independent variables and the logarithm of the dependent variable. For example :
  $\log(y) = a_0 + a_1 x_1 + a_2 x_2 \ldots + a_N x_N$

  where y is the dependent variable; x_i, i=1,...,N are independent variables and $\{a_i, i=0,\ldots,N\}$ are parameters (coefficients) of the model.
- For example, log linear models are widely used to analyze categorical data represented as a contingency table. In this case, the main reason to transform frequencies (counts) or probabilities to their log-values is that, provided the independent variables are not correlated with each other, the relationship between the new transformed dependent variable and the independent variables is a linear (additive) one.

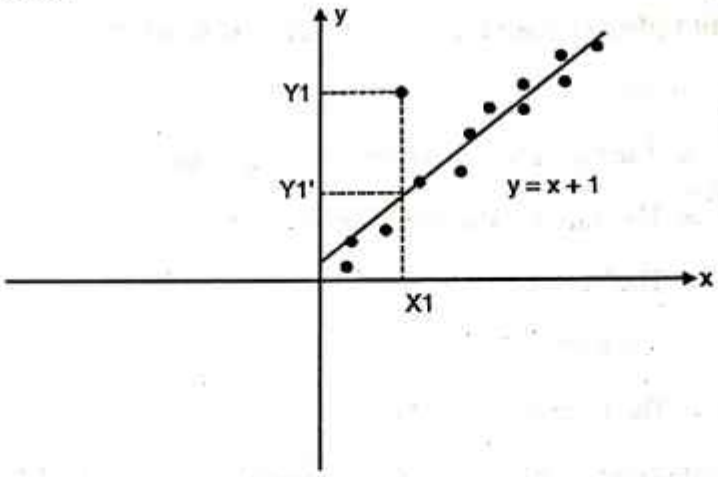


Fig. 3.12.2 : Regression example

3.12.5 Inconsistent Data

- The state in which the data quality of the existing data is understood and the desired quality of the data is known refers to consistent data quality.
- It is a state in which the existent data quality is been modified to meet the current and future business demands.

Scanned by CamScanner

Syllabus Topic : Integration

3.13 Data Integration

→ (MU - May 2016)

A coherent data store (e.g. a Data warehouse) is prepared by collecting data from multiple sources like multiple databases, data cubes or flat files.

Issues in data integration

1. **Schema integration**
   - Integrate metadata from different sources.
   - Entity identification problem : identify real world entities from multiple data sources. e.g. A.cust-id ≡B.cust-#.
2. **Detecting and resolving data value conflicts**
   - As the data is collected from multiple sources, attribute values are different for the same real world entity.
   - Possible reasons include different representations, different scales, e.g. metric vs. British units.
3. **Redundant data occur due to integration of multiple databases**
   - Attributes may be represented in different names in different sources of data.
   - An attribute may be derived attribute in another table, e.g. yearly income.
   - With the help of co-relational analysis, detection of redundant data is possible.
   - The redundancies or inconsistencies may be reduced by careful integration of the data from multiple sources, which will help in improving mining speed and quality.

3.13.1 Entity Identification Problem

- Schema integration is an issue as to integrate metadata from different sources is a difficult task.
- Identify real world entities from multiple data sources and their matching is the entity identification problem. For example, Roll number in one database and enrolment number in another database refers to the same attribute.
- Such conflicts may create problem for schema integration.

Scanned by CamScanner

- Detecting and resolving data value conflicts for the same real world entity, attribute values from different sources are different.

3.13.2 Redundancy and Correlation Analysis

- Data redundancy occurs when data from multiple sources is considered for integration.
- Attribute naming may be a problem as same attributes may have different names in multiple databases.
- An attribute may be derived attribute in another table e.g. "yearly income".
- Redundancy can be detected using correlation analysis.
- To reduce or avoid redundancies and inconsistencies data integration must be carried out carefully. This will also improve mining algorithm speed and quality.
- X^2 (*Chi-square*) test can be carried out on nominal data to test how strongly the two attributes are related.
- Correlation coefficient and covariance may be used with numeric data, this will give a variation between the attributes.

The X^2 (Chi-square)

- It is used to test hypotheses about the shape or proportions of a population distribution by means of sample data.
- For nominal data, a correlation relationship between two attributes, P and Q, can be discovered by an X^2 (Chi-square) test.
- These nominal variables, also called "attribute variables" or "categorical variables", classify observations into a small number of categories, which are not numbers. It doesn't work for numeric data.
- Examples of nominal variables include Gender (the possible values are male or female), Marital Status (Married, unmarried or divorced), etc.
- The Chi-square test is used to test the probability of independence of a distribution of data but does not gives you any details about the relationship between them.
- Chi-square test is defined by,

$$X^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

Where X^2

E

O

- **Degrees of free**

DF

where, r is the n
for the other cat

- **Expected freq**
attribute. The fo

$E_{r,c}$

  - Where $Er_{,}$ attribute Y
  - n_r is the su
  - n_c is the su
  - n is the to

Ex. 3.13.1 : By t
data
(You
Tab

Co

Le

Soln. : State th

1. H_0 : Null hy
2. H_a : Alterna

Scanned by CamScanner

Where X^2 = Chi-square

E = Frequency expected which is the amount of subjects that you would *expect* to find in each category based on known information.

O = Frequency observed which is the amount of subjects you *actually* found to be in each category in the present data.

- **Degrees of freedom :** The degrees of freedom (DF) is equal to :

$$DF = (r - 1) * (c - 1)$$

where, r is the number of levels for one categorical variable and c is the number of levels for the other categorical variable.

- **Expected frequencies :** It is the count which is computed for each level of categorical attribute. The formula for expected frequency is

$$E_{r,c} = (n_r * n_c) / n$$

- Where $Er_{,c}$ is the expected frequency count for level r of attribute X and level c of attribute Y,
- n_r is the sum of sample observations at level r of attribute X,
- n_c is the sum of sample observations at level c of attribute Y,
- n is the total size of sample data.

Ex. 3.13.1 : By taking the random sample of 1000 buyers, a survey was conducted. Sample data were classified by gender (male or female) and by buying preference (Young Age, Middle Age, or Old Age). Results are shown in the contingency Table P. 3.13.1

Table P. 3.13.1

	Buying Preferences			Row
	Young Age	Middle Age	Old Age	Total
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

Level of significance is 0.05. Interpret the result.

Soln. : State the hypotheses

- H_0 : Null hypothesis : Gender and buying preferences are independent.

H_a : Alternative hypothesis : Gender and buying preferences are not independent.

Scanned by CamScanner

Analyze sample data

- The degrees of freedom DF is,

$$DF = (r-1) * (c-1) = (2-1) * (3-1) = 2$$

$r = 2$ (the number of levels for row i.e. male and female)

$c = 3$ (the number of levels for column categorical variable

i.e. Young age, Middle age and Old age.)

- By using given contingency table, calculate the expected frequency $E_{r,c}$ for each level $E_{r,c} = (n_r * n_c) / n$.

	Buying preferences			Row total
	Young age	Middle age	Old age	
Male (O)	200	150	50	400
Male(E)	(400 * 450) / 1000 = 180	(400 * 450) / 100 = 180	(400 * 100) / 1000 = 40	400
Male(O-E)	20	−30	10	
Male(O-E)²	400	900	100	
Male(O-E)²/E	2.22	5	2.5	
Female(O)	250	300	50	600
Female(E)	(600 * 450) / 1000 = 270	(600 * 450) / 1000 = 270	(600 * 100) / 1000 = 60	600
Female(O-E)	−20	30	−10	
Female(O-E)²	400	900	100	
Female (O-E)²/E	1.48	3.33	1.67	

- Using formula,

$$X^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

$$X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$$

- The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 16.2.
- We use the Chi-Square distribution table to find $P(X^2 > 16.2) = 0.0003$.

Results Interpretation

The significance level
than the significance
interpret that there is a

The correlation coeffic

- The correlation coeff

r_{pq} =

Where n is the num
are the respective
product.

- If $r_{pq} > 0$, p and q
stronger correlatio
- Correlation meas
objects, x and y a

- Covariance is g

Cov (X,Y)

3.13.3 Tuple Du

- In data integr
occur due to i
- Detection of t
  - Specify t
  - Compar
  - Objects
  - The clo

Scanned by CamScanner

Results Interpretation

The significance level given in problem statement is 0.05 and the P-value (0.0003) is less than the significance level (0.05). So, the null hypothesis cannot be accepted and we can interpret that there is a relationship between gender and buying preference.

The correlation coefficient and covariance

- The correlation coefficient is given by,

$$r_{p \cdot q} = \frac{\Sigma (p-\bar{p})(q-\bar{q})}{n \cdot \sigma_p \sigma_q} = \frac{\Sigma ((pq) - n\,\bar{p}\,\bar{q})}{n \cdot \sigma_p \sigma_q}$$

Where n is the number of tuples, $\bar{p}$ and $\bar{q}$ are the respective means of p and q, σ_p and σ_p are the respective standard deviation of p and q and Σ (pq) is the sum of the pq cross-product.

- If $r_{p,q} > 0$, p and q are positively correlated (p's values increase as q's). The higher, the stronger correlation. $r_{p,q} = 0$: independent ; $r_{pq} < 0$: negatively correlated.
- Correlation measures the linear relationship between objects. First standardize data objects, x and y and then calculate the correlation between them by taking dot product.

$$X'_k = (X_k - \text{mean}(X))/\text{std}(X)$$

$$Y'_k = (Y_k - \text{mean}(Y))/\text{std}(Y)$$

$$\text{Correlation}(X,Y) = X' \cdot Y'$$

- Covariance is given by

$$\text{Cov}(X,Y) = E((X-\bar{X})(Y-\bar{Y})) = E(X \cdot Y) - \bar{X}\,\bar{Y}$$

3.13.3 Tuple Duplication

- In data integration tuple duplication should be checked. This type of duplication may occur due to incorrect data entry or updation of data.
- Detection of tuple duplication :
  - Specify the relevant attributes.
  - Compare tuples pair wisely using a similarity measure.
  - Objects with similarity above a given threshold are considered as duplicates.
  - The closures of duplicates are computed, in which each is assigned one <sourceID>.

Scanned by CamScanner

3.13.4 Data Value Conflict Detection and Resolution

- Data conflicts can arise because of incomplete data, invalid data and out-of-date data.
- It is thus critical for data integration systems to resolve conflicts from various sources an identify true values from false ones.
- Two kinds of data conflicts :

  1. Uncertainty 2. Contradiction

Uncertainty	Contradiction
It is an attribute level data conflict.	It is an attribute level data conflict.
If there is missing information or null values, then Uncertainties occurs.	If there is contradicting information or different attribute values, then contradiction occurs.
If information for the attribute value is not available, then uncertainty is introduced.	When data is collected from various sources, then contradiction is introduced for the same properties of the same objects.
If the set of values includes special NULL value and one other NON NULL value, then uncertainty is present.	If at least two different non-null values appear in the set of values, then contradiction is present.
For age attribute, if the set of values for customer1 are{30,Null,30}={30,Null}. Then there is uncertainty as there is one Null value for age.	For age attribute, if the set of values for customer1 are {30,35,30}={30,35}. Then contradiction is there, as there are two different values for age.

Syllabus Topic : Data Reduction - Attribute Subset Selection, Histograms, Clustering and Sampling

3.14 Data Reduction

➔ (MU - May 2016)

3.14.1 Need for Data Reduction

1. Reducing the number of attributes

- **Data cube aggregation :** This process involves applying OLAP operations like roll-up, slice or dice operations.

- **Removing irrelevan** wrapper methods ma
- **Principle compon** representing the data

2. Reducing the number

- **Binning (histogran** bins, this will result
- **Clustering :** Grou clusters.
- **Aggregation or ge**

3. Reducing the numbe

To reduce the number o

3.14.2 Data Reduction

Following are the data

(A) Data cube aggrega
(B) Dimensionality re
(C) Data compressior
(D) Numerosity redu

3.14.2(A) Data Cube

- It reduces the data to detailed) level neces
- Queries regarding a possible.

Example

Total annual sales o

Scanned by CamScanner

- **Removing irrelevant attributes :** In this attribute selection methods like filtering and wrapper methods may be used, it also involves searching the attribute space.
- **Principle component analysis (numeric attributes only) :** This involves representing the data in a compact form by using a lower dimensional space.

2. **Reducing the number of attribute values**
   - **Binning (histograms) :** This involves representing the attributes into groups called as bins, this will result into lesser number of attributes.
   - **Clustering :** Grouping the data based on their similarity into groups called as clusters.
   - **Aggregation or generalization.**
3. **Reducing the number of tuples**

   To reduce the number of tuples, sampling may be used.

3.14.2 Data Reduction Technique

Following are the data reduction techniques :

(A) Data cube aggregation

(B) Dimensionality reduction

(C) Data compression

(D) Numerosity reduction

3.14.2(A) Data Cube Aggregation

- It reduces the data to the concept level needed in the analysis and uses the smallest (most detailed) level necessary to solve the problem.
- Queries regarding aggregated information should be answered using data cube when possible.

Example

Total annual sales of TV in USA is aggregated quarterly as shown in Fig. 3.14.1.

Scanned by CamScanner

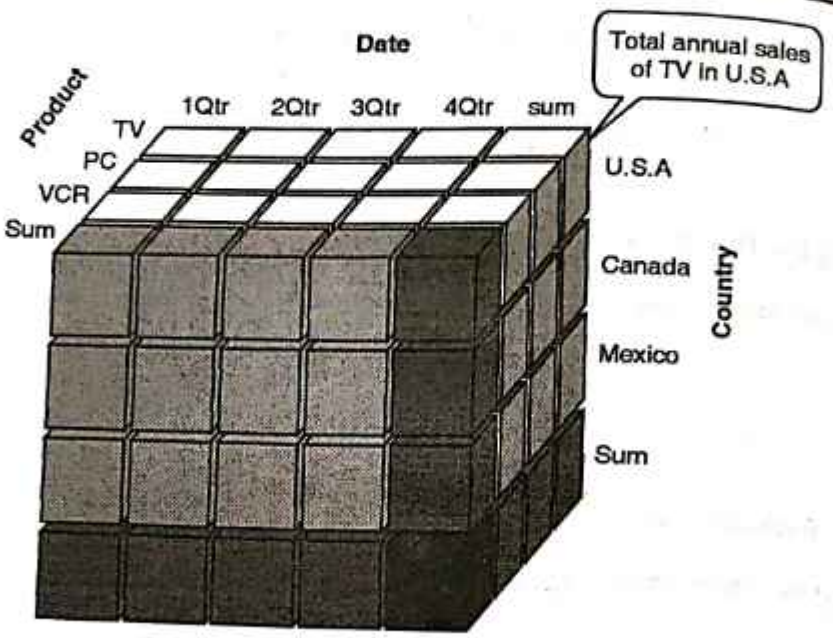


Fig. 3.14.1 : Example of data cube

3.14.2(B) Dimensionality Reduction

- In the mining task during analysis, the data sets of information may contain large number of attributes that may be irrelevant or redundant.
- Dimensionality reduction is a process in which attributes are removed and the resulting dataset is smaller in size.
- This process helps in reducing the time and space complexity required by a data mining technique.
- Data visualization becomes an easy task.
- It also involves deleting inappropriate features or reducing the noisy data.

Attribute subset selection

How to find a good subset of the original attributes ?

Attribute subset selection refers to a process in which minimum set of attributes are selected in such a way that their distribution represents the same as the original data set distribution considering all the attributes.

Scanned by CamScanner

Different attribute subset selection techniques

1. **Forward selection**
   - Start with empty set of attributes.
   - Determine the best of the original attributes and add it to the set.
   - At each step, find the best of the remaining original attributes and add it to the set.
2. **Stepwise backward elimination**
   - Starts with the full set of attributes.
   - At each step, it removes the worst attribute remaining in the set.
3. **Combination of forward selection and backward elimination**
   - The procedure combines and selects the best attribute and removes the worst among the remaining attributes.
   - For all above method stopping criteria is different and it requires a threshold on the measure used to stop the attribute selection process.
4. **Decision tree induction**
   - ID3, C4.5 intended for classification.
   - Construct a flow chart like structure.
   - A decision tree is a tree in which :
     - Each internal node tests an attribute.
     - Each branch corresponds to attribute value.
     - Each leaf node assigns a classification.

3.14.2(C) Data Compression

- Data compression is the process of reducing the number of bits needed to either store or transmit the data. This data can be text, graphics, video, audio, etc. This can be usually be done with the help of encoding techniques.
- Data compression techniques can be classified into either lossy or lossless techniques. In lossy technique there is a loss of information whereas in lossless there is no loss.

Scanned by CamScanner

Lossless compression

- Lossless compression consists of those techniques guaranteed to generate an exact duplication of the input dataset after a compress/decompress cycle.
- Lossless compression is essentially a coding technique. There are many different kinds of coding algorithms, such as Huffman coding, run-length coding and arithmetic coding.

Lossy compression

- In lossy compression techniques at the cost of data quality one can achieve higher compression ratio.
- These types of techniques are useful in applications where data loss is affordable. They are mostly applied to digitized representations of analog phenomenon.
- Two methods of lossy data compression :

1. Wavelet transforms
2. Principle component analysis

1. The wavelet transform

A clustering approach which applies wavelet transform to the feature space :

- The orthogonal wavelet transform when applied over a signal results in time scale decomposition through its multi resolution aspect.
- It clusters the functional data into homogenous groups.
- Both grid-based and density-based.

Input parameters

- Number of grid cells for each dimension.
- The wavelet and the number of applications of wavelet transform.
- Clustering approach using Wavelet transform.
- Impose a multidimensional grid like structure on to the data for summarisation.
- Use an n-dimensional feature space for representing spatial data objects.
- Dense regions may be identified by applying the wavelet transform over the feature space.

Scanned by CamScanner

- Applying wavelet transform multiple times results in clusters of different scales.
- Clusters are identified by using hat-shape filters and also suppress weaker information in their boundary.

Major features

- It also results in Effective removal of outliers.
- The technique is Cost efficient.
  Complexity O(N).
- At different scales arbitrary shaped clusters are detected.
- The method is not sensitive to noise or input order.
- It is applicable only to low dimensional data.

2. Principle components analysis

- Principle Component Analysis (PCA) creates a representation of the data with orthogonal basis vectors, i.e. eigenvectors of the covariance matrix of the data. This can also be derived using Singular value decomposition(SVD) method. By this projection original dataset is reduced with little loss of information.
- PCA is often presented using the eigenvalue/eigenvector approach of the covariance matrices. But in efficient computation related to PCA, it is the Singular Value Decomposition (SVD) of the data matrix that is used.
- A few scores of the PCA and the corresponding loading vectors can be used to estimate the contents of a large data matrix.
- The idea behind this is that by reducing the number of eigenvectors used to reconstruct the original data matrix, the amount of required storage space is reduced.

3.14.2(D) Numerosity Reduction

Numerosity reduction technique refers to reducing the volume of data by choosing smaller forms for data representation.

Different techniques used for numerosity reduction are :

1. Histograms

- It replaces data with an alternative, smaller data representation.
- Approximate data distributions.

Scanned by CamScanner

- Divide data into buckets and store average (sum) for each bucket.
- A bucket represents an attribute-value/frequency pair.
- Can be constructed optimally in one dimension using dynamic programming.
- Related to quantization problems.

Different types of histogram

i. **Equal-width histograms :** It divides the range into N intervals of equal size.

ii. **Equal-depth (frequency) partitioning :** It divides the range into *N* intervals, each containing approximately same number of samples.

iii. **V-optimal :** Different Histogram types for a given number of buckets are considered and the one with least variance is chosen.

iv. **MaxDiff :** After the sorting process applied to the data, borders of the buckets are defined where the adjacent values have maximum difference.

Example

1,1,5,5,5,5,5,8,8,10,10,10,10,12,14,14,14,15,15,15,
15,15,15,18,18,18,18,18,18,18,18,20,20,20,20,20,
20,20,21,21,21,21,25,25,25,25,25,28,28,30,30,30

Histogram of above data sample is,

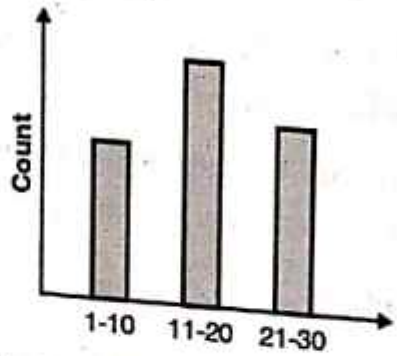


Fig. 3.14.2 : Example of histogram

2. Clustering

- Clustering is a data mining technique used to group the elements based on their similarity without prior knowledge of their class labels.
- It is a technique that belongs to undirected data mining tools.

Scanned by CamScanner

- The goal of undirected data mining is to explore structure in the data. No target variable is to be predicted, therefore there is no difference been made between independent and dependent variables.
- Categorization of clusters based on clustering techniques is given below :
  - Any example belonging to a single cluster would be termed as exclusive cluster.
  - Any example may belong to many clusters in such a case it is said to be overlapping.
  - Any example belongs to a cluster with certain probability then it is said to be probabilistic.
  - A Hierarchical representation may be used for clusters in which clusters may be at highest level of hierarchy and subsequently refined at lower levels to form subclusters.

3. Sampling

- Sampling is used in preliminary investigation as well as final analysis of data.
- Sampling is important in data mining as processing the entire data set is expensive and time consuming.

Types of sampling

i. Simple random sampling

There is an equal probability of selecting any particular item.

ii. Sampling without replacement

As each item is selected, it is removed from the population.

iii. Sampling with replacement

The objects selected for the sample is not removed from the population. In this technique the same object may be selected multiple times.

iv. Stratified sampling

The data is split into partitions and samples are drawn from each partition randomly.

Scanned by CamScanner

Syllabus Topic : Data Transformation and Data Discretization, Normalization, Binning

3.15 Data Transformation and Data Discretization

3.15.1 Data Transformation

Q. Explain the data transformation in detail.

- Operational databases keep changing with the requirements, a data warehouse integrating data from these multiple sources typically faces the problem of inconsistency.
- To deal with these inconsistent data, transformation process may be employed.
- The most commonly used process is "Attribute Naming Inconsistency", as it is very common to use different names to the same attribute in different sources of data.
- E.g. Manager Name may be MGM_NAME in one database, MNAME in the other.
- In this one set of data names is considered and used consistently in the data warehouse.
- Once the naming consistency is done, they must be converted to a common format.
- The conversion process involves the following :
  - (i) ASCII to EBCDIC or vice versa conversion process may be used for characters.
  - (ii) To ensure consistency uppercase representation may be used for mixed case text.
  - (iii) A common format may be adopted for numerical data.
  - (iv) Standardisation must be applied for data format.
  - (v) A common representation may be used for measurement e.g. (Rs/$).
  - (vi) A common format must be used for coded data (e.g. Male/Female, M/F).
- The above conversions are automated and many tools are available for the transformation e.g. DataMapper.

Data transformation can have the following activities

- **Smoothing :** It involves removal of noise from the data.
- **Aggregation :** It involves summarisation and data cube construction.
- **Generalization :** In generalization data is replaced by higher level concepts using concept hierarchy.

- **Normalization :** In norm
- **Example :** To transform

Scaling by using mean
when there are outliers)

- **Attribute/feature con**
used for data mining p

3.15.2 Data Discretizat

- The range of a contin
- Categorical attributes
- By Discretization the
- Dividing the range
given continuous att
- Actual data values r
- Discretization proce

3.15.3 Data Transfo

- Data Transformati
entire set of values
- Following method

1. Min-Max
2. Z-score
3. Decimal scal

1. **Min-Max norm**
original data. Th
Following form
range [minA, m

Scanned by CamScanner

- **Normalization :** In normalization, attribute scaling is performed for a specified range.

  Example : To transform V in [min, max] to V' in [0,1], apply

  $$V' = (V\text{-Min}) / (\text{Max-Min})$$

  Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers) :

  $$V' = (V\text{-Mean}) / \text{Std. Dev.}$$

- **Attribute/feature construction :** In this process new attributes may be constructed and used for data mining process

3.15.2 Data Discretization

- The range of a continuous attribute is divided into intervals.
- Categorical attributes are accepted by only a few classification algorithms.
- By Discretization the size of the data is reduced and prepared for further analysis.
- Dividing the range of attributes into intervals would reduce the number of values for a given continuous attribute.
- Actual data values may be replaced by interval labels.
- Discretization process may be applied recursively on an attribute.

3.15.3 Data Transformation by Normalization

- Data Transformation by Normalization or standardization is the process of making an entire set of values have a particular property.
- Following methods may be used for normalization :

1. Min-Max
2. Z-score
3. Decimal scaling

1. **Min-Max normalization :** Min-max normalization results in a linear alteration of the original data. The values are within a given range.

Following formula may be used to perform mapping a v value, of an attribute A from range [minA, maxA] to a new range [new_minA,new_maxA],

Scanned by CamScanner

$$v' = (v - minA)/(maxA - minA) * (new_maxA - new_minA) + new_minA$$

$$v = 73600 \text{ in } [12000, 98000]$$

$$v' = 0.716 \text{ in } [0,1] \text{ (new range)}$$

2. **Z-score :** In Z-score normalization, data is normalized based on the mean and standard deviation. Z-score is also known as Zero mean normalization.

$$v' = (v - meanA) / std_devA$$

Where, MeanA = sum of the all attribute value of A

std_devA = Standard deviation of all values of A

Example :

If sample data {10, 20, 30}, then

Mean = 20

std_dev = 10

So v' = (-1, 0, 1)

3. **Decimal scaling :** Based on the maximum absolute value of the attributes the decimal point is moved. This process is called as Decimal Scale Normalization

$$v'(i) = v(i)/10^k \text{ for the smallest } k \text{ such that } \max(|v'(i)|)<1.$$

Example : For the range between – 991 and 99,

10^k is 1000 (k=3 as we have maximum 3 digit number in the range)

$$v'(-991) = -0.991 \text{ and } v'(99) = 0.099$$

3.15.4 Discretization by Binning

- This is the data smoothing technique.
- Discretization by binning has two approaches :
  (a) Equal-width (distance) partitioning
  (b) Equal-depth (frequency) partitioning or Equal-height binning
- Both this binning approaches are given in section 3.12.4.

Scanned by CamScanner

3.15.5 Discretization by Histogram Analysis

In Discretization by Histogram divide the data into buckets and store average (sum) for each bucket in smaller data representation.

Different types of histogram

1. Equal-width histograms
2. Equal-depth (frequency) partitioning
3. V-optimal
4. MaxDiff

All the above mentioned methods are given in section 3.14.2(D) in Numerosity Reduction.

Syllabus Topic : Concept Hierarchy Generation

3.16 Concept Hierarchy Generation

The amount of data may be reduced using concept hierarchies. The low level detailed data (for example numerical values for age) may be represented by higher-level data (e.g. Young, Middle aged or Senior).

Concept hierarchy generation for categorical data

- The users or experts may perform a partial/total ordering of attributes explicitly at schema level :

  E.g. street < city < state < country

- Specification of a hierarchy for a set of values by explicit data grouping :

  E.g. {Acton, Canberra, ACT} < Australia

- Ordering of only a partial set of attributes :

  E.g. only street < city, not others

- By analysing number of distinct values the hierarchies or attribute levels may be generated automatically.

E.g. for a set of attributes : {street, city, state, country}

E.g. weekday, month, quarter, year

Scanned by CamScanner

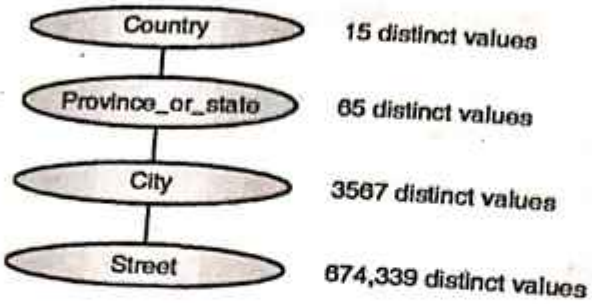


Fig. 3.16.1 : Concept hierarchy example

Syllabus Topic : Concept Description - Attribute Oriented Induction for Data Characterization

3.17 Concept Description : Attribute Oriented Induction for Data Characterization

Data mining can be classified into two categories

1. Descriptive data mining
2. Predictive data mining

1. **Descriptive data mining :** It is used to describe the main features of a collection of data in quantitative terms. It describes concepts or task-relevant data sets in concise, summarize, informative form and presents interesting properties of data.

2. **Predictive data mining : It combines database analysis with multivariate statistics and artificial intelligence.** Moreover, users like the ease and flexibility of having data sets described at different levels of granularity and form different angles. Such descriptive data mining is called **concept description.** Concept description generates description for **characterization and comparison** of data.

Data characterization provides a concise summarization of the given collection of data. It is also called as **class characterization.**

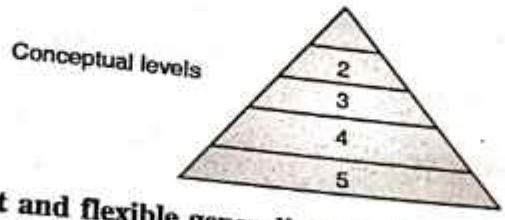
Data comparison provides descriptions comparing two or more collections of data. It is also called as concept or **class comparison.**

Scanned by CamScanner

3.18 Data Generalization and Summarization-based Characterization

3.18.1 Data Generalization

Data Generalization is a process which abstracts a large set of task-relevant data in a database from a low conceptual level to higher ones.



Methods for efficient and flexible generalization of large data sets can be categorized according to two approaches :

(i) Data cube approach (or OLAP approach)

- The data cube approach can be considered as a data warehouse-based, precomputation-oriented, materialized-view approach.
- Computations and results in data cubes.
- It performs off-line aggregation before an OLAP or data mining query is submitted for processing.
- **Strengths**
  - An efficient implementation of data generalization.
  - Computation of various kinds of measures, e.g., count( ) or sum( ).
  - Roll-up and drill-down.
- **Limitations**
  - Only dimensions of simple non-numeric data and measures of simple aggregated numeric values.
  - Lack of intelligent analysis.

(ii) Attribute-Oriented Induction (AOI)

- The attribute-oriented induction approach, at least in its initial proposal, is a relational database, query-oriented, generalization-based, on-line technique.
- Collect the task-relevant data using a relational database query.

Scanned by CamScanner

- Data generalization is performed in two ways :
  - Attribute removal or (Why a OR)
  - Attribute generalization.
- Data aggregation is performed by merging identical, generalized tuples, accumulating their respective counts.
- Presentation of the generalized relation such as charts or rules.

3.18.2 How Attribute-Oriented Induction is Performed?

3.18.2(A) Data Generalization

It can be performed in two ways : attribute removal and attribute generalization.

(i) Attribute Removal is based on the following rule

- If there is no generalization operator on the attribute then remove that attribute.
- Its higher level concepts are expressed in terms of other attributes in the relation the remove the attribute from the initial working relation.

(ii) Attribute Generalization is based on the following rule

If there is a large set of distinct values for an attribute in the initial working relation, the use generalization operator on that attribute.

3.18.2(B) Attribute Generalization Control

It is subjective to how high the attribute is generalized. While attribute oriented generalization a balance should be maintained so that it should not be "too-high" or "too-low" generalization. The control of this process is called **attribute generalization control.**

Two common approaches of this process are :

(I) Attribute generalization threshold control

- Set one generalization threshold for all of the attributes, or set one threshold for each attribute.
- The number of distinct values in an attribute should be less or equal to attribute threshold, If not then further the attribute removal or attribute generalization should be performed.
- Typical threshold value range is from 2 to 8.

(ii) Generalized relatio

- Set a threshold f
- The number of the threshold, If
- Typical thresho
- These two tec control techni control to furt

3.18.2(C) Example

This example sh working relation of T

Name	Gender
John	M
Mack	M
Nancy ...	F ...
Removed	Retaine

For each attr

- **Name :** Thi
- **Gender :** gender.
- **Major :** It
- **Birth_pla** If the nu threshold
- **Birth_da**

Scanned by CamScanner

(II) Generalized relation threshold control

- Set a threshold for generalized relation.
- The number of (distinct) tuples in the generalized relation should be less or equal to the threshold. If not then further generalization should be performed.
- Typical threshold value range is from 10 to 30.
- These two techniques can be applied in sequence : first apply the attribute threshold control technique to generalize each attribute, and then apply relation threshold control to further reduce the size of the generalized relation.

3.18.2(C) Example of Attribute Oriented Induction

This example shows that how attribute-oriented induction is performed on the initial working relation of Table 3.18.1.

Table 3.18.1 : Initial Relation

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
John	M	CS	Vancouver, BC,Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Mack	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Nancy ...	F ...	Physics ...	Seattle, WA, USA	25-8-70 ...	125 Austin Ave., Burnaby	420-5232 ...	3.83 ...
Removed	Retained	Sci,Eng,Bus	Country	Age range	City	Removed	Excl,VG

For each attribute of the relation, the generalization proceeds as follows :

- **Name :** This attribute is removed as there are large number of distinct values for name
- **Gender :** This attribute is retained as there are only two distinct values (M and F) for gender.
- **Major :** It can be generalized to the values {arts, science, engineering, business}
- **Birth_place :** This attribute can be generalized as it has a large number of distinct values. If the number of distinct values for country is less than the attribute generalization threshold, then birth_place should be generalized to birth_country (or country).
- **Birth_date :** It can be generalized to age or age_range.

Scanned by CamScanner

- **Residence :** As number of values for city in residence attribute is less than threshol value, it can be generalizes to city.
- **Phone# :** This attribute has large distinct values so it should be removed in generalizatio
- **GPA :** Grade Point Average (GPA) has distinct numerical values but it can be generalize into numerical intervals like {3.75-4.0, 3.5-3.75, ...}, which in turn can be grouped in descriptive values such as {Excellent (Excl), Very Good(VG),.....}. The attribute ca therefore be generalized.

A generalized relation obtained by attribute-oriented induction on the data o Table 3.18.1 is as given in Table 3.18.2.

Table 3.18.2 : Generalized Relation

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
...	...	...	...	...	...	...

3.19 University Questions and Answers

May 2010

Q. 1 Explain data mining as a step in KDD. Give the architecture of typical DM system.
(Ans. : Refer sections 3.5 and 3.3) **(10 Marks)**

Dec. 2010

Q. 2 Explain data mining as a step in KDD. Give the architecture of typical DM system.
(Ans. : Refer sections 3.5 and 3.3) **(10 Marks)**

May 2011

Q. 3 Describe the steps in the KDD process with a suitable bock diagram.
(Ans. : Refer section 3.3) **(5 Marks)**

Dec. 2011

Q. 4 What is data mining ? What are techniques and applications of data mining ? Explain the architecture of typical data mining system.
(Ans. : Refer sections 3.1, 3.4, 3.7 and 3.3) **(10 Marks)**

Scanned by CamScanner

May 2012

Q. 5 Explain the steps in KDD with a suitable block diagram. *(Ans. : Refer section 3.5)* **(5 Marks)**

Dec. 2012

Q. 6 With a neat diagram describe the KDD process. *(Ans. : Refer section 3.5)* **(10 Marks)**

Q. 7 Describe through a short note : Visualization techniques for Data warehousing and mining. *(Ans. : Refer section 3.10)* **(10 Marks)**

Dec. 2013

Q. 8 Explain Data mining as a step in KDD. Explain the architecture of a typical DM system. *(Ans. : Refer sections 3.5 and 3.3)* **(10 Marks)**

May 2016

Q. 9 Discuss :

i. The steps in KDD process *(Ans. : Refer section 3.5)*

ii. The architecture of a typical DM system *(Ans. : Refer section 3.3)* **(10 Marks)**

Q. 10 Discuss different steps involved in data preprocessing. *(Ans. : Refer sections 3.12, 3.13 and 3.14)* **(10 Marks)**

□□□

Chapter Ends

Scanned by CamScanner

Module 4

CHAPTER 4

Classification, Prediction and Clustering

Syllabus :

Classification, Prediction and Clustering : Basic Concepts, Decision Tree using Information Gain, Induction : Attribute Selection Measures, Tree pruning, Bayesian Classification : Naive Bayes, Classifier Rule - Based Classification : Using IFTHEN Rules for classification, Prediction : Simple linear regression, Multiple linear regression Model Evaluation & Selection : Accuracy and Error measures, Holdout, Random Sampling, Cross Validation, Bootstrap.

Clustering : Distance Measures, Partitioning Methods (k-Means, k-Medoids), Hierarchical Methods(Agglomerative, Divisive).

Syllabus Topic : Basic Concepts

4.1 Basic Concept : Classification

- Classification constructs the classification model based on training data set and using that model classifies the new data.
- It predicts the value of classifying attribute or class label.
- **Typical applications :**
  - Classify credit approval based on customer data.
  - Target marketing of product.
  - Medical diagnosis based on symptoms of patient.
  - Treatment effectiveness analysis of patient based on the treatment given.

Scanned by CamScanner

- **Various classification techniques**
  - Regression
  - Decision trees
  - Rules
  - Neural networks

4.1.1 Classification Problem

- Suppose a database D is given as D= $\{t_1, t_2,\ldots,t_n\}$ and a set of desired classes are $C=\{C_1,\ldots,C_m\}$, the **Classification problem** is to define the mapping m in such a way that which tuple of database D belongs to which class of C.
- Actually divides D into **equivalence classes.**
- **Prediction** is similar, but may be viewed as having infinite number of classes. Prediction models continuous-valued functions, i.e., predicts unknown or missing values.

4.1.2 Classification Example

How teacher gives grades to students based on their marks obtained :

- If x >= 90 then grade = A.
- If 80<=x<90 then grade = B.
- If 70<=x<80 then grade = C.
- If 60<=x<70 then grade = D.
- If x<60 then grade = E.

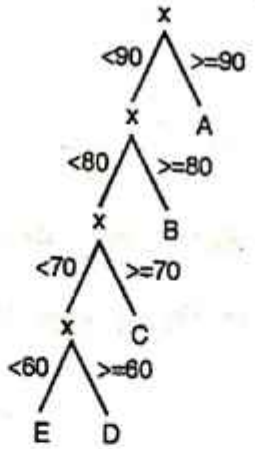


Fig. 4.1.1 : Classification of grading

4.1.3 Classification is a Two Step Process

1. Model construction

- Every sample tuple or object has assigned a predefined class label
- Those set of sample tuples or subset data set is known as training data set.
- The constructed model based on training data set is represented as classification rules, decision trees or mathematical formulae.

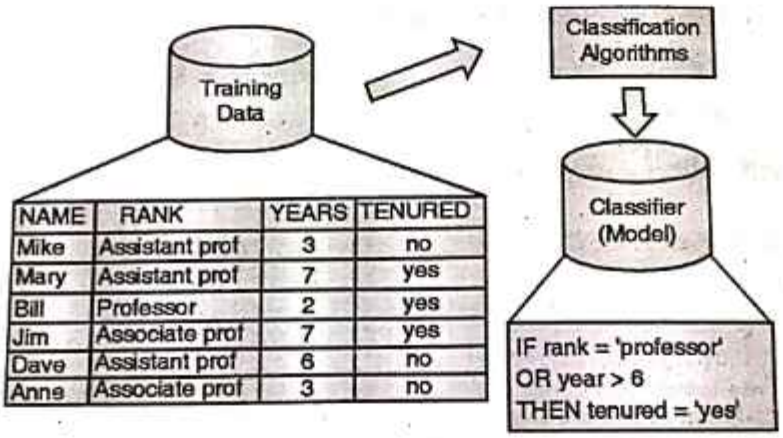
Scanned by CamScanner

2. Model usage

- For classifying unknown objects or new tuple use the constructed model.
- Compare the class label of test sample with the resultant class label.
- Estimate accuracy of the model by calculating the percentage of test set samples that are correctly classified by the model constructed.
- Test sample data and training data samples are always different, otherwise over-fitting will occur.

Example

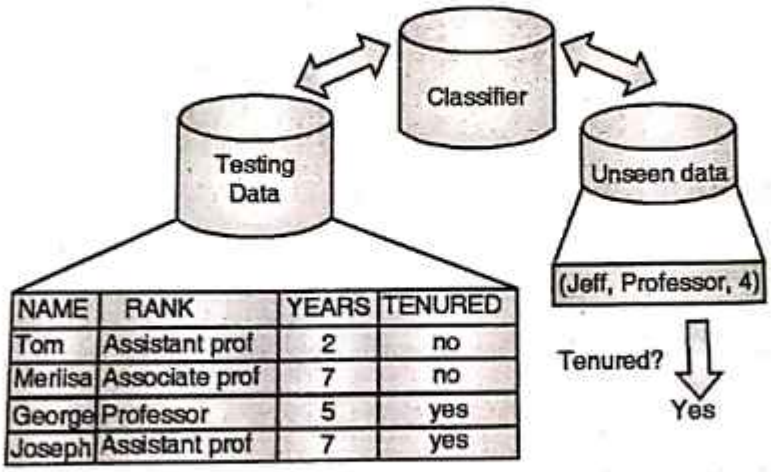
Classification process : (1) Model construction



NAME	RANK	YEARS	TENURED
Mike	Assistant prof	3	no
Mary	Assistant prof	7	yes
Bill	Professor	2	yes
Jim	Associate prof	7	yes
Dave	Assistant prof	6	no
Anne	Associate prof	3	no

Fig. 4.1.2 : Learning : Training data are analyzed by a classification algorithm

Classification process : (2) Model usage (Use the model in prediction)



NAME	RANK	YEARS	TENURED
Tom	Assistant prof	2	no
Merlisa	Associate prof	7	no
George	Professor	5	yes
Joseph	Assistant prof	7	yes

Fig. 4.1.3 : Classification : Test data are used to estimate the accuracy of the classification rule

Scanned by CamScanner

For example : How to perform classification task for classification of medical patients by their disease ?

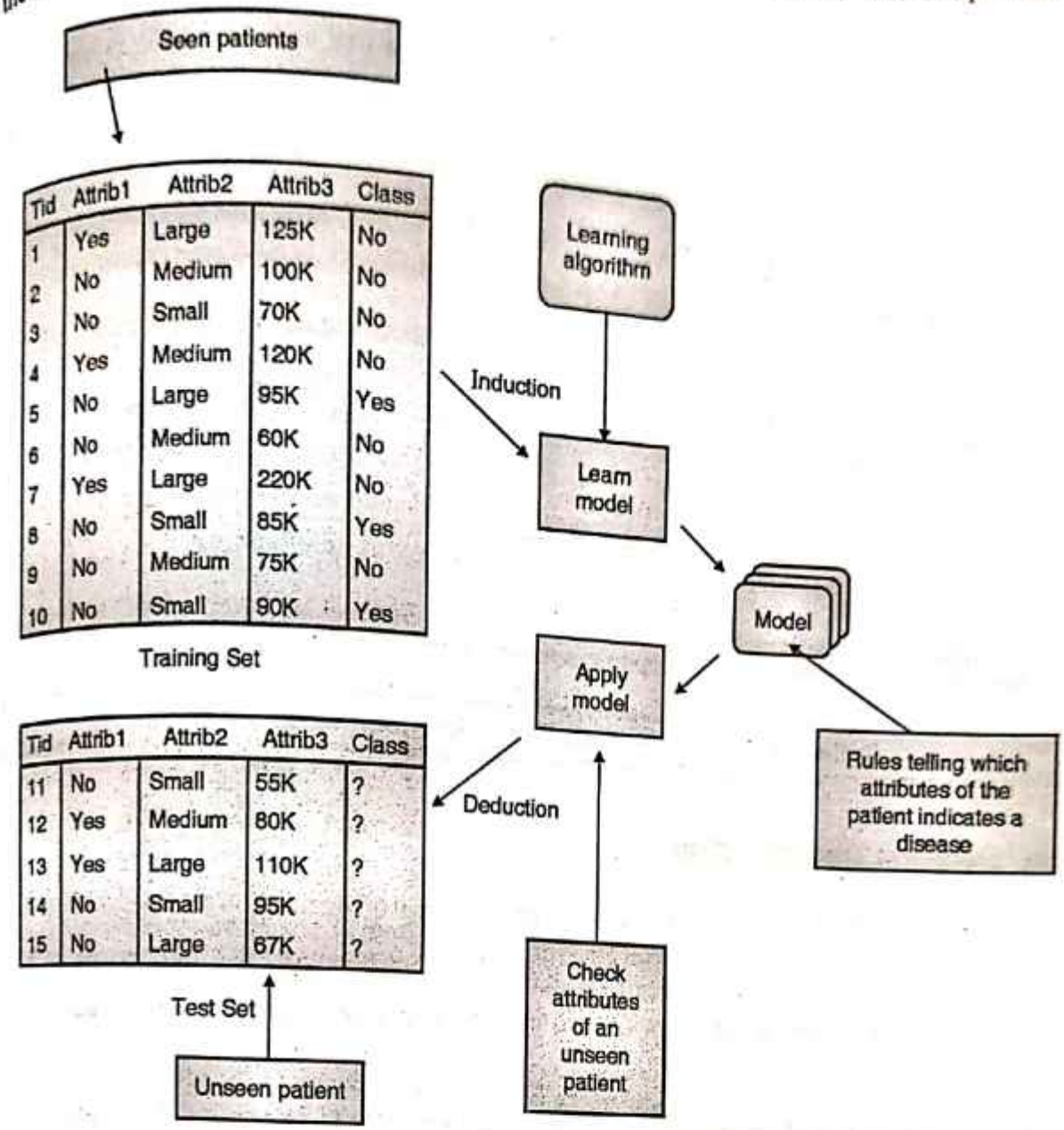


Fig. 4.1.4 : Example of classification model

4.1.4 Difference between Classification and Prediction

Sr. No.	Classification	Prediction
1.	Classification is a major type of prediction problem where classification is used to predict discrete or nominal values.	Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled sample.

Scanned by CamScanner

Sr. No.	Classification	Prediction
2.	Classification is the use of prediction to predict class labels.	It is used to assess the values or value ra of an attribute that a given sample is likely have.
3.	E.g. Group patients based on their known medical data and treatment outcome then it's a classification.	E.g. if a classification model is used predict the treatment outcome for a ne patient, then it would be a prediction.

4.2 Classification Methods

Classifications methods are given below :

1. Decision Tree Induction : Attribute selection measures, tree pruning
2. Bayesian Classification : Naïve Bayes classifier

Syllabus Topic : Decision Tree using Information Gain

4.2.1 Decision Tree Induction

- Training dataset should be class-labelled for learning of decision trees in decision tre induction.
- A decision tree represents rules and it is very a popular tool for classification an prediction.
- Rules are easy to understand and can be directly used in SQL to retrieve the records from database.
- To recognize and approve the discovered knowledge acquired from decision model is very crucial task.
- There are a many algorithms to build decision trees :
  - ID3 (Iterative Dichotomiser 3)
  - C4.5 (Successor of ID3)
  - CART (Classification and Regression Tree)
  - CHAID (CHi-squared Automatic Interaction Detector)

Scanned by CamScanner

4.2.1(A) Appropriate Problems for Decision Tree Learning

Decision tree learning is appropriate for the problems having the characteristics given below :

- Instances are represented by a fixed set of attributes (e.g. gender) and their values (e.g. male, female) described as **attribute-value pairs.**
- If the attribute has small number of disjoint possible values (e.g. high, medium, low) or there are only two possible classes (e.g. true, false) then decision tree learning is easy.
- Extension to decision tree algorithm also handles real value attributes (e.g. salary).
- Decision tree gives a class label to each instance of dataset.
- Decision tree methods can be used even when some training examples have unknown values (e.g. humidity is known for only a fraction of the examples).
- Learned functions are either represented by a decision tree or re-represented as sets of if-then rules to improve readability.

4.2.1(B) Decision Tree Representation

Decision tree classifier has tree type structure which has leaf nodes and decision nodes.

- **A leaf node** is the last node of each branch and indicates class label or value of target attribute
- **A decision node** is the node of tree which has leaf node or sub-tree. Some test to be carried on the each value of decision node to get the decision of class label or to get next sub-tree.

Example : Decision tree representation for play tennis.

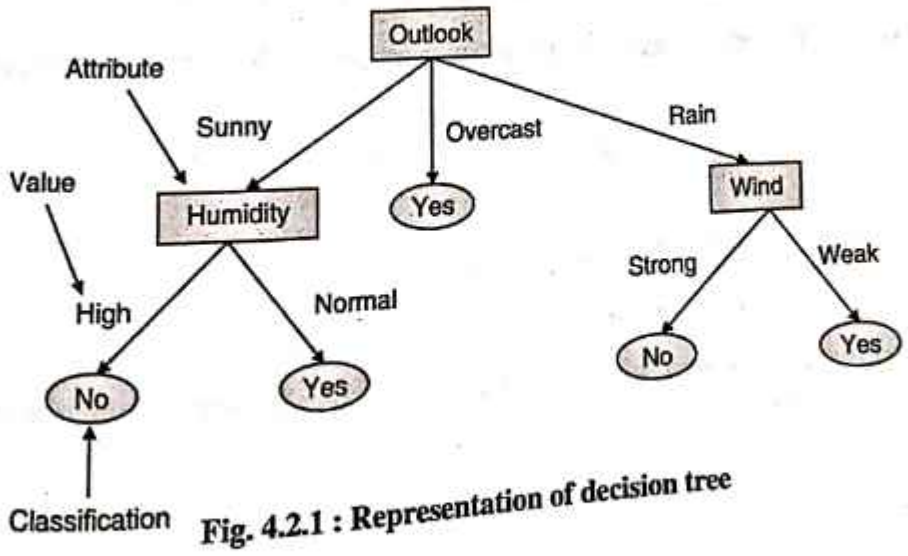


Fig. 4.2.1 : Representation of decision tree

Scanned by CamScanner

Other representation for play tennis

- Logical expression for Play tennis = Yes is given below,

  (Outlook = sunny ∧ humidity = normal) ∨ (outlook = overcast) ∨ outlook = rain ∧ wind = weak)

- If-then rules :
  - IF outlook = sunny ∧ humidity = normal THEN play tennis = Yes
  - IF outlook = sunny ∧ humidity = high THEN play tennis = No
  - IF outlook = overcast THEN play tennis = Yes
  - IF outlook = rain ∧ wind = weak THEN play tennis = Yes
  - IF outlook = rain ∧ wind = strong THEN play tennis = No

Syllabus Topic : Induction - Attribute Selection Measure

4.2.1(C) Attribute Selection Measure

(1) Gini index (IBM Intelligent Miner)

- Suppose all attributes are continuous-valued.
- Assume that each value of attribute has many possible split.
- It can be adapted for categorical attributes.
- An alternative method to information gain is called the **gini index.**
- Gini is used in CART (Classification and Regression Trees), IBM's Intelligent Miner system, SPRINT (Scalable PaRallelizable INduction of decision Trees).
- If a data set T contains examples from n classes, gini index, gini(T) is defined as

$$\text{gini}(T) = 1 - \sum_{j=1}^{n} p_j^2$$

  Where, p_j is the relative frequency of class j in T.

  gini(T) is minimized if the classes in T are skewed.

- After splitting T into two subsets T_1 and T_2 with sizes N_1 and N_2, the gini index of the split data is defined as,

Scanned by CamScanner

$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

- For each attribute, each of the possible binary splits is considered. For a discrete-valued attribute, the attribute providing smallest $gini_{split}(T)$ is chosen to split the node. For continuous-valued attributes, each possible split-point must be considered.

Example using gini index :

- Consider the following dataset D,

Outlook	Temperature	Humidity	Windy	Play ?
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
overcast	Mild	High	True	Yes
overcast	Hot	Normal	False	Yes
rainy	Mild	High	True	No

- The total for node D is :

$$gini(D) = 1 - sum(p1^2, p2^2, \ldots pn^2) \quad \ldots(4.2.1)$$

Where, p1....n are the frequency ratios of class 1....n in D.

In the above example, there are only two classes : 1) Yes 2) No

Therefore n = 2

Scanned by CamScanner

So the gini index for the entire set :

$$= 1 - ([9/14]^2 + [5/14]^2) = 1 - (0.413 + 0.127)$$

$$= 0.459$$

- To find the splitting criterion for the tuples in D, compute the gini index for e attribute.
- The gini value of a split of D into subsets $D_1, D_2, \ldots\ldots, D_n$ is :

$$\text{Split (D)} = N_1/N \text{ gini } (D_1) + N_2/N \text{ gini } (D_2) + \ldots\ldots\ldots\ldots + N_n/N \text{ gini } (D_n) \quad \ldots(4.2.2)$$

Where, N = Size of dataset D

$N_1, N_2, \ldots\ldots N_n$ = Size of each subset.

- Consider First attribute as outlook from dataset D

E.g. Outlook splits into 5 tuples for sunny, 4 tuples for overcast, 5 tuples for rainy :

So $N_1 = 5$, $N_2 = 4$ and $N_3 = 5$

$$\text{Split} = 5/14 \text{ gini (sunny)} + 4/14 \text{ gini (overcast)} + 5/14 \text{ gini (rainy)} \quad \ldots(4.2.3)$$

- Now consider the dataset D_1 with only tuples having outlook = "sunny"

Outlook	Temperature	Humidity	Windy	Play ?
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Sunny	Mild	Normal	True	Yes

Using Equation (4.2.1),

$$\text{Gini(sunny)} = 1 - \text{sum}([2/5]^2, [3/5]^2) = 1 - 0.376 = 0.624$$

Similarly calculate for "overcast" and "rainy"

$$\text{Overcast} = 1 - \text{sum}(4/4^2, 0/4^2) = 0.0$$

$$\text{Rainy} = 1 - \text{sum}([3/5]^2, [2/5]^2) = 0.624$$

Scanned by CamScanner

From Equation (4.2.3)

$$\text{Split} = 5/14 \text{ gini (sunny)} + 4/14 \text{ gini (overcast)} + 5/14 \text{ gini (rainy)}$$

$$\text{Split} = (5/14 * 0.624) + 0 + * (5/14 * 0.624)$$

$$\text{Split} = 0.446$$

The attribute that generates the smallest gini split value is chosen to split the node on.

(2) Information gain (ID3/C4.5)

- All attributes are believed to be categorical.
- It can be adapted for continuous-valued attributes.
- The attribute which has the highest information gain is selected for split.
- Assume there are two classes, P and N.
- Suppose we have S samples, out of these p samples belongs to class P and n samples belongs to class N.
- The amount of information, needed to decide if random example in S belongs to P or N is defined as

$$I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

- Assume that using attribute A, a set S will be partitioned into sets $\{S_1, S_2, \ldots, S_v\}$
- If S_i contains p_i examples of P and n_i examples of N, the entropy, or the expected information needed to classify objects in all subtrees S_i is

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

Entropy (E) :

- Expected amount of information (in bits) needed to assign a class to a randomly drawn object in S under the optimal, shortest-length code.
- **Calculate information gain i.e. gain (A) :** Measures reduction in entropy achieved because of the split. Choose the split that achieves most reduction (maximizes GAIN)

$$\text{Gain}(A) = I(p, n) - E(A)$$

(3) Gain ratio

- It is an alteration of the information gain that reduces its favouritism on high-branch attributes.

Scanned by CamScanner

- Gain ratio should be big when data is evenly spread and small when all data belong to one branch. So it considers number of branches and size of branches when it selects attribute to split.
- Intrinsic information is the entropy of distribution of instances into branches.

$$\text{Intrinsic Info}(S,A) = -\sum \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

- Gain ratio normalizes info gain by using following formula:

$$\text{Gain Ratio}(S,A) = \frac{\text{Gain}(S,A)}{\text{Intrinsic Info}(S,A)}$$

4.2.1(D) Algorithm for Inducing a Decision Tree

The Basic ideas behind ID3 :

- **C4.5** is an extension of ID3.
- **C4.5** accounts for unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation and so on.
- **C4.5** is a designed by Quinlan to address the following issues not given by ID3 :
  - It avoids over fitting the data.
  - It determines the depth of decision tree and reduces the error pruning.
  - It also handles continuous value attributes e.g. Salary or temperature.
  - It works for missing value attribute and handles suitable attribute selection measure.
  - It gives better efficiency of computation. The algorithm to generate decision tree is given by Jiawei Han et al. as below :

Algorithm : Generate_decision_tree : Generate a decision tree from the training tuples of data partition, D.

Input :

- Data partition, D, which is a set of training tuples and their associated class labels;
- Attribute_list, the set of candidate attributes;
- Attribute_selection_method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting_attribute and, possibly, either a split-point or splitting subset.

Scanned by CamScanner

Output :

A decision tree.

Method :

1. create a node N;
2. if tuples in D are all of the same class, C, then
   return N as a leaf node labeled with the class C;
3. if attribute_list is empty then
   return N as a leaf node labeled with the majority class in D; // majority voting
4. Apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
5. label node N with splitting_criterion;
6. if splitting_attribute is discrete-valued and
   Multiway splits allowed then // not restricted to binary trees
   Attribute_list ← attribute_list – splitting_attribute; // remove splitting_attribute
7. for each outcome j of splitting_criterion
   // partition the tuples and grow subtrees for each partition
8. let D_j be the set of data tuples in D satisfying outcome j; // a partition
9. if D_j is empty then
   attach a leaf labeled with the majority class in D to node N;
10. else attach the node returned by Generate_decision_tree (D_j, attribute_list) to node N; endfor
11. return N;

Strengths and Weakness of Decision Tree Methods

- The strengths of decision tree methods are :
  - Decision trees are able to generate understandable rules.
  - Decision trees perform classification without requiring much computation.
  - Decision trees are able to handle both continuous and categorical variables.

Scanned by CamScanner

- Decision trees provide a clear indication of which fields are most important f prediction or classification.

– The weaknesses of decision tree methods :

- Not suitable for prediction of continuous attribute.
- Perform poorly with many class and small data.
- Computationally expensive to train.

Syllabus Topic : Tree Pruning

4.2.1(E) Tree Pruning

– Because of noise or outliers, the generated tree may over fit due to many branches.

– To avoid over fitting, prune the tree so that it is not too specific.

Prepruning

- Start pruning in the beginning while building the tree itself.
- Stop the tree construction in early stage.
- Avoid splitting a node by checking the threshold with the goodness measure falling below a threshold.
- Selection of correct threshold is difficult in prepruning.

Postpruning

- Build the full tree then start pruning, remove the branches.
- Use different set of data than training data set to get the best pruned tree.

4.2.1(F) Examples of ID3

Ex. 4.2.1 : Apply ID3 on the following training dataset from all electronics customer database and extract the classification rule from the tree

Table P. 4.2.1 : Training data of customer

Age	Income	Student	Credit_rating	Class : buys_computer
≤ 30	High	No	Fair	No
≤ 30	High	No	Excellent	No
31...40	High	No	Fair	Yes

Age	Income
> 40	Medium
> 40	Low
> 40	Low
31...40	Low
≤ 30	Mediu
≤ 30	Low
> 40	Medi
≤ 30	Medi
31...40	Med
31...40	Hig
> 40	Me

Soln. :

Class P : buys_c

Class N : buys_

Total number of r

Count the numbe

So number of re

So Information

I (p, n)

Step 1 : Comput

For age ≤ 30

p_i = with "y

Therefore ,

Scanned by CamScanner

Age	Income	Student	Credit_rating	Class : buys_computer
> 40	Medium	No	Fair	Yes
> 40	Low	Yes	Fair	Yes
> 40	Low	Yes	Excellent	No
31...40	Low	Yes	Excellent	Yes
≤ 30	Medium	No	Fair	No
≤ 30	Low	Yes	Fair	Yes
> 40	Medium	Yes	Fair	Yes
≤ 30	Medium	Yes	Excellent	Yes
31...40	Medium	No	Excellent	Yes
31...40	High	Yes	Fair	Yes
> 40	Medium	No	Excellent	No

Soln. :

Class P : buys_computer = "yes"

Class N : buys_computer = "no"

Total number of records 14.

Count the number of records with "yes" class and "no" class.

So number of records with "yes" class = 9 and "no" class = 5

So Information gain $= I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$

$$I(p, n) = I(9, 5)$$
$$= -(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$$
$$= 0.940$$

Step 1 : Compute the entropy for age

For age ≤ 30,

p_i = with "yes" class = 2 and n_i = with "no" class = 3

Therefore , $I(p_i, n_i) = I(2,3) = 0.971$.

Scanned by CamScanner

Similarly for different age ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.1(a)

Age	p_i	n_i	$I(p_i, n_i)$
≤ 30	2	3	0.971
31...40	4	0	0
> 40	3	2	0.971

So, the expected information needed to classify a given sample if the samples are partitioned according to age is,

Calculate entropy using the values from the Table P. 4.2.1(a) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{age}) = \frac{5}{14} I(2, 3) + \frac{4}{14} I(4, 0) + \frac{5}{14} I(3, 2) = 0.694$$

Hence

$$\text{Gain(age)} = I(p, n) - E(\text{age})$$
$$= 0.940 - 0.694 = 0.246$$

Similarly,

$$\text{Gain (income)} = 0.029$$
$$\text{Gain (student)} = 0.151$$
$$\text{Gain (credit_rating)} = 0.048$$

Now the age has the highest information gain among all the attributes, so select age as test attribute and create the node as age and show all possible values of age for further splitting.

Since Age has three possible values, the root node has three branches ($\leq 30, 31...40, > 40$).

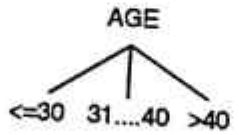


Fig. P. 4.2.1

Scanned by CamScanner

Step 2 :

The next question is "what attribute should be tested at the Age branch node?" Since we have used Age at the root, now we have to decide on the remaining three attributes : income, student, or credit_rating.

Consider Age : ≤ 30 and count the number of tuples from the original given training set

$S_{\leq 30} = 5$ (Age: ≤ 30)

Table P. 4.2.1(b)

Age	Income	Student	Credit_rating	Buys_computer
≤ 30	High	No	Fair	No
≤ 30	High	No	Excellent	No
≤ 30	Medium	No	Fair	No
≤ 30	Low	Yes	Fair	Yes
≤ 30	Medium	Yes	Excellent	Yes

Note : Refer Table P. 4.2.1(b)

Total number of Yes tuple = 2 and total number of No tuple = 3

$I(p, n) = I(2, 3) = -(2/5)\log_2(2/5) - (3/5)\log_2(3/5) = 0.971$

(i) Compute the entropy for income : (High, medium, low)

For Income = High,

p_i = with "yes" class = 0 and n_i = with "no" class = 2

Therefore , $I(p_i, n_i) = I(0,2) = -(0/2)\log_2(0/2) - (2/2)\log_2(2/2) = 0.$

Similarly for different age ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.1(c)

Income	p_i	n_i	$I(p_i, n_i)$
High	0	2	0
Medium	1	1	1
Low	1	0	0

Scanned by CamScanner

Calculate entropy using the values from the Table P. 4.2.1(c) and the formula given a

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Income}) = 2/5 * I(0, 2) + 2/5 * I(1, 1) + 1/5 * I(1, 0) = 0.4$$

Note : $S_{\leq 30}$ is the total training set.

Hence

$$Gain(S_{\leq 30}, \text{Income}) = I(p, n) - E(\text{Income})$$

$$= 0.971 - 0.4 = 0.571$$

(ii) Compute the entropy for Student : (No , yes)

For Student = No,

p_i = with "yes" class = 0 and n_i = with "no" class = 3

Therefore, $I(p_i, n_i) = I(0,3) = -(0/3)\log_2(0/3) - (3/3)\log_2(3/3) = 0.$

Similarly for different outlook ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.1(d)

Student	p_i	n_i	$I(p_i, n_i)$
No	0	3	0
Yes	2	0	0

Calculate Entropy using the values from the Table P. 4.2.1(d) and the formula give below

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Student}) = 3/5 * I(0, 3) + 2/5 * I(2, 0) = 0$$

Note : $S_{\leq 30}$ is the total training set.

Hence

$$Gain(S_{\leq 30}, \text{Student}) = I(p, n) - E(\text{Student})$$

$$= 0.971 - 0 = 0.971$$

Scanned by CamScanner

(iii) Compute the entropy for credit_rating : (Fair, excellent)

For credit_rating = Fair,

p_i = with "yes" class = 1 and n_i = with "no" class = 2

Therefore $I(pi, ni) = I(1, 2) = -(1/3)\log2(1/3) - (2/3)\log2(2/3)$

$= 0.918$

Similarly for different outlook ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.1(e)

Credit_rating	p_i	n_i	$I(p_i, n_i)$
Fair	1	2	0.918
Excellent	1	1	1

Calculate entropy using the values from the Table P. 4.2.1(e) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Credit_rating}) = 3/5 * I(1, 2) + 2/5 * I(1, 1) = 0.951$$

Note : $S_{\leq 30}$ is the total training set.

Hence

$$\text{Gain}(S_{\leq 30}, \text{credit_rating}) = I(p, n) - E(\text{credit_rating})$$

$$= 0.971 - 0.951 = 0.02$$

Therefore,

$$\text{Gain}(S_{\leq 30}, \text{student}) = 0.970$$

$$\text{Gain}(S_{\leq 30}, \text{income}) = 0.570$$

$$\text{Gain}(S_{\leq 30}, \text{credit_rating}) = 0.02$$

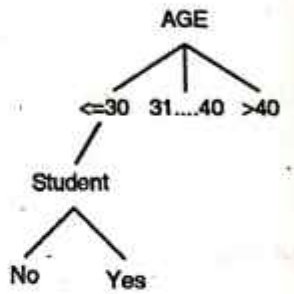


Fig. P. 4.2.1(a)

Student has the highest gain; therefore, it is below Age : " ≤ 30".

Scanned by CamScanner

Step 3 :

Consider now only income and credit rating for age : 31…40 and count the number of tuples from the original given training set

$$S_{31...40} = 4 \text{ (age : 31...40)}$$

Table P. 4.2.1(f)

Age	Income	Student	Credit_rating	Buys_computer
31…40	High	No	Fair	Yes
31…40	Low	Yes	Excellent	Yes
31…40	Medium	No	Excellent	Yes
31…40	High	Yes	Fair	Yes

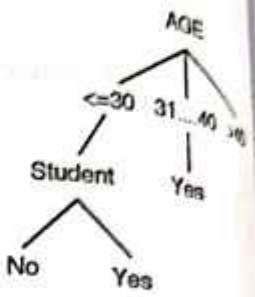


Fig. P. 4.2.1(h)

Since for the attributes income and credit_rating , buys_computer = yes, so assign class 'yes' to 31…40

Step 4 :

Consider income and credit_rating for age: >40 and count the number of tuples from the original given training set

$$S_{>40} = \text{ (age: >40)}$$

Table P. 4.2.1(g)

Age	Income	Student	Credit_rating	Buys_computer
> 40	Medium	No	Fair	Yes
> 40	Low	Yes	Fair	Yes
> 40	Low	Yes	Excellent	No
> 40	Medium	Yes	Fair	Yes
> 40	Medium	No	Excellent	No

Consider the Table P. 4.2.1(g) as the new training set and calculate the Gain for income and credit_rating

Class P : buys_computer = "yes"

Class N : buys_computer = "no"

Total number of records 5.

Scanned by CamScanner

Count the number of records with "yes" class and "no" class.

So number of records with "yes" class = 3 and "no" class = 2

So Information gain = $I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$

$$I(p, n) = I(3, 2) = -(3/5)\log_2(3/5) - (2/5)\log_2(2/5) = 0.970$$

(d) **Compute the entropy for credit_rating**

For credit_rating = Fair

p_i = with "yes" class = 3 and n_i = with "no" class = 0

Therefore, $I(p_i, n_i) = I(3,0) = 0$.

For credit_rating = Excellent

p_i = with "yes" class = 0 and n_i = with "no" class = 2

Therefore, $I(p_i, n_i) = I(0,2) = 0$

Similarly for different age ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.1(h)

Credit_rating	p_i	n_i	$I(p_i, n_i)$
Fair	3	0	0
Excellent	0	2	0

Calculate entropy using the values from the Table P. 4.2.1(h) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{Credit_rating}) = \frac{3}{5} I(3,0) + \frac{2}{5} I(0,2) = 0$$

Hence

$$\text{Gain}(S_{>40}, \text{credit_rating}) = I(p, n) - E(\text{credit_rating})$$

$$= 0.970 - 0 = 0.970$$

Scanned by CamScanner

(ii) Compute the entropy for income : (High, medium, low)

For Income = High,

p_i = with "yes" class = 0 and n_i = with "no" class = 0

Therefore, $I(p_i, n_i) = I(0,0) = 0$

Similarly for different outlook ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.1(i)

Income	p_i	n_i	$I(p_i, n_i)$
High	0	0	0
Medium	2	1	0.918
Low	1	1	1

Calculate Entropy using the values from the Table P. 4.2.1(i) and the formula given below

$$E(\text{Income}) = 0/5 * I(0, 0) + 3/5 * I(2, 1) + 2/5 * I(1, 1) = 0.951$$

Note : $S_{>40}$ is the total training set.

Hence

$$Gain(S_{>40}, \text{income}) = I(p, n) - E(\text{income}) = 0.970 - 0.951 = 0.019$$

Therefore,

$$Gain(S_{>40}, \text{income}) = 0.019$$

$$Gain(S_{>40}, \text{Credit_rating}) = 0.970$$

Credit_rating has the highest gain; therefore, it is below Age : "> 40".

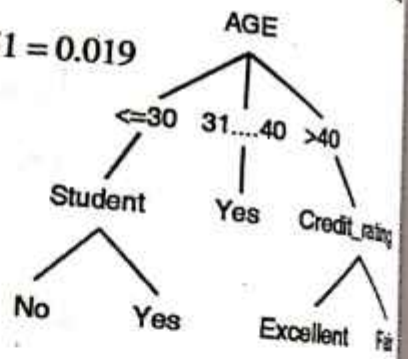


Fig. P. 4.2.1(c)

Output : A Decision Tree for "buys_computer"

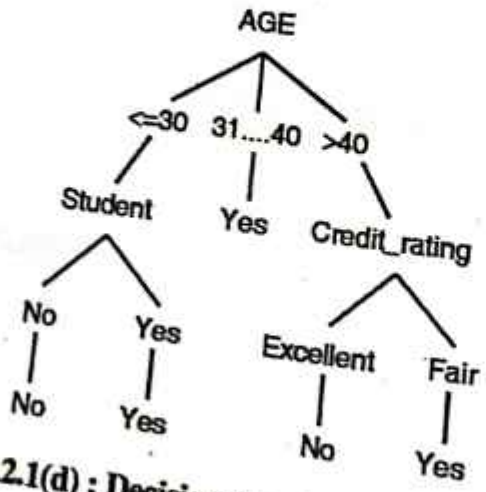


Fig. P. 4.2.1(d) : Decision tree for "buys computer"

Scanned by CamScanner

Extracting classification rules from trees

Example

IF age = "≤ 30" AND student = "no" THEN buys_computer = "no"

IF age = "≤ 30" AND student = "yes" THEN buys_computer = "yes"

IF age = "31…40" THEN buys_computer = "yes"

IF age = "> 40" AND credit_rating = "excellent" THEN buys_computer = "no"

IF age = "> 40" AND credit_rating = "fair" THEN buys_computer = "yes"

Ex. 4.2.2 : The weather attributes are outlook, temperature, humidity, and wind speed. They can have the following values :

Outlook = {sunny, overcast, rain}

Temperature = {hot, mild, cool}

Humidity = {high, normal}

wind = {weak, strong}

Sample data set S are :

Table P. 4.2.2 : Training data set for Play Tennis

Day	Outlook	Temperature	Humidity	Wind	Play ball
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

We need to find which attribute will be the root node in our decision tree. The gain is calculated for all four attributes using formula of gain(A).

Scanned by CamScanner

Soln. :

Class P : Playball = "yes"

Class N : Playball = "no"

Total number of records 14.

Count the number of records with "yes" class and "no" class.

So number of records with "yes" class = 9 and "no" class = 5

$$\text{So Information gain} = I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

$$I(p, n) = I(9, 5) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$$

$$= (-0.643) * (-0.637) + (-0.357) * (-1.485)$$

$$= 0.409 + 0.530 = 0.940$$

Step 1 : Compute the entropy for outlook : (Sunny, overcast, rain)

For outlook = sunny,

p_i = with "yes" class = 2 and n_i = with "no" class = 3

Therefore, $I(p_i, n_i) = I(2,3) = -(2/5)\log_2(2/5) - (3/5)\log_2(3/5) = 0.971$.

Similarly for different outlook ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.2(a)

Outlook	p_i	n_i	$I(p_i, n_i)$
Sunny	2	3	0.971
Overcast	4	0	0
Rain	3	2	0.971

Calculate entropy using the values from the Table P. 4.2.2(a) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{outlook}) = \frac{5}{14} I(2, 3) + \frac{4}{14} I(4, 0) + \frac{5}{14} I(3, 2) = 0.694$$

Note : T is the total training set.

Hence $\text{Gain}(T, \text{outlook}) = I(p, n) - E(\text{outlook}) = 0.940 - 0.694 = 0.246$

Scanned by CamScanner

Similarly,

$$\text{Gain (T, Temperature)} = 0.029$$
$$\text{Gain (T, Humidity)} = 0.151$$
$$\text{Gain (T, Wind)} = 0.048$$

Outlook shows the highest gain, so it is used as the decision attribute in the root node.

As Outlook has only values "sunny, overcast, rain", the root node has three branches.

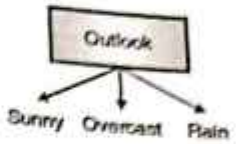


Fig. P. 4.2.2

Step 2 :

As attribute outlook at root, we have to decide on the remaining three attribute for sunny branch node.

Consider outlook = Sunny and count the number of tuples from the original given training set

$S_{sunny} = \{1, 2, 8, 9, 11\} = 5$ (From Table P. 4.2.2, outlook = sunny)

Table P. 4.2.2(b)

Day	Outlook	Temperature	Humidity	Wind	Play ball
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Note : Refer Table P. 4.2.2

Total number of Yes tuple = 2 and total number of No tuple = 3

$I(p, n) = I(2, 3) = -(2/5)\log_2(2/5) - (3/5)\log_2(3/5) = 0.971$

(i) Compute the entropy for temperature : (Hot, mild, cool)

For Temperature = Hot,

p_i = with "yes" class = 0 and n_i = with "no" class = 2

$$\text{Therefore, } I(p_i, n_i) = I(0,2) = -(0/2)\log_2(0/2) - (2/2)\log_2(2/2) = 0$$

Scanned by CamScanner

Similarly for different outlook ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.2(c)

Temperature	p_i	n_i	$I(p_i, n_i)$
Hot	0	2	0
Mild	1	1	1
Cool	1	0	0

Calculate Entropy using the values from the Table P. 4.2.2(c) and the formula g below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Temperature}) = 2/5 * I(0, 2) + 2/5 * I(1, 1) + 1/5 * I(1, 0) = 0.4$$

Note : T_{sunny} is the total training set.

Hence

$$\text{Gain}(T_{sunny}, \text{temperature}) = I(p, n) - E(\text{temperature})$$

$$= 0.971 - 0.4$$

$$= 0.571$$

(ii) Compute the entropy for humidity : (High, normal)

For Humidity = High,

p_i = with "yes" class = 0 and n_i = with "no" class = 3

Therefore, $I(pi, ni) = I(0,3) = -(0/3)\log2(0/3) - (3/3)\log2(3/3) = 0$

Similarly for different outlook ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.2(d)

Humidity	p_i	n_i	$I(p_i, n_i)$
High	0	3	0
Normal	2	0	0

Calculate Entropy using the
below :

E

Note : T_{sunny} is the total train

Hence

Gain(T_{s}

(iii) Compute the entropy

For wind = weak,

p_i = with "yes" class

Therefore, $I(p_i, n_i)$ =

Similarly for differen

Calculate Entropy

Note : T_{sunny} is the

Hence

Therefore,

Scanned by CamScanner

Calculate Entropy using the values from the Table P. 4.2.2(d) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Humidity}) = 3/5 * I(0, 3) + 2/5 * I(2, 0) = 0$$

Note : T_{sunny} is the total training set.

Hence

$$\text{Gain}(T_{sunny}, \text{Humidity}) = I(p, n) - E(\text{Humidity}) = 0.971 - 0 = 0.971$$

(iii) Compute the entropy for wind : (Weak, strong)

For wind = weak,

p_i = with "yes" class = 1 and n_i = with "no" class = 2

Therefore, $I(p_i, n_i) = I(1,2) = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3) = 0.918$

Similarly for different outlook ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.2(e)

Wind	p_i	n_i	$I(p_i, n_i)$
Weak	1	2	0.918
Strong	1	1	1

Calculate Entropy using the values from the Table P. 4.2.2(e) and the formula given as:

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Wind}) = 3/5 * I(1, 2) + 2/5 * I(1, 1) = 0.951$$

Note : T_{sunny} is the total training set.

Hence

$$\text{Gain}(T_{sunny}, \text{Wind}) = I(p, n) - E(\text{Wind})$$

$$= 0.971 - 0.951 = 0.02$$

Therefore,

$$\text{Gain}(T_{sunny}, \text{Humidity}) = 0.970$$

Scanned by CamScanner

$$Gain(T_{sunny}, Temperature) = 0.570$$

$$Gain(T_{sunny}, Wind) = 0.02$$

Humidity has the highest gain; therefore, it is below Outlook = "sunny".

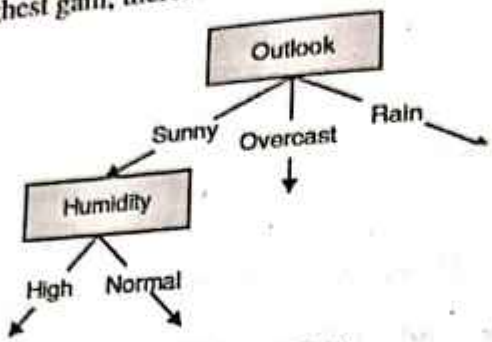


Fig. P. 4.2.2(a)

Step 3 :

Consider now only temperature and wind for outlook = Overcast and count the numbe tuples from the original given training set

$$T_{overcast} = \{3,7,12,13\} = 4 \text{ (From Table P. 4.2.2, outlook = overcast)}$$

Table P. 4.2.2(f)

Day	Outlook	Temperature	Humidity	Wind	Play ball
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

Since for the attributes temperature and wind, playball = yes, so assign class 'yes' t overcast.

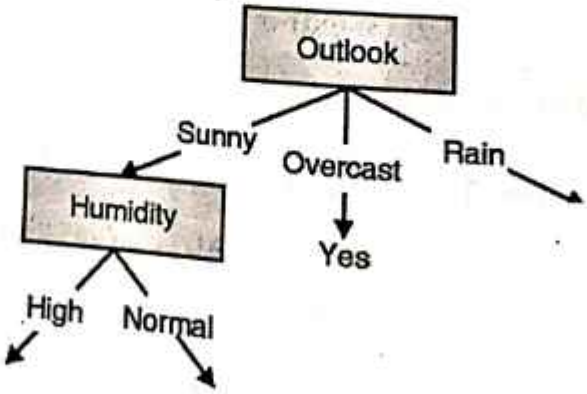


Fig. P. 4.2.2(b)

Scanned by CamScanner

Step 4 :

Consider temperature and wind for outlook = Rain and count the number of tuples from the original given training set

T_{rain} = {4, 5, 6, 10, 14}

= 5 (From Table P. 4.2.2, outlook = rain)

Table P. 4.2.2(g)

Day	Outlook	Temperature	Humidity	Wind	Play ball
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Consider the Table P. 4.2.2(g) as the new training set and calculate the Gain for temperature and Wind.

Class P : Playball = "yes"

Class N : Playball = "no"

Total number of records 5

Count the number of records with "yes" class and "no" class.

So number of records with "yes" class = 3 and "no" class = 2

So Information gain = $I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$

$I(p, n) = I(3, 2) = -(3/5)\log_2(3/5) - (2/5)\log_2(2/5) = 0.970$

(i) **Compute the entropy for Wind**

For Wind = Weak

p_i = with "yes" class = 3 and n_i = with "no" class = 0

Therefore, $I(p_i, n_i) = I(3,0) = 0$.

For Wind = strong

p_i = with "yes" class = 0 and n_i = with "no" class = 2

Scanned by CamScanner

Therefore, $I(p_i, n_i) = I(0, 2) = 0$

Similarly for different outlook ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.2(h)

Wind	p_i	n_i	$I(p_i, n_i)$
Weak	3	0	0
Strong	0	2	0

Calculate entropy using the values from the Table P. 4.2.2(h) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Wind}) = \frac{3}{5} I(3, 0) + \frac{2}{5} I(0, 2) = 0$$

$$\text{Hence } \text{Gain}(T_{rain}, \text{Wind}) = I(p, n) - E(\text{Wind}) = 0.970 - 0 = 0.970$$

(ii) Compute the entropy for Temperature : (Hot, mild , cool)

For Temperature = Hot,

p_i = with "yes" class = 0 and n_i = with "no" class = 0

Therefore, $I(p_i, n_i) = I(0,0) = 0$

Similarly for different outlook ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.2(i)

Temperature	p_i	n_i	$I(p_i, n_i)$
Hot	0	0	0
Mild	2	1	0.918
Cool	1	1	1

Calculate Entropy using the values from the Table P. 4.2.2(i) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

Scanned by CamScanner

$$E(\text{Temperature}) = 0/5 * I(0, 0) + 3/5 * I(2, 1) + 2/5 * I(1, 1)$$
$$= 0.951$$

Note : T_{Rain} is the total training set.

Hence

$$\text{Gain}(T_{Rain}, \text{temperature}) = I(p, n) - E(\text{temperature})$$
$$= 0.970 - 0.951 = 0.019$$

Therefore,

$$\text{Gain}(T_{rain}, \text{Temperature}) = 0.019$$
$$\text{Gain}(T_{rain}, \text{Wind}) = 0.970$$

Wind has the highest gain; therefore, it is below outlook = "rain".

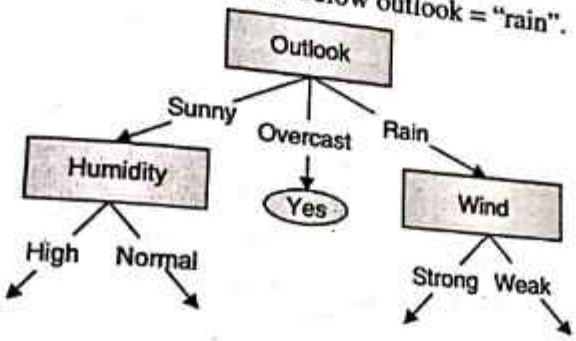


Fig. P. 4.2.2(c)

Therefore the final decision tree is :

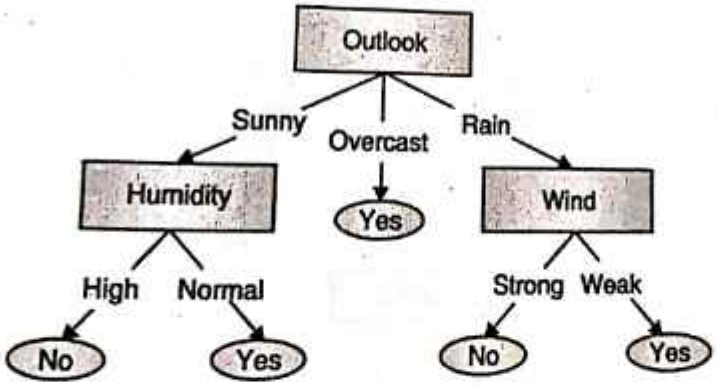


Fig. P. 4.2.2(d) : Decision tree for play tennis

The decision tree can also be expressed in rule format

- IF outlook = sunny AND humidity = high THEN playball = no

Scanned by CamScanner

- IF outlook = Sunny AND humidity = normal THEN playball = yes
- IF outlook = overcast THEN playball = yes
- IF outlook = rain AND wind = strong THEN playball = no
- IF outlook = rain AND wind = weak THEN playball = yes

Ex. 4.2.3 : A sample training dataset for stock market is given below. Profit is the attribute and value is based on age, contest and type.

Age	Contest	Type	Profit
Old	Yes	Swr	Down
Old	No	Swr	Down
Old	No	Hwr	Down
Mid	Yes	Swr	Down
Mid	Yes	Hwr	Down
Mid	No	Hwr	Up
Mid	No	Swr	Up
New	Yes	Swr	Up
New	No	Hwr	Up
New	No	Swr	Up

Soln. :

In the stock market case the decision tree is :

Fig. P. 4.2.3

Scanned by CamScanner

Ex. 4.2.4 : Using the following training data set. Create classification model using decision-tree and hence classify following tuple.

Table P. 4.2.4

Tid	Income	Age	Own House
1.	Very High	Young	Yes
2.	High	Medium	Yes
3.	Low	Young	Rented
4.	High	Medium	Yes
5.	Very high	Medium	Yes
6.	Medium	Young	Yes
7.	High	Old	Yes
8.	Medium	Medium	Rented
9.	Low	Medium	Rented
10.	Low	Old	Rented
11.	High	Young	Yes
12.	medium	Old	Rented

Soln. :

Class P : Own house = "yes"

Class N: Own house = "rented"

Total number of records 12

Count the number of records with "yes" class and "rented" class.

So number of records with "yes" class = 7 and "no" class = 5

So Information gain $= I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$

$$I(p, n) = I(7, 5) = -(7/12)\log_2(7/12) - (5/12)\log_2(5/12) = 0.979$$

Step 1 : Compute the entropy for Income : (Very high, high, medium, low)

For Income = Very high,

p_i = with "yes" class = 2 and n_i = with "no" class = 0

Therefore, $I(p_i, n_i) = I(2,0) = 0$

Scanned by CamScanner

Similarly for different Income ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.4(a)

Income	p_i	n_i	$I(p_i, n_i)$
Very high	2	0	0
High	4	0	0
Medium	1	2	0.918
Low	0	3	0

Calculate entropy using the values from the Table P. 4.2.4(a) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{Income}) = 2/12 * I(2,0) + 4/12*I(4,0) + 3/12*I(0,3) = 0.229$$

Note : S is the total training set.

Hence

$$\text{Gain}(S, \text{Income}) = I(p, n) - E(\text{Income})$$

$$= 0.979 - 0.229 = 0.75$$

Step 2 : Compute the entropy for Age : (Young , medium, old)

Similarly for different age ranges $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.4(b)

Age	p_i	n_i	$I(p_i, n_i)$
Young	3	1	0.811
Medium	3	2	0.971
Old	1	2	0.918

Calculate entropy using the values from the Table P. 4.2.4(b) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

Scanned by CamScanner

$$E(Age) = 4/12 * I(3,1) + 5/12*I(3,2) + 3/12*I(1,2)$$
$$= 0.904$$

Note : S is the total training set.

Hence

$$Gain(S, age) = I(p, n) - E(age)$$
$$= 0.979 - 0.904$$
$$= 0.075$$

Income attribute has the highest gain, therefore it is used as the decision attribute in the root node.

Since income has four possible values, the root node has four branches (very high, high, medium, low).

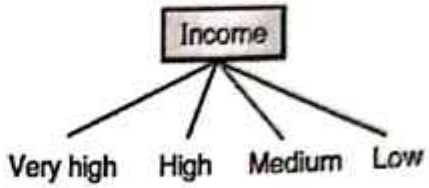


Fig. P. 4.2.4

Step 3 :

Since we have used income at the root, now we have to decide on the age attribute.

Consider income = "very high" and count the number of tuples from the original given training set

$$S_{very\ high} = 2$$

Since both the tuples have class label ="yes" , so directly give "yes" as a class label below "very high".

Similarly check the tuples for income= "high" and income = "low", are having the class label "yes" and "rented" respectively.

Now check for income = "medium", where number of tuples having "yes" class label is 1 and tuples having "rented" class label are 2.

So put the age label below income = "medium".

So the final decision tree is :

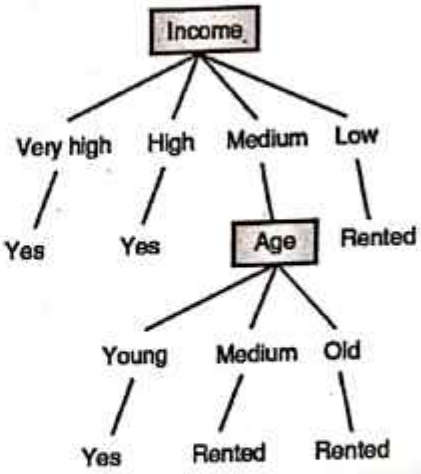


Fig. P. 4.2.4(a)

Ex. 4.2.5 : Data Set : A set of classified objects is given as below. Apply ID3 to generate tree.

Scanned by CamScanner

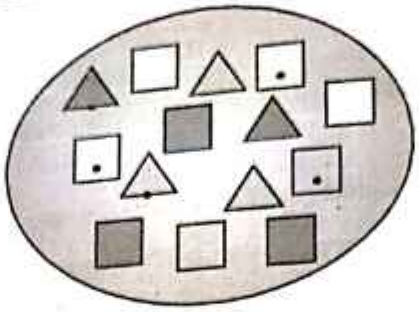


Fig. P. 4.2.5

Table P. 4.2.5

	Attribute			
Sr. No	Colour	Outline	Dot	Shape
1	Green	Dashed	No	Triangle
2	Green	Dashed	Yes	Triangle
3	Yellow	Dashed	No	Square
4	Red	Dashed	No	Square
5	Red	Solid	No	Square
6	Red	Solid	Yes	Triangle
7	Green	Solid	No	Square
8	Green	Dashed	No	Triangle
9	Yellow	Solid	Yes	Square
10	Red	Solid	No	Square
11	Green	Solid	Yes	Square
12	Yellow	Dashed	Yes	Square
13	Yellow	Soild	No	Square
14	Red	Dashed	yes	Triangle

Soln. :

Class N : Shape = "Triangle"

Class P: Shape = "Square"

Total number of records 14

Count the number of records with "triangle" class and "square" class.

Scanned by CamScanner

So number of records with "triangle" class = 5 and "square" class = 9

$$P(\text{square}) = 9/14$$

$$P(\text{triangle}) = 5/14$$

So information gain = I (p, n) = $-\frac{p}{p+n}\log2\frac{p}{p+n} - \frac{n}{p+n}\log2\frac{n}{p+n}$

$$I(p, n) = I(9, 5) = -(9/14)\log2(9/14) - (5/14)\log2(5/14)$$

$$= 0.940$$

Step 1 : Compute the entropy for Color : (Red, green, yellow)

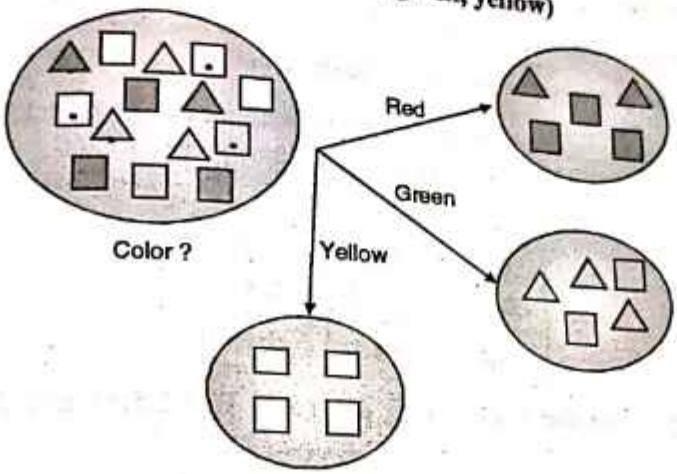


Fig. P. 4.2.5(a)

For color = Red,

p_i = with "square" class = 3 and n_i = with "triangle" class = 2

Therefore, $I(p_i, n_i) = I(3,2) = 0.971$

Similarly for different Color values, $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.5(a)

Color	p_i	n_i	$I(p_i, n_i)$
Red	3	2	0.971
Green	2	3	0.971
Yellow	4	0	0

Scanned by CamScanner

Calculate Entropy using the values from the Table P. 4.2.5(a) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{Color}) = 5/14 * I(3, 2) + 5/14*I(2, 3) + 4/14*I(4, 0)$$

$$= 0.694$$

Note : S is the total training set.

Hence $\text{Gain}(S, \text{color}) = I(p, n) - E(\text{Color}) = 0.940 - 0.694 = 0.246$

Step 2 : Compute the entropy for outline : (Dashed, solid)

Similarly for different outline values, $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.5(b)

Outline	p_i	n_i	$I(p_i, n_i)$
Dashed	3	4	0.985
Solid	6	1	0.621

Calculate Entropy using the values from the Table P. 4.2.5(b) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{Outline}) = 7/14 * I(3,4) + 7/14*I(6,1) = 0.803$$

Note : S is the total training set.

Hence

$$\text{Gain}(S, \text{Outline}) = I(p, n) - E(\text{Outline})$$

$$= 0.940 - 0.803 = 0.137$$

Step 3 : Compute the entropy for dot : (no, yes)

Similarly for different dot values, $I(p_i, n_i)$ is calculated as given below :

Scanned by CamScanner

Table P. 4.2.5(c)

Outline	p_i	n_i	$I(p_i, n_i)$
No	6	2	0.811
Yes	3	3	1

Calculate entropy using the values from the Table P. 4.2.5(c) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Dot}) = 8/14 * I(6,2) + 6/14 * I(3,3)$$

$$= 0.892$$

Note : S is the total training set.

Hence

$$\text{Gain}(S, \text{dot}) = I(p, n) - E(\text{dot})$$

$$= 0.940 - 0.892 = 0.048$$

Therefore,

$$\text{Gain}(S, \text{color}) = 0.246$$

$$\text{Gain}(S, \text{outline}) = 0.137$$

$$\text{Gain}(S, \text{dot}) = 0.048$$

As color has highest gain, it should be the root node.

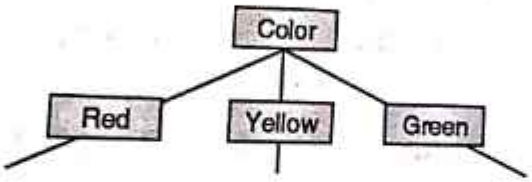


Fig. P. 4.2.5(b)

Step 4 : As attribute color is at the root, we have to decide on the remaining two attribute for red branch node.

Consider color = red and count the number of tuples from the original given training set

Scanned by CamScanner

Table P. 4.2.5(d)

	Attribute			Shape
	Color	Outline	Dot	
1.	Red	Dashed	No	Square
2.	Red	Solid	No	Square
3.	Red	Solid	Yes	Triangle
4.	Red	Solid	No	Square
5.	Red	Dashed	Yes	Triangle

Note : Refer Table P. 4.2.5(d) :

Total number of tuple with "square" class = 3 and total number of No tuple with "triangle" class = 2

$I(p, n) = I(3,2) = -(3/5)\log_2(3/5) - (2/5)\log_2(2/5) = 0.971$

Compute the entropy for outline : (Dashed, solid)

Similarly for different outline values, $I(p_i, n_i)$ is calculated as given below:

Table P. 4.2.5(e)

Outline	p_i	n_i	$I(p_i, n_i)$
Dashed	1	1	1
Solid	2	1	0.918

Calculate Entropy using the values from the Table P. 4.2.5(e) and the formula given as :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Outline}) = 2/5 * I(1,1) + 3/5 * I(2,1)$$

$$= 0.951$$

Hence

$$\text{Gain}(S_{red}, \text{Outline}) = I(p, n) - E(\text{Outline})$$

$$= 0.971 - 0.951 = 0.02$$

Scanned by CamScanner

Compute the entropy for Dot : (no,yes)

Similarly for different Dot values, $I(p_i, n_i)$ is calculated as given below :

Table P. 4.2.5(f)

Outline	p_i	n_i	$I(p_i, n_i)$
No	3	0	0
Yes	0	2	0

Calculate entropy using the values from the Table P. 4.2.5(f) and the formula given below

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(Dot) = 3/5 * I(3,0) + 2/5 * I(0,2)$$

$$= 0$$

Hence

$$Gain(S_{red}, Dot) = I(p, n) - E(Dot)$$

$$= 0.971 - 0 = 0.971$$

Dot has the highest gain; therefore, it is below Color = "Red"

Check the tuples with Dot = "yes" from sample S_{red}, it has class triangle

Check the tuples with Dot = "no" from sample S_{red}, it has class square

So the partial tree for red color sample is as given in Fig. P. 4.2.5(c).

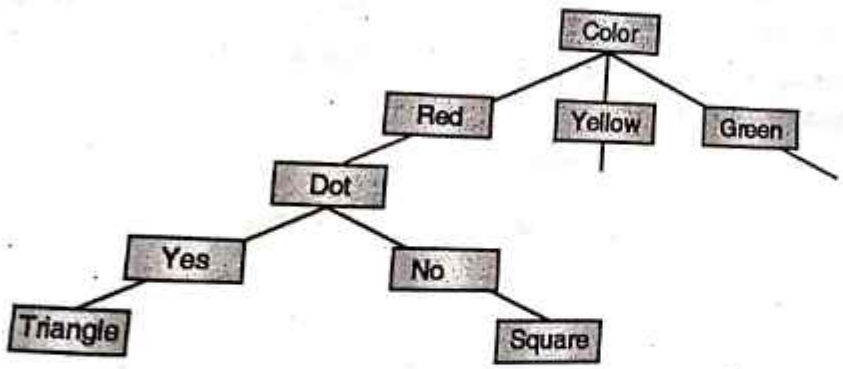


Fig. P. 4.2.5(c)

Scanned by CamScanner

Step 5 : Consider Color = Yellow and count the number of tuples from the original given training set.

Table P. 4.2.5(g)

	Attribute			Shape
	Color	Outline	Dot	
1.	Yellow	Dashed	No	Square
2.	Yellow	Solid	Yes	Square
3.	Yellow	Dashed	Yes	Square
4.	Yellow	Solid	No	Square

As all the tuples belong to yellow color have class label square, so directly assign a class label below the node color = "yellow" as square.

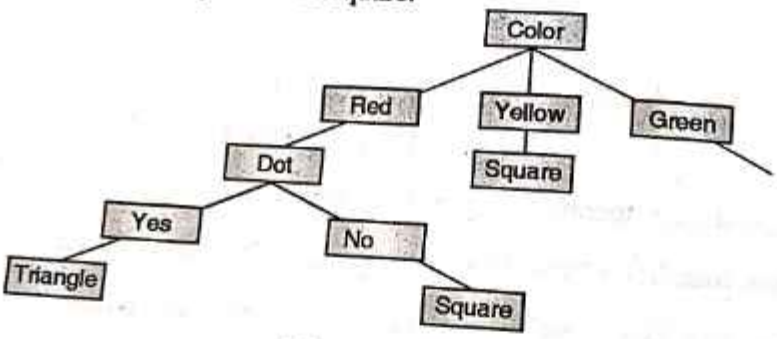


Fig. P. 4.2.5(d)

Step 6 : Consider Color = green and count the number of tuples from the original given training set, as only attribute outline has left, it becomes a node below color = "green".

Table P. 4.2.5(h)

	Attribute			Shape
	Color	Outline	Dot	
1.	green	dashed	no	Triangle
2.	green	dashed	Yes	triangle
3.	green	solid	No	square
4.	green	dashed	no	triangle
5.	green	solid	yes	Square

Scanned by CamScanner

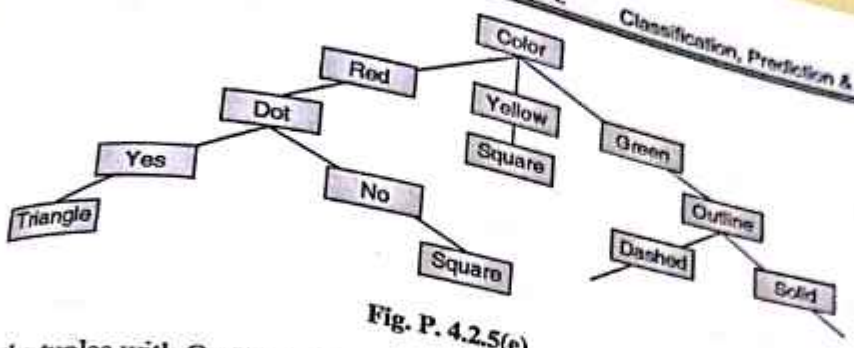


Fig. P. 4.2.5(e)

Check the tuples with Outline = "dashed" from sample S_{green}, it has class triangle.

Check the tuples with outline = "solid" from sample S_{green}, it has class square.

Therefore the final Decision Tree is

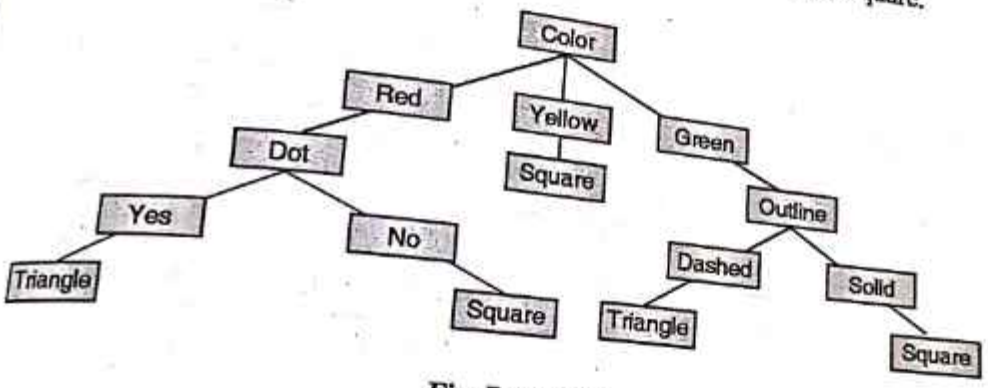


Fig. P. 4.2.5(f)

Ex. 4.2.6 : Apply statistical based algorithm to obtain the actual probabilities of each event to classify the new tuple as a tall. Use the following data.

Person ID	Name	Gender	Height	Class
1	Kristina	Female	1.6m	Short
2	Jim	Male	2m	Tall
3	Maggie	Female	1.9m	Medium
4	Martha	Female	1.85m	Medium
5	John	Male	2.8m	Tall
6	Bob	Male	1.7m	Short
7	Clinton	Male	1.8m	Medium
8	Nyssa	Female	1.6m	Short
9	Kathy	Female	1.65m	Short

Scanned by CamScanner

Soln. :

P(Short) = 4/9 P(Medium) = 3/9 P(Tall) = 2/9

Divide the height attribute into six ranges as given below :

[0,1.6], [1.6,1.7], [1.7,1.8], [1.8,1.9], [1.9,2.0], [2.0,infinity]

Gender attribute has only two values Male and Female.

Total Number of short person = 4, Medium = 3, Tall = 2

Prepare the probability table as given below :

Attribute	Value	Count			Probabilities		
		Short	Medium	Tall	Short	Medium	Tall
Gender	Male	1	1	2	1 / 4	1/3	2/2
	Female	3	2	0	3 /4	2/3	0/2
Height	[0,1.6]	2	0	0	2 /4	0	0
	[1.6,1.7]	2	0	0	2/4	0	0
	[1.7,1.8]	0	1	0	0	1/3	0
	[1.8,1.9]	0	2	0	0	2 /3	0
	[1.9,2.0]	0	0	1	0	0	1 /2
	[2.0,infinity]	0	0	1	0	0	1 /2

Use above values to classify new tuple as a tall :

Consider new tuple as t = {Adam, M, 1.95m}

$$P(t|Short) = 1/4 * 0 = 0$$

$$P(t|Medium) = 1/3 * 0 = 0$$

$$P(t|Tall) = 2/2 * 1/2 = 0.5$$

Therefore likelihood of being short = p(t|short)* P(short) = 0 * 4/9 = 0

Likelihood of being Medium = 0 * 3/9 = 0

Likelihood of being Tall = 2/9 * 1/2 = 0.11

Then estimate P(t) by adding individual likelihood values since t will be either short or medium or tall.

$$P(t) = 0 + 0 + 0.11 = 0.11$$

Scanned by CamScanner

Finally Actually probabilities of each event

$$P(\text{Short} \mid t) = (P(t|\text{short}) * p(\text{short}))/P(t) = (0 * 4/9)/0.11 = 0$$

Similarly $P(\text{Medium}|t) = (0 * 3/9)/0.11 = 0$

$$P(\text{Tall}|t) = (0.5 * 2/9)/0.11 = 1$$

New tuple is a Tall as it has the highest probability.

Ex. 4.2.7 : The training data is supposed to be a part of a transportation study regarding mode choice to select Bus, Car or Train among commuters along a major route in a city.

Attributes				Classes
Gender	Car ownership	Travel cost ($)/km	Income level	Transportation mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Cheap	Medium	Train
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car

Suppose we have new unseen records of a person from the same location where the data sample was taken. The following data are called *test data* (in contrast to *training data*) because we would like to examine the classes of these data.

Person name	Gender	Car ownership	Travel cost ($)/km	Income level	Transportation mode
Alex	Male	1	Standard	High	?
Buddy	Male	0	Cheap	Medium	?
Cherry	Female	1	Cheap	High	?

The question is what transportation mode would Alex, Buddy and Cheery use?

Soln. :

Class P : Transportation mode = "Bus"

Class Q : Transportation mode = "Train"

Scanned by CamScanner

Class N : Transportation mode = "Car"

Total no. of records : 10

No. of records with "Bus" class = 4

No. if records with "Train" class = 3

No. of records with "Car" class = 3

So,

$$\text{Information Gain} = I(p,q,n) = -(p/(p+q+n))\log_2(p/(p+q+n)) - (q/(p+q+n))\log_2(q/(p+q+n)) - (n/(p+q+n))\log_2(n/(p+q+n))$$

$$I(p,q,n) = I(4,3,3) = -(0.4)(-1.322) - (0.3)(-1.737) - (0.3)(-1.737)$$

$$I(4,3,3) = 0.5288 + 0.5211 + 0.5211$$

$$\mathbf{I(4,3,3) = 1.571}$$

Step 1 : Compute the entropy of gender : (Male, Female)

For gender = Male $\quad p_i = 3$

$q_i = 1 \quad n_i = 1$

Therefore,

$$I(p_i, q_i, n_i) = I(3,1,1) = -(3/5)\log_2(3/5) - (1/5)\log_2(1/5) - (1/5)\log_2(1/5)$$

$$= 1.371$$

Similarly for different gender $I(p_i, q_i, n_i)$ is calculated as given below :

Table P. 4.2.7

Gender	p_i	q_i	n_i	$I(p_i, n_i)$
Male	3	1	1	1.371
Female	1	2	2	1.522

Calculate Entropy using the values from the Table P. 4.2.7 and the formula given below

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

Scanned by CamScanner

$$E(\text{gender}) = (5/10) * I(3,1,1) + (5/10) * I(1,2,2)$$
$$= 1.447$$

Note : S is the total training set.

Hence

$$\text{Gain}(S, \text{gender}) = I(p, q, n) - E(\text{gender})$$
$$= 1.571 - 1.447 = 0.124$$

Similarly,

$$\text{Gain}(S, \text{Car Ownership}) = 0.535$$
$$\text{Gain}(S, \text{Travel Cost (\$)/Km}) = 1.21$$
$$\text{Gain}(S, \text{Income Level}) = 0.696$$

Travel cost ($)/Km attribute has the highest gain, therefore it is used as the decision attribute in the root node.

Since travel cost ($)/Km has three possible values, the root node has three branches (Cheap, Standard, Expensive).

Since for all the attributes of Travel Cost ($)/Km = expensive, Transportation mode = "Car", so assign class 'Car' to expensive.

Since for all the attributes of Travel Cost ($)/Km = Standard, Transportation mode = "Train", so assign class 'Train' to standard.

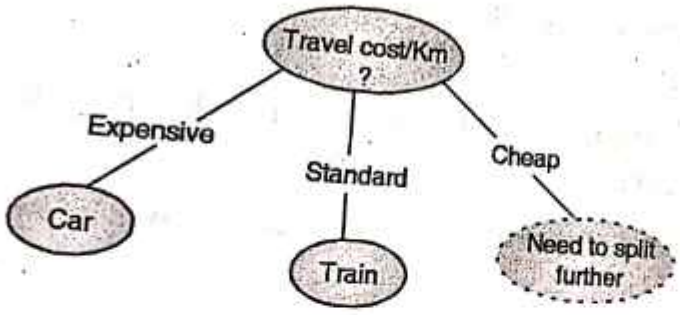


Fig. P. 4.2.7

Consider travel cost ($)/Km = Cheap and count the number of tuples from the original given training set

$$S_{cheap} = 5$$

Scanned by CamScanner

Table P. 4.2.7(a)

Attributes				Classes
Gender	Car ownership	Travel cost ($)/km	Income level	Transportation mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Cheap	Medium	Train
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus

Note : Refer Table P. 4.2.7(a) :

Total No. of Bus tuple = 4 and total no of Train tuple = 1,

and total no of Car tuple = 0

$$I(p, q, n) = (4, 1, 0)$$

$$= -(4/5)\log_2(4/5) - (1/5)\log_2(1/5) - (0/5)\log_2(0/5)$$

$$= 0.722$$

(i) Compute the entropy for gender : (Male, female)

For gender = Male,

p_i = with "Bus" class = 3, q_i = with "Train" class = 0 and n_i with "car" class = 0

Therefore ,

$$I(p_i, q_i, n_i) = I(3,0,0) = -(3/3)\log_2(3/3) - (0/3)\log_2(0/3) - (0/3)\log_2(0/3)$$
$$= 0.$$

Similarly for different genders $I(p_i, q_i)$ is calculated as given below :

Table P. 4.2.7(b)

Gender	p_i	q_i	n_i	$I(p_i, q_i, n_i)$
Male	3	0	0	0
Female	1	1	0	1

Scanned by CamScanner

Calculate entropy using the values from the Table P. 4.2.7(b) and the formula given below

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{gender}) = 3/5 * I(3, 0,0) + 2/5 * I(1, 1,0) = 0.4$$

Note : S_{cheap} is the total training set.

Hence

$$\text{Gain}(S_{cheap}, \text{gender}) = I(p,q, n) - E(\text{gender}) = 0.722 - 0.4 = 0.322$$

(ii) **Compute the entropy for Car ownership : (0, 1, 2)**

For Car ownership = 0,

p_i = with "bus" class = 2 ,q_i = with "train" class = 0 and n_i with "car" class = 0

Therefore,

$$I(p_i, q_i, n_i) = I(2,0,0) = -(2/2)\log_2(2/2) - (0/2)\log_2(0/2) - (0/2)\log_2(0/2) = 0.$$

Similarly for different outlook ranges $I(p_i, q_i, n_i)$ is calculated as given below :

Table P. 4.2.7(c)

Car ownership	p_i	q_i	n_i	$I(p_i, q_i, n_i)$
0	2	0	0	0
1	2	1	0	0.918
2	0	0	0	0

Calculate Entropy using the values from the Table P. 4.2.7(c) and the formula given below

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{Car ownership}) = 2/5 * I(2, 0,0) + 3/5 * I(2, 1,0) + 0/5 * I(0,0,0) = 0.551$$

Note : S_{cheap} is the total training set.

Hence

$$\text{Gain}(S_{cheap}, \text{car ownership}) = I(p,q, n) - E(\text{car ownership}) = 0.722 - 0.551$$

$$= 0.171$$

Scanned by CamScanner

(iii) Compute the entropy for income level : (Low ,medium, high)

For income level = Low,

p_i = with "Bus" class = 2 ,q_i = with "Train" class = 0 and n_i with "car" class = 0

Therefore,

$$I(p_i, q_i, n_i) = I(2,0,0) = -(2/2)\log_2(2/2) - (0/2)\log_2(0/2) - (0/2)\log_2(0/2)$$
$$= 0.$$

Similarly for different outlook ranges $I(p_i, q_i, n_i)$ is calculated as given below :

Table P. 4.2.7(d)

Income level	p_i	q_i	n_i	$I(p_i, q_i, n_i)$
Low	2	0	0	0
Medium	2	1	0	0.918
High	0	0	0	0

Calculate Entropy using the values from the Table P. 4.2.7(d) and the formula given below :

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

E (Income Level) = 2/5 * I (2, 0, 0) + 3/5* I(2, 1,0) + 0/5* I(0,0,0) = 0.551

Note : S_{cheap} is the total training set.

Hence

$$Gain(S_{cheap}, \text{Income level}) = I(p,q,n) - E(\text{Income level})$$
$$= 0.722 - 0.551 = 0.171$$

Therefore, since gender has the highest gain, it comes below cheap.

For all gender = Male, Transportation mode= bus

Scanned by CamScanner

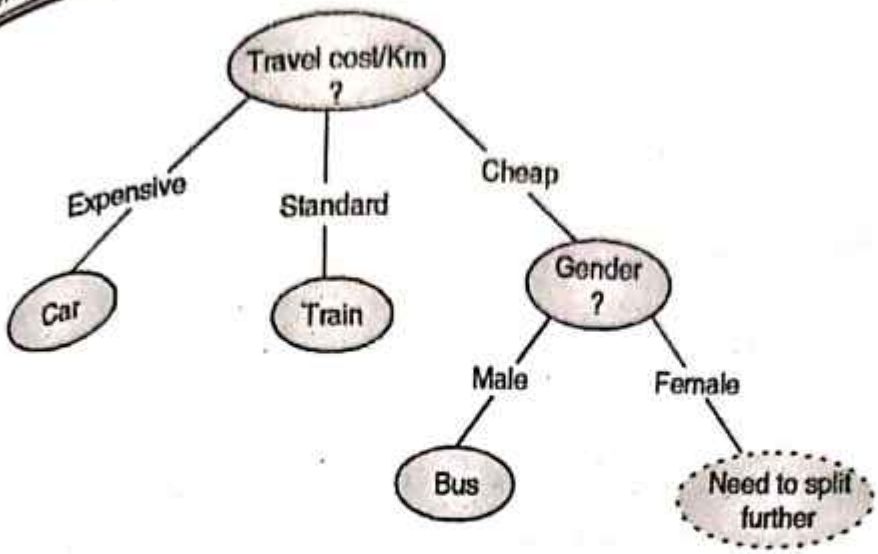


Fig. P. 4.2.7(a)

$S_{female} = 2$

Gender	Car ownership	Income level	Transportation mode
Female	1	Medium	Train
Female	0	Low	Bus

Suppose we select attribute car ownership, we can update our decision tree into the final

ression.

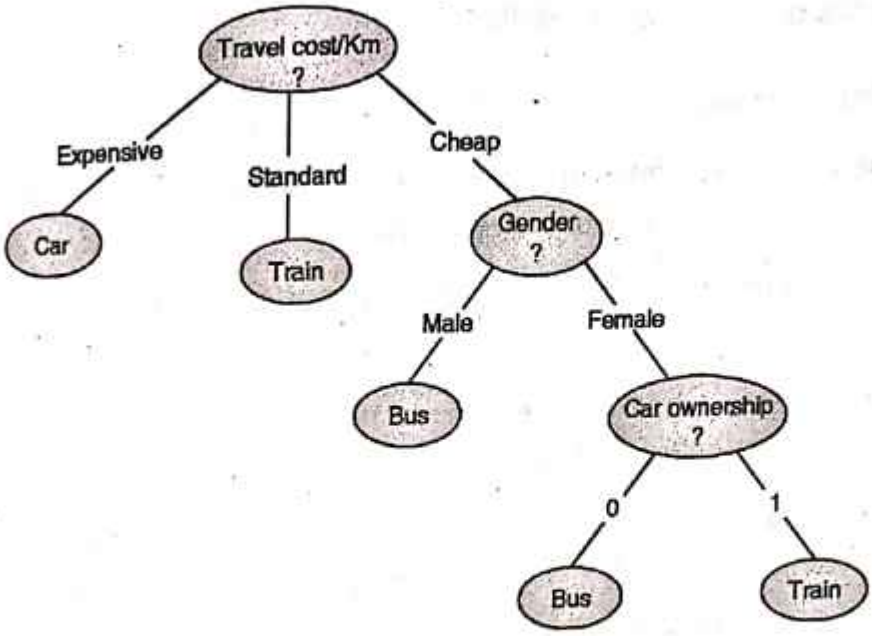


Fig. P. 4.2.7(b)

Scanned by CamScanner

Syllabus Topic : Bayesian Classification - Naïve Bayes

4.2.2 Bayesian Classification : Naïve Bayes Classifier

4.2.2(A) Bayes Theorem

- It is also known as Bayes Rule.
- Bayes theorem is used to find conditional probabilities.
- The conditional probability of an event is a likelihood obtained with the additional information that some other event has previously occurred.
- P(X|Y) is the conditional probability of event X occurring for the event Y which has already occurred.

$$P(X|Y) = P(X \text{ and } Y)/ P(A)$$

- An initial probability is called as **apriori probability** which we get before any additional information is obtained.
- The probability is called as a **posterior probability** value which we get or revised after any additional information is obtained.

4.2.2(B) Basics of Bayesian Classification

- **Probabilistic learning :** Explicit probabilities are calculated for Hypothesis.
- **Incremental :** The probability of a hypothesis whether it is correct can be incrementally increased or decreased by each training example.
- **Probabilistic prediction :** Multiple hypothesis can be predicted by their probability weight.
- **Meta-classification :** The outputs of several classifiers can be combined, e.g. by multiplying the probabilities that all classifiers predict for a given class.
- **Standard :** The computationally intractable Bayesian methods provide a standard of optimal decision making against which other methods can be measured.

Given training data D, posteriori probability of a hypothesis h, P(h|D) follows the Bayes theorem,

$$P(h \mid D) = \frac{P(D \mid h)\, P(h)}{P(D)}$$

P(h) : Independent probability of *h*: prior probability

Scanned by CamScanner

$P(D)$: Independent probability of D

$P(D|h)$: Conditional probability of D given h : likelihood

$P(h|D)$: Conditional probability of h given D : posteriori probability

Practical difficulties

- Require initial knowledge of many probabilities.
- Significant computational cost.

4.2.2(C) Naive Bayes Classifier : Example

Ex. 4.2.8 : Training set is given for play-tennis example.

Outlook	Temperature	Humidity	Windy	Class
Sunny	hot	High	false	No
Sunny	hot	High	true	No
Overcast	hot	High	false	Yes
Rain	mild	High	false	Yes
Rain	cool	Normal	false	Yes
Rain	cool	Normal	true	No
Overcast	cool	Normal	true	Yes
Sunny	mild	High	false	No
Sunny	cool	Normal	false	Yes
Rain	mild	Normal	false	Yes
Sunny	mild	Normal	true	Yes
Overcast	mild	High	true	Yes
Overcast	hot	Normal	false	Yes
Rain	mild	High	true	No

Soln. :

Given a training set, we can compute the probabilities as follows :

- The classification problem may be formalized using a-posteriori probabilities :

$P(C|Y)$ is the probability that the sample tuple $Y = \langle y_1, \ldots, y_k \rangle$ is of class C.

- Assign to sample Y the class label C such that $P(C|Y)$ is maximal.

Scanned by CamScanner

- From the above given sample data, calculate the probabilities for play tennis(P) and don't play tennis(N) for all attributes.

Outlook		
P(sunny|Yes) = 2/9	P(sunny|No) = 3/5	
P(overcast|Yes) = 4/9	P(overcast|No) = 0	P(Yes) = 9/14
P(rain|Yes) = 3/9	P(rain|No) = 2/5	P(No) = 5/14

Temperature	
P(hot|Yes) = 2/9	P(hot|No) = 2/5
P(mild|Yes) = 4/9	P(mild|No) = 2/5
P(cool|Yes) = 3/9	P(cool|No) = 1/5
Humidity	
P(high|Yes) = 3/9	P(high|No) = 4/5
P(normal|Yes) = 6/9	P(normal|No)= 2/5
Windy	
P(true|Yes) = 3/9	P(true|No) = 3/5
P(false|Yes) = 6/9	P(false|No) = 2/5

- An unseen sample Y = <rain, hot, high, false>

$$P(Y|Yes) \cdot P(Yes) = P(rain|Yes) \cdot P(hot|Yes) \cdot P(high|Yes) \cdot P(false|Yes) \cdot P(Yes)$$
$$= 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$$

$$P(Y|No) \cdot P(No) = P(rain|No) \cdot P(hot|No) \cdot P(high|No) \cdot P(false|No) \cdot P(No)$$
$$= 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = 0.018286$$

- Choose the class so that it **maximizes** this probability. This means that the new instance will be classified as no.(don't play)
- Sample Y is classified in class No (i.e. don't play)

An unseen sample = <sunny, cool, high, true >

$$P(Y|Yes) \cdot P(Yes) = P(sunny \mid Yes) \cdot P(cool \mid Yes) \cdot P(high \mid Yes) \cdot P(true \mid Yes) \cdot P(Yes)$$

Scanned by CamScanner

$$= 2/9 \cdot 3/9 \cdot 3/9 \cdot 3/9 \cdot 9/14 = 0.0053$$

$$P(Y|No) \cdot P(No) = P(sunny|No) \cdot P(cool|No) \cdot P(high|No) \cdot P(true|No) \cdot P(No)$$

$$= 3/5 \cdot 1/5 \cdot 4/5 \cdot 3/5 \cdot 5/14 = 0.0206$$

Now choose the class so that it **maximizes** this probability. This means that the new instance will be classified as no. (don't play)

Ex. 4.2.9 : Car theft example : Attributes are color, type, origin and the subject, stolen can be either yes or no.

Data set :

Car No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Soln. :

We want to classify a <Red, Domestic, SUV> i.e. unseen sample. Note there is no example of a <Red, Domestic, SUV> in our data set.

P(Yes) = 5/10
P(No) = 5/10

Color	
P(Red|Yes) = 3/5	P(Red|No) = 2/5
P(Yellow|Yes) = 2/5	P(Yellow|No) = 3/5

Scanned by CamScanner

Type	
P(SUV|Yes) = 1/5	P(SUV|No) = 3/5
P(SPORTS|Yes) = 4/5	P(SPORTS|No) =2/5
Origin	
P(Domestic|Yes) =2/5	P(Domestic|No) =3/5
P(Imported|Yes) =3/5	P(Imported|No) = 2/5

An unseen sample X = <Red, Domestic, SUV>

$$P(X|Yes) \cdot P(Yes) = P(Red|\ Yes) \cdot P(Domestic|\ Yes) \cdot P(Suv|\ Yes) \cdot P(Yes)$$

$$= 3/5 \times 2/5 \times 1/5 \times 5/10 = 0.024$$

$$P(X|No) \cdot P(No) = P(Red|\ No) \cdot P(Domestic|\ No) \cdot P(Suv|\ No) \cdot P(No)$$

$$= 2/5 \times 3/5 \times 3/5 \times 5/10 = 0.072$$

Since 0.072 > 0.024, our example gets classified as 'NO'.

Ex. 4.2.10 : Consider the following data set S, which contains observations of several cases of sunburn :

Name	Hair	Height	Weight	Dublin	Result
Sarah	Blonde	Average	Light	No	Sunburned
Dana	Blonde	Tall	Average	Yes	None
Alex	Brown	Short	Average	Yes	None
Annie	Blonde	Short	Average	No	Sunburned
Emily	Red	Average	Heavy	No	Sunburned
Pete	Brown	Tall	Heavy	No	None
John	Brown	Average	Heavy	No	None
Katie	Brown	Short	Light	Yes	None

Unseen sample X = <brown, tall, average no> Predict the result value as sunburned or none ?

Scanned by CamScanner

Hair	
P(Blonde|Sunburned) = 2/3	P(Blonde| None) = 1/5
P(Brown| Sunburned) = 0	P(Brown| None) = 4/5
P(Red| Sunburned) =1/3	P(Red| None) =0
Height	
P(Average| Sunburned) = 2/3	P(Average| None) = 0
P(Tall| Sunburned) = 0	P(Tall| None) = 2/5
P(Short| Sunburned) = 1/3	P(Short| None) = 2/5
Weight	
P(Light| Sunburned) =1/3	P(Light| None) = 1/5
P(Average| Sunburned) =1/3	P(Average| None) = 2/5
P(Heavy| Sunburned) =1/3	P(Heavy| None) = 2/5
Dublin	
P(No| Sunburned) = 3/3	P(No| None) = 2/5
P(Yes| Sunburned) = 0	P(Yes| None) = 3/5

P(Sunburned) = 3/8
P(None) = 5/8

- An unseen sample X = < brown , tall , average ,no >

$$P(X|Sunburned) \cdot P(Sunburned) = P(Brown|\ Sunburned) \cdot P(tall|\ Sunburned) \cdot P(average|\ Sunburned) \cdot P(No|\ Sunburned) \cdot P(sunburned) = 0$$

$$P(X|None) \cdot P(None) = P(Brown|\ None) \cdot P(tall|\ None) \cdot P(average|\ None) \cdot P(No|\ None) \cdot P(None)$$

$$= 0.032$$

nce 0.032 > 0, our example gets classified as 'NONE'.

Scanned by CamScanner

Ex. 4.2.11 : Predict a class label of an unknown sample using Naive Bayesian classifi on the following training dataset from all electronics customer database.

Age	Income	Student	Credit_rating	Class: buys_computer
≤ 30	High	No	Fair	No
≤ 30	High	No	Excellent	No
31...40	High	No	Fair	Yes
> 40	Medium	No	Fair	Yes
> 40	Low	Yes	Fair	Yes
> 40	Low	Yes	Excellent	No
31...40	Low	Yes	Excellent	Yes
≤ 30	Medium	No	Fair	No
≤ 30	Low	Yes	Fair	Yes
> 40	Medium	Yes	Fair	Yes
≤ 30	Medium	Yes	Excellent	Yes
31...40	Medium	No	Excellent	Yes
31...40	High	Yes	Fair	Yes
> 40	Medium	No	Excellent	No

Soln. :

The unknown sample is x={ age= "≤ 30" ,Income= "Medium", Student= "Yes", Credit rating= "Fair"}

Age	
P(≤ 30lyes) = 2/9	P(≤ 30lNo) = 3/5
P(31...40lyes) = 4/9	P(31...40l No) = 0
P(> 40lyes) = 3/9	P(> 40l No) = 2/5
Income	
P(Highlyes) = 2/9	P(HighlNo) = 2/5
P(Mediumlyes) = 4/9	P(Medium| No) = 2/5
P(lowlyes) = 3/9	P(low| No) = 1/5

Scanned by CamScanner

Student	
P(No|yes) = 3/9	P(No|No) = 4/5
P(yes|yes) = 6/9	P(yes|No) = 1/5
Credit_Rating	
P(fair|yes) = 6/9	P(fair|No) = 2/5
P(excellent|yes) =3/9	P(excellent|No) = 3/5

P(yes) = 9/14
P(No) = 5/14

- An unseen sample X = < age= "≤ 30", Income= "Medium", Student= "Yes", Credit rating= "Fair" >

$P(X|Yes) \cdot P(Yes)$ = P(Age ≤ "30"| Yes) · P(Income=Medium| Yes)

· P(Student="yes"| Yes) · P(Credit Rating="fair"| Yes)

· P(yes) = 2/9 . 4/9. 6/9. 6/9. 9/14 = 0.028

$P(X|No) \cdot P(No)$ = P(Age ≤ "30"| No) · P(Income=Medium| No)

· P(Student="yes"| No) · P(Credit Rating="fair"| No)

· P(No) = 0.007

- Since 0.028 > 0.007, Therefore the naive Bayesian classifier predicts Buys computer = "Yes" for sample X.

Ex. 4.2.12 : Using Naive Bayesian classification on the following given training set, classify the unseen tuple (Refund = No, Married, Income = 1 K)

rid	Refund	Marital status	taxable	Evade
1	Yes	Single	125 K	No
2	No	Married	100 K	No
3	No	Single	120 K	No
4	Yes	Married	70 K	No
5	No	Divorced	95 K	Yes
6	No	Married	60 K	No
7	Yes	Divorced	220 K	No

Scanned by CamScanner

rid	Refund	Marital status	taxable	Evade
8	No	Single	85 K	Yes
9	No	Married	75 K	No
10	No	Single	90 K	Yes

P(No) = 7/10

P(Yes) = 3/10

Soln. :

Given a test Record

$$X = (\text{Refund} = \text{No, Married, Income} = 120\text{ K})$$

$$P(X|\text{Class=No}) = P(\text{Refund=No}|\text{Class=No}) \times P(\text{Married}|\text{ Class=No}) \times P(\text{Income=120K}|\text{ Class=No})$$

$$= 4/7 \times 4/7 \times 1/7 = 0.0466$$

$$P(X|\text{Class=Yes}) = P(\text{Refund=No}|\text{ Class=Yes}) \times P(\text{Married}|\text{ Class=Yes}) \times P(\text{Income=120K}|\text{ Class=Yes})$$

$$= 3/3 \times 0 \times 0 = 0$$

Since $P(X|No)\,P(No) > P(X|Yes)\,P(Yes)$

$$0.0466 \times 7/10 > 0 \times 3/10$$

Therefore $P(No|X) > P(Yes|X) \geq \text{Class} = \text{No}$

Ex. 4.2.13 : Consider the training set for the class of mammals and non mammals, using Naive Bayesian classification classify the unseen tuple (Give Birth = Yes, Can fly = No, Live in water = Yes , have legs = No)

Name	Give birth	Can fly	Live in water	Have legs	Class
Human	Yes	No	No	Yes	Mammals
Python	No	No	No	No	Non-mammals
Salmon	No	No	Yes	No	Non-mammals
Whale	Yes	No	Yes	No	Mammals
Frog	No	No	Sometimes	Yes	Non-mammals
Komodo	No	No	No	Yes	Non-mammals
Bat	Yes	Yes	No	Yes	Non-mammals
Pigeon	No	Yes	No	Yes	Mammals
Cat	Yes	No	No	Yes	Mammals

Scanned by CamScanner

Name	Give birth	Can fly	Live in water	Have legs	Class
Leopard shark	Yes	No	Yes	No	Non-mammals
Turtle	No	No	Sometimes	Yes	Non-mammals
Penguin	No	No	Sometimes	Yes	Non-mammals
Porcupine	Yes	No	No	Yes	Mammals
Eel	No	No	Yes	No	Non-mammals
Salamander	No	No	Sometimes	Yes	Non-mammals
Gila monster	No	No	No	Yes	Non-mammals
Platypus	No	No	No	Yes	Non-mammals
Owl	No	Yes	No	Yes	Mammals
Dolphin	Yes	No	Yes	No	Non-mammals
Eagle	No	Yes	No	Yes	Non-mammals

Soln. :

Unseen record is given as :

Give birth	Can fly	Live in water	Have legs	Class
Yes	No	Yes	No	?

A: attributes M: mammals N: non-mammals

$$P(A \mid M) = 5/7 \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.05$$

$$P(A \mid N) = 2/13 \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0084$$

$$P(A \mid M)\,P(M) = 0.05 \times \frac{7}{20} = 0.0175$$

$$P(A \mid N)\,P(N) = 0.008 \times \frac{13}{20} = 0.0052$$

$$P(A|M)P(M) > P(A|N)P(N)$$

Unseen record belongs to class mammals.

Ex. 4.2.14 : Consider the following dataset that helps to predict the RISK of a loan application based on the applicant's CREDIT HISTORY, DEBT and INCOME.

CREDIT HISTORY	DEBT	INCOME	RISK
Bad	Low	0 to 15	High
Bad	High	15 to35	High
Bad	Low	0 to15	High

Scanned by CamScanner

CREDIT HISTORY	DEBT	INCOME	RISK
Unknown	High	15 to 35	High
Unknown	High	0 to15	High
Good	High	0 to15	High
Bad	Low	Over 35	Moderate
Unknown	Low	15 to 35	Moderate
Good	High	15 to 35	Moderate
Unknown	Low	Over 35	Low
Unknown	Low	Over 35	Low
Good	Low	Over 35	Low
Good	High	Over 35	Low
Good	High	Over 35	Low

Soln. :

- Predict the risk for unseen Tuple X = <unknown, high, over 35, moderate>.
- Write down the rule used by Naive Bayes to classify instances, and apply it to the following instance : <Credit History=bad; Debt=low; Income=15to35>. Which class will be returned by Naive Bayes?

Ex. 4.2.15 : Using given table, create classification model using any algorithm and hence classify following tuple <income = medium, credit = good>.

Transaction Id	Income	Credit	Decision
1	Very High	Excellent	AUTHORIZE
2	High	Good	AUTHORIZE
3	Medium	Excellent	AUTHORIZE
4	High	Good	AUTHORIZE
5	Very High	Good	AUTHORIZE
6	Medium	Excellent	AUTHORIZE
7	High	Bad	REQUEST ID
8	Medium	Bad	REQUEST ID
9	High	Bad	REJECT
10	Low	Bad	CALL POLICE

Scanned by CamScanner

P(Very High| AUTHORIZE) = 2/6	P(Very High| REQUEST ID) = 0	P(Very High| REJECT) = 0	P(Very High| CALL POLICE) = 0
P(High| AUTHORIZE) = 2/6	P(High| REQUEST ID) = 1/2	P(High| REJECT) = 1/1	P(High| CALL POLICE) = 0
P(Medium| AUTHORIZE) = 2/6	P(Medium| REQUEST ID) = 1/2	P(Medium| REJECT) = 0	P(Medium| CALL POLICE) = 0
P(Low| AUTHORIZE) = 0	P(Low| REQUEST ID) = 0	P(Low| REJECT) = 0	P(Low| CALL POLICE) = 1/1
P(Excellent| AUTHORIZE)= 3/6	P(Excellent| REQUEST ID) = 0	P(Excellent| REJECT) = 0	P(Excellent| CALL POLICE) = 0
P(Good| AUTHORIZE) = 3/6	P(Good| REQUEST ID) = 0	P(Good| REJECT) = 0	P(Good| CALL POLICE) = 0
P(Bad| AUTHORIZE) = 0	P(Bad| REQUEST ID) = 2/2	P(Bad| REJECT) = 1/1	P(BAd| CALL POLICE) = 1/1

$$P(\text{AUTHORIZE}) = 6/10$$

$$P(\text{REQUEST ID}) = 2/10$$

$$P(\text{REJECT}) = 1/10$$

$$P(\text{CALL POLICE}) = 1/10$$

$$P(X|\text{AUTHORIZE}) \times P(\text{AUTHORIZE}) = 2/6 \times 3/6 \times 6/10 = 0.1$$

$$P(X|\text{REQUEST ID}) \times P(\text{REQUEST ID}) = 0$$

$$P(X|\text{REJECT}) \times P(\text{REJECT}) = 0$$

$$P(X|\text{CALL POLICE}) \times P(\text{CALL POLICE}) = 0$$

Therefore the Naïve Bayesian classifier predicts decision = "AUTHORIZE" for sample X.

Ex. 4.2.16 : Given the training data for height classification, classify the tuple, t = <Rohit, M, 1.95> using Bayesian Classification.

Name	Gender	Height	Output
Kiran	F	1.6m	Short
Jatin	M	2m	Tall
Madhuri	F	1.09m	Medium
Manisha	F	1.88m	Medium
Shilpa	F	1.7m	Short

Scanned by CamScanner

Name	Gender	Height	Output
Bobby	M	1.85m	Medium
Kavita	F	1.6m	Short
Dinesh	M	1.7m	Short
Rahul	M	2.2m	Tall
Shree	M	2.1m	Tall
Divya	F	1.8m	Medium
Tushar	M	1.95m	Medium
Kim	F	1.9m	Medium
Aarti	F	1.8m	Medium
Rajashree	F	1.75m	Medium

Soln. : Divide the height attribute into six ranges as given below :

[0,1.6] , [1.61,1.7], [1.71,1.8], [1.81,1.9], [1.91,2.0], [2.1, infinity]

Gender attribute has only two values Male and Female.

Total Number of short person = 4

Medium = 8

Tall = 3

Prepare the probability table as given below :

Gender		
P(F|Short) = ¾	P(F|Medium) = 6/8	P(F|Tall) = 0
P(M|Short) = ¼	P(M|Medium) = 2/8	P(M|Tall) = 3/3
Height		
P([0,1.6] | Short) = 2/4	P([0,1.6] | Medium) = 1 /8	P([0,1.6] | Tall) = 0
P([1.61,1.7] | Short) =2/4	P([1.61,1.7] | Medium) = 0	P([1.61,1.7] | Tall) = 0
P([1.71,1.8] | Short) = 0	P([1.71,1.8] | Medium) = 3/8	P([1.71,1.8] | Tall) = 0
P([1.81,1.9] | Short) = 0	P([1.81,1.9] | Medium) = 3/8	P([1.81,1.9] | Tall) = 0
P([1.91,2.0] | Short) = 0	P([1.91,2.0] | Medium) = 1/8	P([1.91,2.0] | Tall) = 1/3
P([2.1, infinity] | Short) = 0	P([2.1, infinity] | Medium) = 0	P([2.1, infinity] | Tall) = 0/3

Scanned by CamScanner

P(Short) = 4/15
P(Medium) = 8/15
P(Tall) = 3/15

The unseen tuple is X = < Name = Rohit , Gender = M , Height = 1.95>

$$P(X|Short) \cdot P(Short) = P(M \mid Short) \times P([1.91,2.0] \mid Short) \times P(Short)$$

$$= 1/4 \times 0 \times 4/5 = 0$$

$$P(X|\ Medium) \cdot P(Medium) = P(M \mid Medium) \times P([1.91,2.0] \mid Medium) \times P(Medium)$$

$$= 2/8 \times 1/8 \times 8/15 = 0.016$$

$$P(X|\ Tall) \cdot P(Tall) = P(M \mid Tall) \times P([1.91,2.0] \mid Tall) \times P(Tall)$$

$$= 3/3 \times 1/3 \times 3/15 = 0.067$$

Since 0.016 < 0.067, Therefore the naïve bayesian classifier predicts Rohit is Tall.

4.2.3 Rule based Classification

- Rule-based classification is featured by building rules based on object attributes. Rule-based classification is a powerful tool for feature extraction, often performing better than supervised classification for many feature types.
- Learned model is represented as a set of IF-THEN rules.
- IF-THEN rule is expressed in the form.
- IF condition THEN conclusion.
- The LHS or "IF" part of the rule is called as "rule antecedent" or "precondition".
- The RHS or "THEN" part is called as "rule consequent".
- Example : IF age = young AND salary = high THEN loan = yes
- This can also be written as :

$$(age = young) \wedge (salary = high) \Rightarrow loan = yes$$

Rule R can be accessed by its coverage and accuracy

$$coverage(R) = n_{covers} / |D|$$

$$accuracy(R) = n_{correct} / n_{covers}$$

Scanned by CamScanner

where n_{covers} = number of tuples covered by R

$n_{correct}$ = number of tuples correctly classified by R

$|D|$ = number of tuples in data set D

- If more than one rule is triggered then it need **conflict resolution.**
- Based on size, it has to order. So give highest priority to that triggering rule which has the maximum attribute test.
- Make the decision list based on the ordering of the rules. Rules are organized based on some measure of rule quality or by taking expert opinion.

Extract the rules from decision tree

- Once the decision tree is created, list the rules which are easy to understand than big and complex tree.
- For every path of the tree, create a rule from root node to a leaf node.
- The last node or leaf node gives the class label.

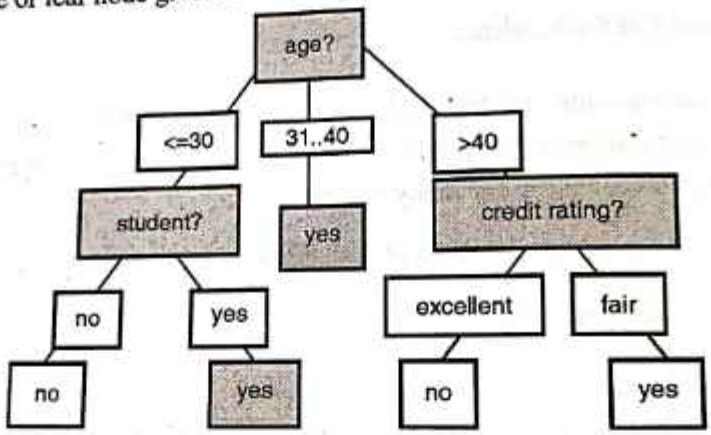


Fig. 4.2.2 : Decision Tree for "Buys_Computer"

Rule extraction from above buys_computer decision-tree

1. IF age = "≤ 30" AND student = "no" THEN buys_computer = "no"
2. IF age = "≤ 30" AND student = "yes" THEN buys_computer = "yes"
3. IF age = "31…40" THEN buys_computer = "yes"
4. IF age = "> 40" AND credit_rating = "excellent" THEN buys_computer = "no"
5. IF age = "> 40" AND credit_rating = "fair" THEN buys_computer = "yes"

Scanned by CamScanner

4.2.4 Other Classification Methods

- k-nearest neighbor classifier
- Case-based reasoning
- Genetic algorithm
- Rough set approach
- Fuzzy set approaches

Syllabus Topic : Prediction - Simple Linear Regression, Multiple Linear Regression

4.3 Prediction

- Suppose an employee needs to predict how much rise he will get in his salary after 5 years. In this case a model is constructed based on his previous salary values that predicts a continuous-valued function or ordered value.
- Prediction is generally about the future values or the unknown events and it models continuous-valued functions.
- Most commonly used methods for prediction is regression.

4.3.1 Structure of Regression Model

- Regression Model represents reality by using the system of equations.
- Regression model explains relationship between variables and also enables quantification of these relationships.
- It determines the strength of relationship between one dependent variable with the other independent variable using some statistical measure.
- Dependent variable is usually denoted by Y.
- The two basic types of regression :

1. Linear regression
2. Multiple regressions

- The general form of regression is :

Linear regression : $Y = m + nX + u$

Scanned by CamScanner

Multiple regression : $Y = m + n_1X_1 + n_2X_2 + n_3X_3 + \ldots + n_tX_t + u$

Where : Y = The dependent variable which we are trying to predict

X = The independent variable that we are using to predict variable Y

m = The intercept

n = The slope

u = The regression residual.

- In multiple regressions each variable is differentiated with subscripted numbers.
- Regression uses a group of random variables for prediction and finds a mathematical relationship between them. This relationship is depicted in the form of a straight line (linear regression) that approximates all the points in the best way.
- Regression may be used to determine for e.g. price of a commodity, interest rates, the price movement of an asset influenced by industries or sectors.

4.3.2 Linear Regression

Regression tries to find the mathematical relationship between variables, if it is a straight line then it is a linear model and if it gives a curved line then it is a non linear model.

Simple linear regression

- The relationship between dependent and independent variable is described by straight line and it has only one independent variable

$$Y = \alpha + \beta X$$

- Two parameters, α and β specify the (Y-intercept and slope of the) line and are to be estimated by using the data at hand.
- The value of Y increases or decreases in a linear manner as the value of X changes accordingly.
- Draw a line relating to Y and X which is well fitted to given data set.
- The ideal situation is that if the line which is well fitted for all the data points and no error for prediction.
- If there is random variation of data points, which are not fitted in a line then construct a probabilistic model related to X and Y.
- Simple linear regression model assumes that data points deviates about the line, as shown in the Fig. 4.3.1.

Scanned by CamScanner

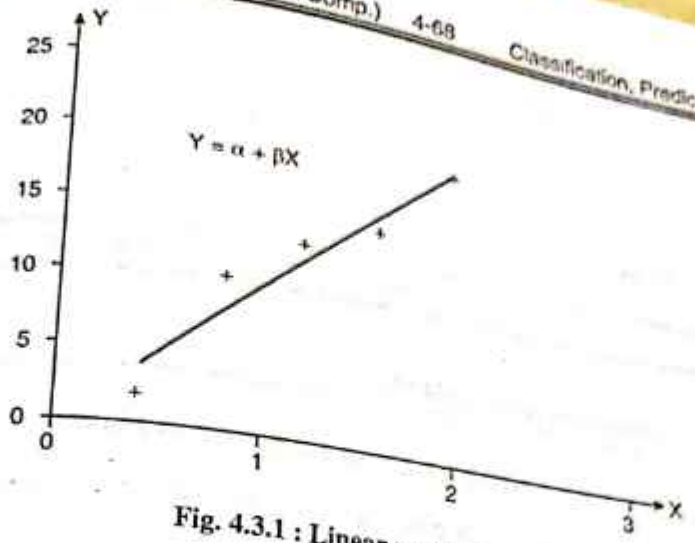


Fig. 4.3.1 : Linear regression

4.3.3 Multiple Linear Regression

- Multiple linear regression is an extension of simple linear regression analysis.
- It uses two or more independent variables to predict the outcome and a single continuous dependent variable

$$Y = a_0 + a_1 X_1 + a_2 X_2 + ... + a_k X_k + e$$

Where, Y is the dependent variable or response variable

$X_1, X_2 .....X_k$ are the independent variables or predictors.

e is random error.

$a_0, a_1, a_2..... a_k$ are the regression coefficients.

4.3.4 Other Regression Model

- In log linear regression a best fit between the data and a log linear model is found.
- Major assumption : A linear relationship exists between the log of the dependent and independent variables.
- Log linear models are models that postulate a linear relationship between the independent variables and the logarithm of the dependent variable, for example :

$$\log(y) = a_0 + a_1 x_1 + a_2 x_2 ... + a_N x_N$$

where y is the dependent variable; x_i, i=1,...,N are independent variables and $\{a_i, i=0,...,N\}$ are parameters (coefficients) of the model.

Scanned by CamScanner

- For example, log linear models are widely used to analyze categorical data represented as a contingency table. In this case, the main reason to transform frequencies (counts) or probabilities to their log-values is that, provided the independent variables are not correlated with each other, the relationship between the new transformed dependent variable and the independent variables is a linear (additive) one.

Syllabus Topic : Model Evaluation and Selection

4.4 Model Evaluation and Selection

- Validation test data is very useful to estimate the accuracy of model.
- Various methods for estimating a classifier's accuracy are given below. All of them are based on randomly sampled partitions of data :
  - Holdout method
  - Random subsampling
  - Cross-validation
  - Bootstrap
- If we want to compare classifiers to select the best one then the following methods are used :
  - Confidence intervals,
  - Cost-benefit analysis and Receiver Operating Characteristic (ROC) Curves.

Syllabus Topic : Accuracy and Error Measures

4.4.1 Accuracy and Error Measures

Q. Explain major factors related to performance of DT based data mining techniques.

Accuracy of a classifier M, acc(M) is the percentage of test set tuples that are correctly classified by the model M.

Basic concepts

1. **Partition the data randomly into three sets :** Training set, validation set and test set.
   - Training set is the subset of data used to train/build the model.

Scanned by CamScanner

- Test set is a set of instances that have not been used in the training process. The model's performance is evaluated on unseen data. Testing just estimates the probability of success on unknown data.
- Validation data is used for parameter tuning but it cannot be the test data. Validation data can be the training data, or a subset of training data.
- **Generalization Error :** Model error on the test data.
- **Success :** Instance (record) class is predicted correctly.
- **Error :** Instance class is predicted incorrectly.
- **The confusion matrix :** It is a useful tool for analyzing how well your classifier can recognize tuples of different classes.
- If we have only two way classification then only four classification outcomes are possible which are given below in the form of a confusion matrix :

		Predicted class		
	Class Label	C_1	C_2	**Total**
Actual class	C_1	True Positives (TP)	False Negatives (FN)	**P**
	C_2	False Positives (FP)	True Negatives (TN)	**N**
	Total	**P'**	**N'**	**All**

- **TP :** Class members which are classified as class members.
- **TN :** Class non-members which are classified as non-members.
- **FP :** Class non-members which are classified as class members.
- **FN :** Class members which are classified as class non-members.
- **P :** Number of positive tuples.
- **N :** The number of negative tuples.
- **P' :** The number of tuples that were labeled as positive.
- **N' :** The number of tuples that were labeled as negative.
- **All :** Total number of tuple i.e. TP + FN + FP + TN or P + N or P' + N'

Scanned by CamScanner

5. **Sensitivity :** True Positive recognition rate which is the proportion of positive tuples that are correctly identified.

$$\text{Sensitivity} = TP/P$$

6. **Specificity :** True Negative recognition rate which is the proportion of negative tuples that are correctly identified.

$$\text{Specificity} = TN/N$$

7. **Classifier accuracy or recognition rate :** Percentage of test set tuples that are correctly classified.

$$\text{Accuracy} = (TP + TN)/\text{All}$$

OR

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

Accuracy is also a function of sensitivity and specificity:

$$\text{Accuracy} = \text{Sensitivity}\,\frac{P}{(P + N)} + \text{Specificity}\,\frac{N}{(P + N)}$$

8. **Error rate :** A percentage of errors made over the whole set of instances (records) used for testing.

$$\text{Error rate} = 1 - \text{accuracy, or Error rate} = (FP + FN)/\text{All}$$

Or

$$\text{Error rate} = \frac{FP + FN}{P + N}$$

9. **Precision :** Percentage of tuples which are correctly classified as positive are actual positive. It is a measure of exactness.

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|}$$

10. **Recall :** Percentage of positive tuples which the classifier labelled as positive. It is a measure of completeness.

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|}$$

Scanned by CamScanner

- **F measure (F_1 or F-score) :** Harmonic mean of precision and recall,

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **F_β :** Weighted measure of precision and recall and assigns ß times as much weight to recall as to precision.

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$

where β is a non-negative real number.

- **Classifiers can also be compared with respect to :**
  - Speed
  - Robustness
  - Scalability
  - Interpretability

- **Re-substitution error rate**
  - Re-substitution error rate is a performance measure and is equivalent to training data error rate.
  - It is difficult to get 0% error rate but it can be minimized, so low error rate is always preferable.

Syllabus Topic : Holdout

4.4.2 Holdout

- In holdout method, data is divided into training data set and testing data set (usually 1/3 for testing, 2/3 for training).

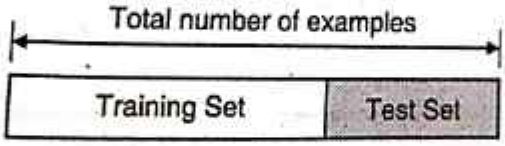


Fig. 4.4.1

- To train the classifier, training data set is used and once the classifier is constructed then use test data set to estimate the error rate of the classifier.

Scanned by CamScanner

- If the training is more than better model is constructed and if the test data is more then more accurate the error estimates.
- **Problem :** The samples might not be representative. For example, some classes might be represented with very few instances or even with no instances at all.
- **Solution :** stratification is the method which ensures that both training and testing data have equal number of samples of same class.

Syllabus Topic : Random Sampling

4.4.3 Random Subsampling

- It is a variation of the holdout method.
- The holdout method is repeated k times.
- Each split randomly selects a fixed number example without replacement.

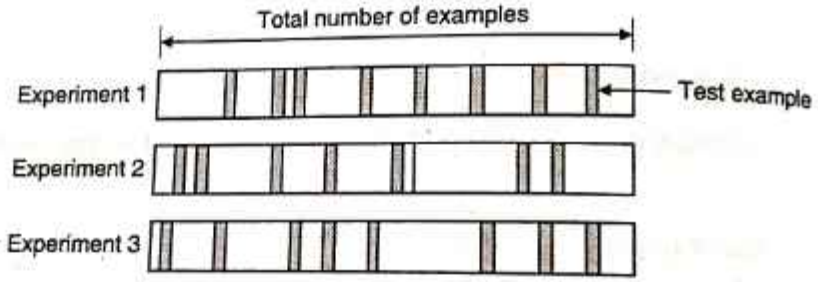


Fig. 4.4.2

- For each data split we retrain the classifier from scratch with the training examples and estimate E_i with the test examples.
- The overall accuracy is calculated by taking the average of the accuracies obtained from each iteration.

$$E = \frac{1}{K} \sum_{i=1}^{K} E_i$$

Syllabus Topic : Cross-Validation

4.4.4 Cross-Validation (CV)

- Avoids overlapping test sets.

Scanned by CamScanner

- **k-fold cross-validation**
  - **First step :** Data is split into k subsets of equal size (usually by random sampling).

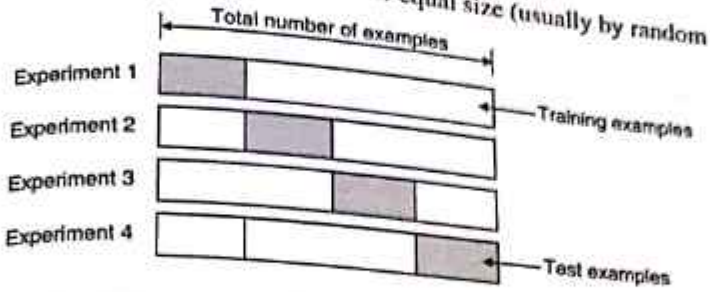


Fig. 4.4.3

  - **Second step :** Each subset in turn is used for testing and the remainder for training.
- The advantage is that all the examples are used for both training and testing.
- The error estimates are averaged to yield an overall error estimate.

$$E = \frac{1}{K} \sum_{i=1}^{K} E_i$$

- **Leave-one-out cross validation**
  - If dataset has N examples, then N experiments to be performed for Leave-one-out cross validation.
  - For every experiment, training uses N-1 examples and remaining example for testing.
- The average error rate on test examples gives the true error.

$$E = \frac{1}{N} \sum_{i=1}^{N} E_i$$

- **Stratified cross-validation:** Subsets are stratified before the cross-validation is performed.
- **Stratified ten-fold cross-validation**
  - This gives accurate estimate of evaluation.
  - The estimate's variance get reduced due to stratification.
  - Ten-fold cross-validation is repeated ten times and finally the results are averaged based on the previous 10 results.

Scanned by CamScanner

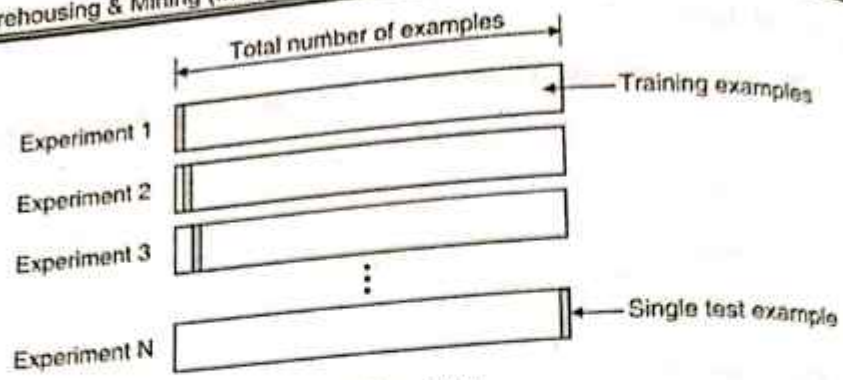


Fig. 4.4.4

Syllabus Topic : Bootstrap

4.4.5 Bootstrapping

- CV uses sampling of data set without replacement. Once the tuple or instance is selected, it cannot be selected again for training or test data.
- The bootstrap uses sampling with replacement to get the training set.
- **Training set :** A dataset of k instances is sampled with replacement k times to form the training set of k instances.
- **Test set :** This is separate dataset from the original dataset which is not the part of training dataset.
- Bootstrapping is the best error estimator for small datasets.

Syllabus Topic : Clustering

4.5 What is Clustering ?

4.5.1 What is Clustering ?

➔ (MU - Dec. 2010, May 2012, Dec. 2012, Dec. 2019)

- Clustering is an unsupervised learning problem.
- Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions.
- It is a process of partitioning data objects into sub classes, which are called as clusters.

Scanned by CamScanner

A cluster contains data objects which have high inter similarity and low intra similarity. We can show this with a simple graphical example :

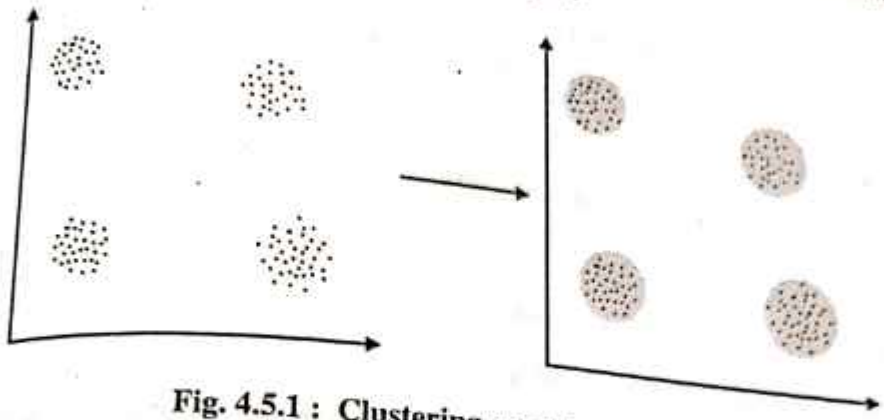


Fig. 4.5.1 : Clustering graphical example

From Fig. 4.5.1, it can be seen that the data objects belong to 4 different clusters. Geometrical distance can be used as similarity criteria to find out which data object would belong to which of the four clusters. This type of clustering is called as distance based clustering.

The other kind of clustering known as conceptual clustering in which data objects in a cluster are a part of it based on a common concept. In other words they fit based on some descriptive concept and not on a similarity measure.

Applications

Clustering algorithms can be applied in many disciplines :

- **Marketing :** Clustering can be used for targeted marketing. For e.g. Given a customer database, containing properties and past buying records. Similar groups of customers can be identified and grouped into one cluster.
- **Biology :** Clustering can also be used in classifying plants and animals into different classes based on their features.
- **Libraries :** Based on different details about books clustering can be used for book ordering.
- **Insurance :** With the help of clustering different groups of policy holders can be identified. For e.g. policy holders with high average claim cost or identifying some frauds.
- **City-planning :** Using details like house type, geographical locations, groups of houses can be identified using clustering.

Scanned by CamScanner

- **Earthquake studies :** Clustering can also be used to identify dangerous zones based on earthquake epicentres.
- **WWW :** Clustering can be used to find groups of similar access patterns using weblog data. It can also be used for classification of documents.

Requirements

The main requirements that a clustering algorithm should satisfy are :

- Scalability.
- Dealing with different types of attributes.
- Discovering clusters with arbitrary shape.
- Minimal requirements for domain knowledge to determine input parameters.
- Ability to deal with noise and outliers.
- Insensitivity to order of input records.
- High dimensionality.
- Interpretability and usability.

Problems

- Because of time complexity, it creates problem to deal with large amount of data items.
- For distance based clustering, the effectiveness of method depends on distance definition.

4.5.2 Categories of Clustering Methods

➔ **(MU - May 2010, May 2012)**

- A good clustering method will produce high quality clusters with :
  - High intra-class similarity
  - Low inter-class similarity
- Major clustering methods can be classified into the following categories :

1. **Partitioning methods :** In Partitioning based approach, various partitions are created and then they are evaluated based on certain criteria.
2. **Hierarchical methods :** The set of data objects are decomposed hierarchically using certain criteria.

3. **Density-bas**
   For e.g. Den
4. **Grid-based**

- Jiawei Han and
  shown in the Ta

Method
Partitioning methods
Hierarchical methods
Density-based methods
Grid-based methods

4.5.3 Differ

Sr. No.	
1.	In con pre

Scanned by CamScanner

3. **Density-based methods :** This approach is based on density (local cluster criteria) For e.g. Density connected points

4. **Grid-based methods :** This approach is based on multi-resolution grid data structure.

Jiawei Han and Kamber has given the overview of above mentioned clustering methods as shown in the Table 4.5.1.

Table 4.5.1 : Overview of various clustering methods

Method	General characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape. – Distance-based. – May use mean or medoid (etc.) to represent cluster center. – Effective for small to medium sized data sets.
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels). – Cannot correct erroneous merges or splits. – May incorporate other techniques like micro-clustering or consider object "linkages".
Density-based methods	– Can find arbitrarily shaped clusters. – Clusters are dense regions of objects in space that are separated by low-density regions. – Cluster density : Each point must have a minimum number of points within its "neighborhood". – May filter out outliers.
Grid-based methods	– Use a multi-resolution grid data structure. – Fast processing time (typically independent of the number of data objects, yet dependent on grid size).

4.5.3 Difference between Classification and Clustering

Sr. No.	Classification	Clustering
1.	In classification, a Training set containing data that have been previously categorized and based on	In clustering, the characteristics of similarity of data is not known in advance so using statistical concepts, we

Scanned by CamScanner

Sr. No.	Classification	Clustering
	this training set, the algorithm finds the category that the new data points belong to.	split the datasets into sub-datasets such that the Sub-datasets have "Similar" data called as clusters.
2.	Classification is *supervised learning*.	Clustering is *unsupervised learning*.
3.	You're given an unseen tuple and you are suppose to set a label or a class to that tuple For example, a company wants to classify their customers. When the company launches a product they want to classify which of their customers will buy their product and which ones will not buy it.	You're given a set of transaction history that gives details about which customer bought what. By using clustering techniques, you can tell the segmentation of your customers.
4.	A common approach for classifiers is to use decision trees to partition and segment records.	There are a variety of algorithms used for clustering, but they all share the property of iteratively assigning records to a cluster unless it indicates that the process has converged to stable segments.

4.6 Types of Data

There are mainly two types of data structures for main memory-based clustering algorithms :

1. Data Matrix or Object by variable structure

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

2. Dissimilarity matrix or by object structure

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Scanned by CamScanner

- Dissimilarity/Similarity metric : Similarity is expressed in terms of a distance function, which is typically metric : $d(i, j)$.
- There is a separate "quality" function that measures the "goodness" of a cluster.
- The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal and ratio variables.
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define "similar enough" or "good enough".
- Type of data in clustering analysis :

(i) Interval-scaled variables
(ii) Binary variables
(iii) Nominal, ordinal, and ratio variables
(iv) Variables of mixed types

4.8.1 Interval-Scaled Variables

- Interval-scaled variables are continuous measurement on a linear scale. Example : weight, height and weather temperature. These attributes allow for ordering, comparing and quantifying the difference between the values. An interval-scaled variables have values whose differences are interpretable
- Consider the measurement for a variable f, this can be performed as follows :
  - Standardize data
  - Calculate the mean absolute deviation :

$$S_f = \frac{1}{h}\left(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|\right)$$

Where m_f is the mean value of f

$$m_f = \frac{1}{h}\left(x_{1f} + x_{2f} + \ldots + x_{nf}\right)$$

  - Calculate the standardized measurement (*z-score*)

$$Z_{if} = \frac{x_{if} - m_f}{s_f}$$

  - Using mean absolute deviation is more robust than using standard deviation.

Scanned by CamScanner

- Distances are normally used to measure the similarity or dissimilarity between two data objects.
- Some popular ones include : *Minkowski distance* :

$$d(i, j) = \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$

- If $q = 2$, d is Euclidean distance :

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2}$$

- Both the Euclidean distance and Manhattan distance satisfy the following mathematic requirements of a distance function :

$$d(i,j) \geq 0$$

$$d(i,i) = 0$$

$$d(i,j) = d(j,i)$$

$$d(i,j) \leq d(i,k) + d(k,j)$$

- Also one can use weighted distance, parametric.

4.6.2 Binary Variable

- A nominal attribute which has either of the two states 0 or 1 is called Binary variable, where 0 means that the attribute is absent and 1 means that it is present.
- **Symmetric binary variable :** If both of its states i.e. 0 and 1 are equally valuable. Here we cannot decide which outcome should be 0 and which outcome should be 1. For example : Marital status of a person is "Married or Unmarried". In this case both are equally valuable and difficult to represent in terms of 0(absent) and 1(present).
- **Asymmetric binary variable :** If the outcome of the states are not equally important. An example of such a variable is the presence or absence of a relatively rare attribute. For example : Person is "handicapped or not handicapped". The most important outcome is usually coded as 1 (present) and the other is coded as 0 (absent).

Scanned by CamScanner

A contingency table for binary data :

		Object n		
		1	0	Sum
object m	1	A	b	a + b
	0	C	d	c + d
	sum	a + c	b + d	P

Here we are comparing two objects, object *m* and object *n*.

a would be the number of variables which are present for both objects.

b would be the number found in object m but not in object *n*.

c is just the opposite to b and *d* is the number that are not found in either object.

- Simple matching coefficient (invariant, if the binary variable is **symmetric**):

$$d(i, j) = \frac{b + c}{a + b + c + d} \quad \text{...(4.6.1)}$$

- Jaccard coefficient (non-invariant if the binary variable is ***asymmetric***):

$$d(i, j) = \frac{b + c}{a + b + c} \quad \text{...(4.6.2)}$$

Example

Table 4.6.1 : A Relational table containing mostly binary values

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute the remaining attributes are asymmetric binary
- Let the values Y and P be set to 1, and the value N be set to 0 as shown in the Table 4.6.2.
- Using formula (Equation 4.6.2) of asymmetric variable.

Scanned by CamScanner

Table 4.6.2

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	1	0	1	0	0	0
Mary	F	1	0	1	0	1	0
Jim	M	1	1	0	0	0	0

Distance between jack and mary (i.e. *d(jack,mary)*) is calculated using formula (Equation 4.6.2) and use contingency table given above :

Consider attributes : Fever, cough, Test-1, Test-2, Test-3, Test-4

Consider jack as object i and mary as object j

a = attribute values 1 in jack and in mary also = 2

b = attribute values 1 in jack but 0 in mary = 0

c = attribute values 0 in jack but 1 in mary = 1

$$d(i, j) = \frac{b+c}{a+b+c}$$

$$d(\text{jack, mary}) = \frac{0+1}{2+0+1} = 0.33$$

Similarly, calculate distance for other combination

$$d(\text{jack, jim}) = \frac{1+1}{1+1+1} = 0.67$$

$$\mathbf{d(jim, mary)} = \frac{\mathbf{1+2}}{\mathbf{1+1+2}} = \mathbf{0.75}$$

So, these measurement shows that, Jack and mary are most likely to have a similar disease as having lowest dissimilarity value among the three pairs. But Jim and mary are unlikely to have a similar disease as having highest dissimilarity value.

4.6.3 Nominal, Ordinal, and Ratio Variables

(1) Nominal Variables

- Nominal variables are also called as Categorical attributes and allow for only qualitative classification.

Scanned by CamScanner

- Every individual item has a certain distinct categories, but quantification or ranking the order of the categories is not possible.
- The nominal variable categories can be numbered arbitrarily.
- Arithmetic and logical operations on the nominal data cannot be performed.
- Typical examples of such attributes are :

Car owner :	1. Yes 2. No
Employment status :	1. Unemployed 2. Employed

Nominal data has no order, and the assignment of numbers to categories is purely arbitrary. Because of lack of order or equal intervals, one cannot perform arithmetic (+, -, /, *) or logical operations (>, <, =) on the nominal data. Typical examples of such variables are:

Gender :	1. Male 2. Female
Marital Status :	1. Unmarried 2. Married 3. Divorcee 4. Widower

(2) Ordinal Variables

- A discrete ordinal variable is a nominal variables, which have meaningful order or rank for its different states.
- The interval between different states is uneven due to which arithmetic operations are not possible, however logical operations may be applied.
- For example, Considering Age as an ordinal attribute, it can have three different states based on an uneven range of age value. Similarly income can also be considered as an ordinal attribute, which is categorised as low, medium, high based on the income value.

Age :	1. Teenage 2. Young 3. Old
Income :	1. Low 2. Medium 3. High

Scanned by CamScanner

(3) Ratio-variables

- Ratio scaled variables are continuous positive measurements on a non linear scale. They are also interval scaled data but are not measured on a linear scale.
- Operations like addition, subtraction can be performed but multiplication and division are not possible.
- For example : For instance, if a liquid is at 40 degrees and we add 10 degrees, it will be 50 degrees. However, a liquid at 40 degrees does not have twice the temperature of a liquid at 20 degrees because 0 degrees does not represent "no temperature"
- There are three different ways to handle the ratio-scaled variables :
  - As interval scale variables. The drawback of handling them as interval scaled is that it can distort the results.
  - As continuous ordinal scale.
  - Transforming the data (for example, logarithmic transformation) and then treating the results as interval scaled variables.

4.6.4 Variable of Mixed Types

- Variable of mixed type may be either interval-scaled, symmetric binary, asymmetric binary, nominal, ordinal or ratio scaled.
- Objects are many times described as a mixture of different variable types and it is used in real time databases.
- In real applications, to compute the dissimilarity between objects of mixed variable types, the preferable approach is to process all variable types together, performing a single cluster analysis.
- It brings all of the meaningful variable onto a common scale of the interval [0.0, 1.0].

Syllabus Topic : Distance Measures

4.7 Distance Measures

➔ **(MU - Dec. 2011)**

- From the scientific and mathematical point of view, *distance* is defined as a quantitative degree of how far apart two objects are. Synonyms for *distance* include dissimilarity.

Scanned by CamScanner

Those distance measures satisfying the metric properties are simply called *metric* while other non-metric distance measures are occasionally called *divergence*. Synonyms for *similarity* include proximity and similarity measures are often called *similarity coefficients*. The choice of distance/similarity measures depends on the measurement type or representation of objects.

Table 4.7.1 : L_p Minkowski family

1.	Euclidean L_2	$d_{Euc} = \sqrt{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^2}$
2.	City block L_1	$d_{CB} = \sum_{i=1}^{d} \lvert P_i - Q_i \rvert$
3.	Minkowski L_p	$d_{Mk} = \sqrt[p]{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^p}$
4.	Chebyshev L_∞	$d_{Cheb} = \max_i \lvert P_i - Q_i \rvert$

- Euclid stated that the shortest distance between two points is a line and thus the equation (1) is predominantly known as Euclidean distance. It was often called Pythagorean metric since it is derived from the Pythagorean Theorem. In the late 19th century, Hermann Minkowski considered the city block distance. Other names for the equation (2) include rectilinear distance, taxicab norm, and Manhattan distance. Hermann also generalized the formulae (1) and (2) to the equation (3) which is coined after Minkowski. When *p* goes to infinite, the equation (4) can be derived and it is called the chessboard distance in 2D, the minimax approximation, or the Chebyshev distance named after Pafnuty Lvovich Chebyshev.

- These distances (similarities) can be based on a single dimension or multiple dimensions, with each dimension representing a rule or condition for grouping objects.

- For example, if we were to cluster fast foods, we could take into account the number of calories they contain, their price, subjective ratings of taste, etc. The most straightforward way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances.

- If we had a two or three-dimensional space this measure is the actual geometric distance between objects in the space (i.e., as if measured with a ruler).

Scanned by CamScanner

Syllabus Topic : Partitioning Methods (k-means, k-medoids)

4.8 Partitioning Methods

→ (MU - Dec. 2010)

Partitioning methods construct a partition of a database D of n objects into a set of K clusters.

Different partitioning methods

1. **Global optimal method :** Exhaustively enumerate all partitions.
2. **Heuristic methods :** K-means and K-medoids algorithms.
   - **K-means :** Each cluster is represented by the centre of the cluster.
   - **K-medoids or PAM (Partitioning Around Medoids) :** Each cluster is represented by one of the objects in the cluster.

4.8.1 K-means Clustering : (Centroid Based Technique)

→ **(MU - May 2010, Dec. 2012, May 2013, Dec. 2013, May 2014, Dec. 2014, May 2016)**

- In 1967, J. MacQueen and then in 1975 J. A. Hartigan and M. A. Wong developed K-means clustering algorithm.
- In K-means approach the data objects are classified based on their attributes or features into k number of clusters. The number of clusters i.e. K is an input given by the user.
- K-means is one of the simplest unsupervised learning algorithms.
- Define K centroids for K clusters which are generally far away from each other.
- Then Group the elements into clusters, which are nearer to the centroid of that cluster.
- After this first step, again calculate the new centroid for each cluster based on the elements of that cluster.
- Follow the same method and group the elements based on new centroid.
- In every step, the centroid changes and elements move from one cluster to another.
- Do the same process till no element is moving from one cluster to another i.e. till two consecutive steps with same centroid and same elements are obtained.

Scanned by CamScanner

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function is given below,

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - c_j\|^2$$

Where, $x_i^{(j)}$ = A data point

c_j = The cluster centre

n = Number of data points

k = Number of clusters

$\|x_i^{(j)} - c_j\|^2$ = Distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j

K-means algorithm

- k : number of clusters
- n : sample feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$
- m_i : the mean of the vectors in cluster i
- Assume $k < n$.
- Make initial guesses for the means $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_k$.
- Until there are no changes in any mean.
- Use the estimated means to classify the samples into clusters.
  - for i = 1 to k

    Replace $\mathbf{m}_i$ with the mean of all of the samples for cluster i
  - end_for
  - end_until
- Following three steps are repeated until convergence :
- Iterate till no object moves to a different group

1. Find the centroid coordinate.
2. Find the distance of each object to the centroids.
3. Based on minimum distance group the objects.

Scanned by CamScanner

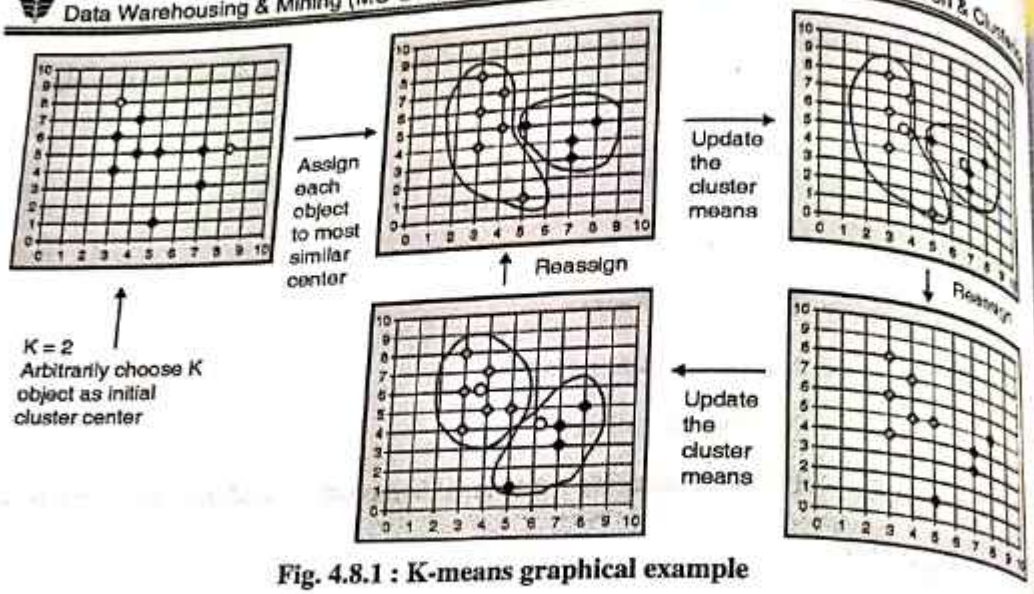


Fig. 4.8.1 : K-means graphical example

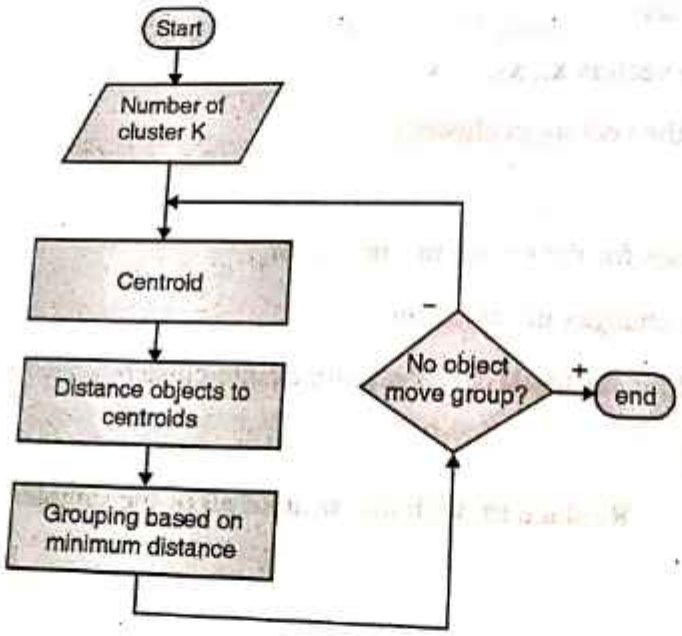


Fig. 4.8.2 : Basic steps for K-means clustering

- Given a cluster $K_i = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$, the cluster mean is $m_i = (1/m)(t_{i1} + \ldots + t_{im})$

Ex. 4.8.1 : Given : {2,4,10,12,3,20,30,11,25}, Assume number of cluster i.e. k = 2.

MU - Dec. 2010, May 2014, 10 Marks

Soln. : Method 1

1. Randomly assign means : $m_1 = 3$, $m_2 = 4$.

Scanned by CamScanner

2. The numbers which are close to mean $m_1 = 3$ are grouped into cluster K_1 and numbers which are close to mean $m_2 = 4$ are grouped into cluster K_2.
3. Again calculate the new mean for new cluster groups.
4. $K_1 = \{2,3\}$, $K_2 = \{4,10,12,20,30,11,25\}$, $m_1 = 2.5$, $m_2 = 16$
5. $K_1 = \{2,3,4\}$, $K_2 = \{10,12,20,30,11,25\}$, $m_1 = 3$, $m_2 = 18$
6. $K_1 = \{2,3,4,10\}$, $K_2 = \{12,20,30,11,25\}$, $m_1 = 4.75$, $m_2 = 19.6$
7. $K_1 = \{2,3,4,10,11,12\}$, $K_2 = \{20,30,25\}$, $m_1 = 7$, $m_2 = 25$
8. $K_1 = \{2,3,4,10,11,12\}$, $K_2 = \{20,30,25\}$
9. Stop as the clusters with these means (in step 7 and 8) are the same. The clusters in the last two groups are identical.
10. So the final answer is $K_1 = \{2,3,4,10,11,12\}$, $K_2 = \{20,30,25\}$.

Method 2

1. Given : {2,4,10,12,3,20,30,11,25}, Randomly assign alternative values to each cluster.
2. Number of cluster = 2, therefore

$K_1 = \{2,10,3,30,25\}$, Mean = 14

$K_2 = \{4,12,20,11\}$, Mean = 11.75

3. Re-assign

$K_1 = \{20,30,25\}$, Mean = 25

$K_2 = \{2,4,10,12,3,11\}$, Mean = 7

4. Re-assign

$K_1 = \{20,30,25\}$, Mean = 25

$K_2 = \{2,4,10,12,3,11\}$, Mean = 7

So the final answer is $K_1 = \{2,3,4,10,11,12\}$, $K_2 = \{20,30,25\}$

Ex. 4.8.2 : Use K-means algorithm to create 3 - clusters for given set of values : {2, 3, 6, 8, 9, 12, 15, 18, 22}

Soln. :

1. 2, 3, 6, 8, 9, 12, 15, 18, 22 – break into 3 clusters (Randomly assign data to three clusters) and calculate the mean value.

Scanned by CamScanner

$K_1 = 2, 8, 15$ – mean = 8.3 ; $K_2 = 3, 9, 18$ – mean = 10

$K_3 = 6, 12, 22$ – mean = 13.3

2. Re-assign

$K_1 = 2, 3, 6, 8, 9$ – mean = 5.6 ; K_2 = mean = 0

$K_3 = 12, 15, 18, 22$ – mean = 16.75

3. Re-assign

$K_1 = 3, 6, 8, 9$ – mean = 6.5 ; $K_2 = 2$ – mean = 2

$K_3 = 12, 15, 18, 22$ – mean = 16.75

4. Re-assign

$K_1 = 6, 8, 9$ – mean = 7.6 ; $K_2 = 2, 3$ – mean = 2.5

$K_3 = 12, 15, 18, 22$ – mean = 16.75

5. Re-assign

$K_1 = 6, 8, 9$ – mean = 7.6 ; $K_2 = 2, 3$ - mean = 2.5

$K_3 = 12, 15, 18, 22$ – mean = 16.75

6. Last two groups are same. So finally we got 3 clusters

Cluster 1 = {6,8,9} , Cluster 2 = { 2,3 } , Cluster 3 = { 12,15,18,22 }

Ex. 4.8.3 : Confer the K-means algorithm with the following data for two clusters. Data Set { 10,4,2,12,3,20,30,11,25,31} **MU - May 2010, 10 Marks**

Soln. :

1. Given : {10,4,2,12,3,20,30,11,25,31}, Randomly assign alternative values to each cluster
2. Number of cluster = 2, therefore

K1 = {10, 2, 3, 30, 25}, Mean = 14

K2 = {4, 12, 20, 11, 31}, Mean = 15.6

3. Re-assign

K1 = {2, 3, 4, 10, 11, 12}, Mean = 7

K2 = {20, 25, 30, 31}, Mean = 26.5

4. Re-assign

K1 = {2, 3, 4, 10, 11, 12}, Mean = 7

Scanned by CamScanner

K2 = {20, 25, 30, 31}, Mean = 26.5

So the final answer is K_1 = {2,3,4,10,11,12}, K_2={20,25,30,31}

Ex. 4.8.4 : Consider four objects with two attribute (X and Y). These four objects are to be grouped together into two clusters. Following are the objects with their attribute values.

Object	X	Y
A	1	1
B	2	1
C	4	3
D	5	4

Soln. :

Each object represents one point with two attributes (X, Y) that can be represented as a coordinate in an attribute space as shown below.

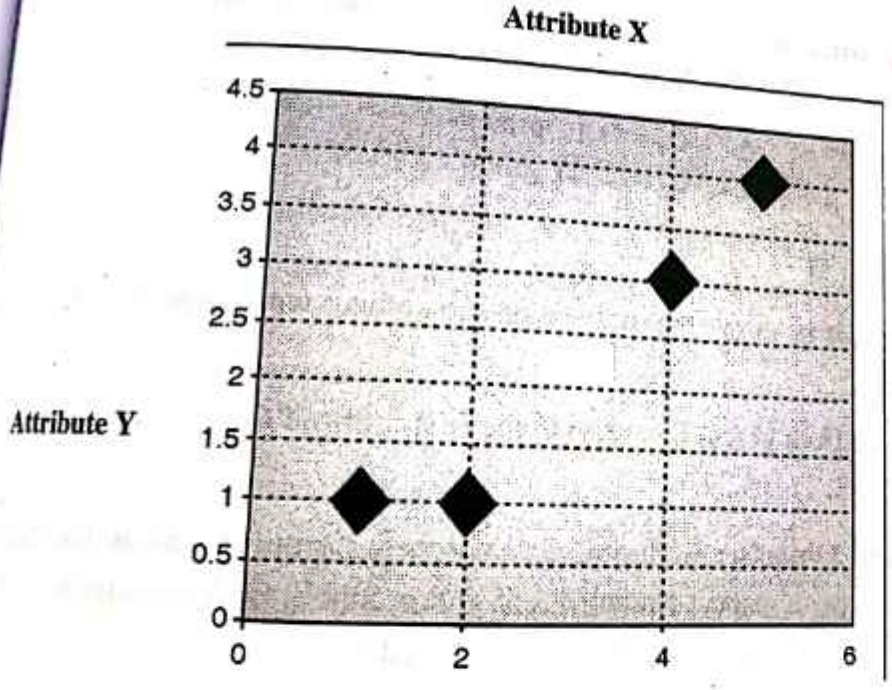


Fig. P. 4.8.4 : Graphical representation of Data Points

1. **Take initial centroids :** Consider object A and object B as the first centroids, then $C_1 = (1,1)$ and $C_2 = (2,1)$

Scanned by CamScanner

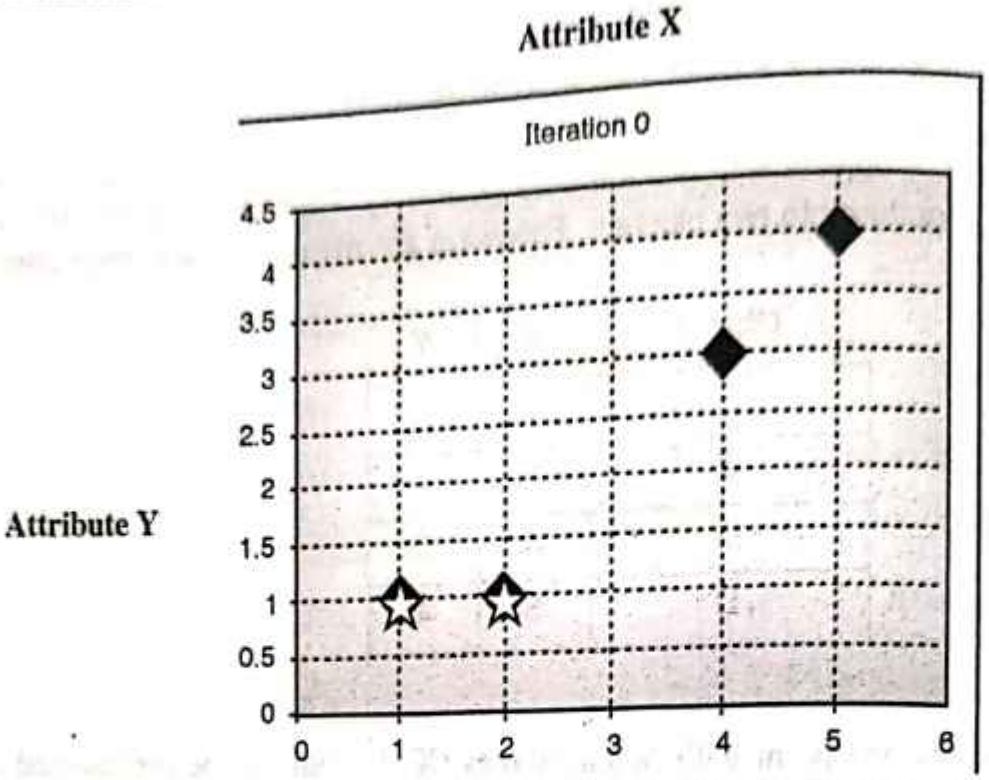


Fig. P. 4.8.4(a) : Randomly selected centroids C_1 and C_2 for two clusters

2. **Objects-centroids distance :** Using Euclidean distance, calculate the distance between cluster centroid to each object. For Iteration 0 we have,

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \begin{matrix} C_1 = (1,1) \text{ group - 1} \\ C_2 = (2,1) \text{ group - 2} \end{matrix}$$

A B C D

The above distance matrix shows the distance of each object with respect to the centroid of cluster C_1 and C_2.

For example, distance from object D = (5,4) to the first centroid $C_1 = (1, 1)$ is 5 and to the second centroid $C_2 = (2, 1)$ is 4.24, etc.

3. **Make the clusters of Objects :** Each object is assigned a group based on the minimum distance of that object with respect to centroid of group. So object A is assigned to group-1 and object B,C,D are assigned grpup-2 and represented by 1.

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} \text{group - 1} \\ \text{group - 2} \end{matrix}$$

A B C D

4. **Determine new centroids :** Based on the object belong to group, calculate the new centroid. As group-1 has only one object, so centroid of group-1 is $C_1 = (1, 1)$.

Scanned by CamScanner

Group-2 has 3 objects so centroid is

$$C_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = \left(\frac{11}{3}, \frac{8}{3}\right)$$

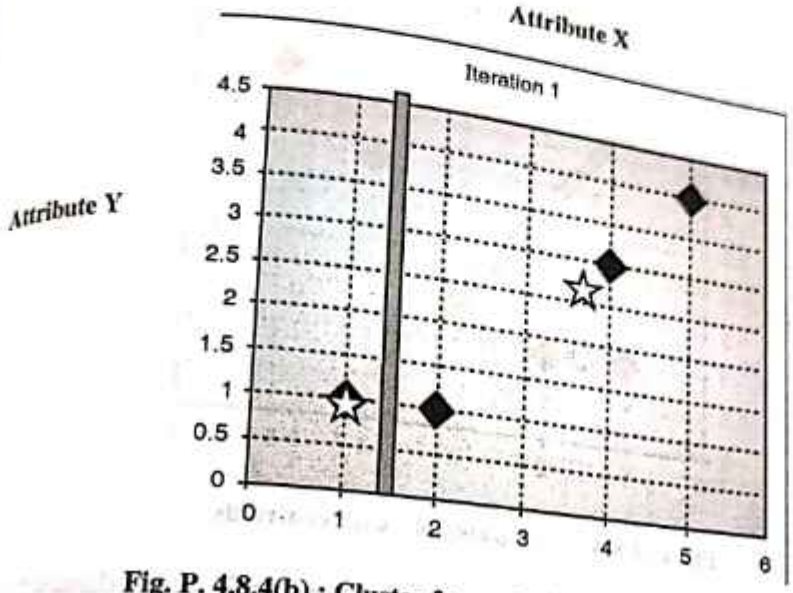


Fig. P. 4.8.4(b) : Cluster formation after First Iteration

5. Calculate **objects-centroids distances :** Compute the distance of all objects with respect to new centroids. So new distance matrix will be D^1 is given below :

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \begin{matrix} C_1 = (1\ 1) & \text{group - 1} \\ C_1 = \left(\frac{11}{3}, \frac{8}{3}\right) & \text{group - 2} \end{matrix}$$

A B C D

6. Make the new clusters of **objects :** Follow the step 3 given above. Based on minimum distance assign the group to each object. Now an object A and B belongs to group-1 and objects C and D belongs to group-2 as given below:

$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} \text{group - 1} \\ \text{group - 2} \end{matrix}$$

A B C D

7. **Again determine centroids :** Repeat the step 4 to calculate new centroid

$$C_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = \left(1\frac{1}{2}, 1\right)$$

Scanned by CamScanner

and $$C_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = \left(4\frac{1}{2}, 3\frac{1}{2}\right)$$

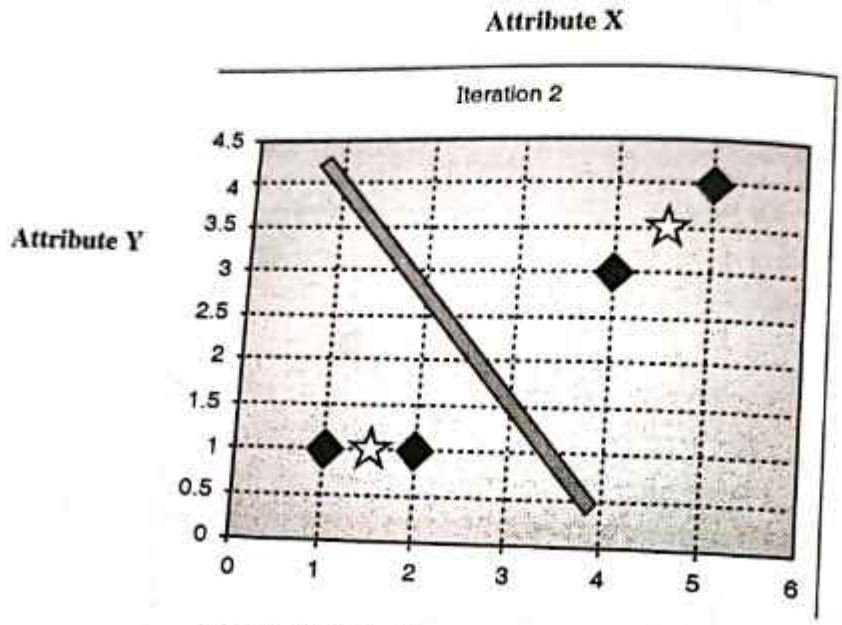


Fig. P. 4.8.4(c) : Two clusters with centroids

8. **Compute the objects-centroids distances :** Repeat step no 2, a new distance matrix for iteration 2 is obtained as shown below :

$$D^2 = \begin{array}{c} \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array} \begin{array}{l} C_1 = \left(1\frac{1}{2}, 1\right) \text{ group - 1} \\ C_2 = \left(4\frac{1}{2}, 3\frac{1}{2}\right) \text{ group - 2} \end{array}$$

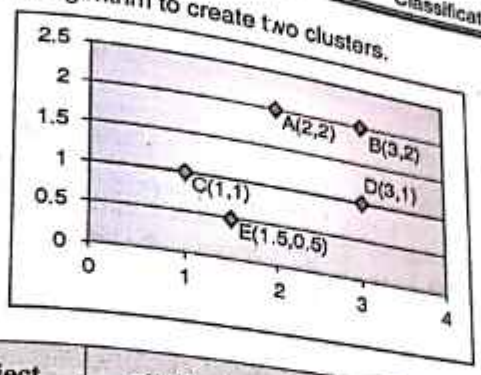
9. **Make the clusters of objects :** Assign each object based on the minimum distance calculated using Euclidean distance.

$$G^2 = \begin{array}{c} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array} \begin{array}{l} \text{group - 1} \\ \text{group - 2} \end{array}$$

Last two iterations shows that object does not move from groups, so stop the iteration of k-means and that will be the final clusters as clustering has reached its stability.

Scanned by CamScanner

Ex. 4.8.5 : Use K-means algorithm to create two clusters.



Soln. :

Object	attribute 1 (X)	attribute 2 (Y)
A	2	2
B	3	2
C	1	1
D	3	1
E	1.5	0.5

For simplicity we can find the adjacency matrix which gives distances of all object from each other.

Using Euclidean Distance we have

$$D(i, j) = \sqrt{|x_2 - x_1|^2 + |y_2 - y_1|^2}$$

$$D(i,j) = D(A,B) = \sqrt{(2-3)^2 + (2-2)^2} = 1,$$

Similarly we can compute for the rest.

	A	B	C	D	E
A	0	1	1.41	1.41	1.58
B	1	0	2.24	1	2.12
C	1.41	2.24	0	2	0.71
D	1.41	1	2	0	1.58
E	1.58	2.12	0.71	1.58	0

1. **Initial value of centroids :** Assume A and C as the first centroids. So the centroids are $C_1(2,2)$ and $C_2(1,1)$

Scanned by CamScanner

2. **Objects-centroids distance :** Using Euclidean distance formula , the distance of each object with respect to centroid C_1 and C_2 is given below :

$D^0 =$

A	B	C	D	E	Cluster centroid
0	1	1.41	1.41	1.58	$C_1(2,2)$ – Group 1
1.41	2.24	0	2	0.71	$C_2(1,1)$ – Group 2

Note : Use adjacency matrix to get the distances or use the Euclidean distance formula for calculation of distances.

The object is represented by column in the distance matrix. The first row represents the distance of each object to the first centroid and the second row to the second centroid.

3. **Objects clustering :** Each object is assigned based on the minimum distance. Thus objects A, B and D is assigned to group 1 and C and E to group 2. A value of 1 is assigned in the distance matrix if an object belongs to that group.

$G^0 =$

A	B	C	D	E	Cluster centroid
1	1	0	1	0	$C_1(2,2)$ – Group 1
0	0	1	0	1	$C_2(1,1)$ – Group 2

4. **Iteration-1, determine centroids :** After assigning the objects to their appropriate groups, now new centroids are calculated. Group 1 has three member thus the centroid C_1 is the average of the coordinates of those three members similarly Group 2 now has two members, thus the centroid is the average coordinate among the two members :

$$C_1 = \left(\frac{2+3+3}{3}, \frac{2+2+1}{3}\right) = (2.67, 1.67)$$

$$C_2 = \left(\frac{1+1.5}{2}, \frac{1+0.5}{2}\right) = (1.25, 0.75)$$

5. **Iteration-1, objects-centroids distances :** The next step is to compute the distance of all objects to the new centroids. Similar to step 2, we have distance matrix at iteration 1 is

$D^1 =$

A	B	C	D	E	Cluster Centroid
0.75	0.47	1.79	0.75	1.65	$C_1(2.67,1.67)$ – Group 1
1.45	2.15	0.32	1.76	0.35	$C_2(1.25,0.75)$ – Group 2

Scanned by CamScanner

For example, distance from A = (2,2) to the first centroid C_1 (2.67,1.67) is $\sqrt{(2.67-2)^2+(1.67-2)^2} = 0.75$, and its distance to the second centroid C_2(1.25,0.75) is $\sqrt{(1.25-2)^2+(0.75-2)^2} = 1.45$, similarly calculate for the points B,C,D,E.

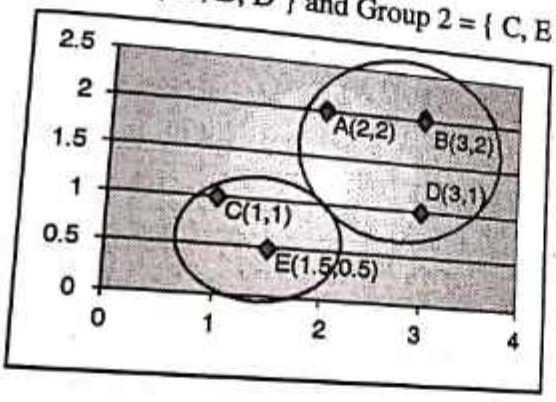
Iteration-1, objects clustering : Similar to step 3, we assign each object based on the minimum distance. The Group matrix is shown below

$G^1 =$

A	B	C	D	E	Cluster centroid
1	1	0	1	0	C_1(2.67,1.67) – Group 1
0	0	1	0	1	C_2 (1.25,0.75) – Group 2

By comparing the above results we observe that $G^0 = G^1$, this shows that object do not move any more to a different group. Thus K means clustering has reached its stability.

So final clusters are Group 1= { A, B, D } and Group 2 = { C, E }



Ex. 4.8.6 : Suppose that the data mining task is to cluster the following points (with (x,y) representing locations) into 3 clusters.

A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9)

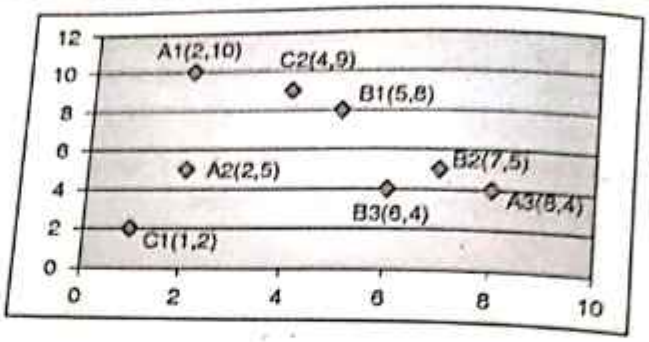
The distance function is Euclidean distance. Suppose initially we assign A1, B1 and C1 as the centre of each cluster respectively, use the K-means algorithm to show only

(a) The three cluster centers after the first round execution.

(b) The final three clusters.

Scanned by CamScanner

Soln. :



1. **Initial value of centroids :** In this we use A1 , B1and C1 as the first centroids. Let X1 and X2, X3 denote the coordinate of the centroids, then X1 = A1(2,10), X2 = B1(5,8) and X3 = C1(1,2)

2. **Objects-centroids distance :** We calculate the distance between cluster centroid to each object. Let us use Euclidean distance, then we have distance matrix at iteration 0 is

$D^0 =$

A1	A2	A3	B1	B2	B3	C1	C2	
0	5	8.48	3.61	7.07	7.21	8.06	2.24	X1=A1(2,10)
3.61	4.24	5	0	3.61	4.12	7.21	1.41	X2=B1(5,8)
8.06	3.16	7.28	7.21	6.71	5.39	0	7.62	X3=C1(1,2)

3. **Objects clustering :** We assign each object based on the minimum distance. Thus, A1 is assigned to group 1, point A3, B1, B2, B3, C2 are assigned to group 2 and A2 and C1 are assigned to group 3. The element of Group matrix below is 1 if and only if the object is assigned to that group.

$G^0 =$

A1	A2	A3	B1	B2	B3	C1	C2	
1	0	0	0	0	0	0	0	X1=A1(2,10)
0	0	1	1	1	1	0	1	X2=B1(5,8)
0	1	0	0	0	0	1	0	X3=C1(1,2)

4. **Iteration-1, determine centroids**

$$X1 = (2,10)$$

$$X2 = \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6,6)$$

Scanned by CamScanner

$$X3 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5,3.5)$$

Iteration-1, objects-centroids distances : The next step is to compute the distance of all objects to the new centroids. Similar to step 2, we have distance matrix at iteration 1 is

$D^1 =$

A1	A2	A3	B1	B2	B3	C1	C2	
0	5	8.48	3.61	7.07	7.21	8.06	2.24	X1(2,10)
5.66	4.12	2.83	2.24	1.41	2	6.40	3.61	X2(6,6)
6.52	1.58	6.52	5.70	5.70	4.52	1.58	6.04	X3(1.5,3.5)

Iteration-1, objects clustering : Similar to step 3, we assign each object based on the minimum distance. The Group matrix is shown below.

$G^1 =$

A1	A2	A3	B1	B2	B3	C1	C2	
1	0	0	0	0	0	0	1	X1(2,10)
0	0	1	1	1	1	0	0	X2(6,6)
0	1	0	0	0	0	1	0	X3(1.5,3.5)

Iteration-2, determine centroids

$$X1 = ((2+4)/2,\ (10+9)/2) = (3, 9.5)$$

$$X2 = ((8+5+7+6)/4, (4+8+5+4)/4) = (6.5, 5.25)$$

$$X3 = ((2+1)/2, (5+2)/2) = (1.5, 3.5)$$

Iteration-2, objects-centroids distances

$D^2 =$

A1	A2	A3	B1	B2	B3	C1	C2	
1.12	2.35	7.43	2.5	6.02	6.26	7.76	1.12	X1(3,9.5)
6.54	4.51	1.95	3.13	0.56	1.35	6.38	7.68	X2(6.5 ,5.25)
6.52	1.58	6.52	5.70	5.70	4.52	1.58	6.04	X3(1.5,3.5)

Iteration-2, objects clustering : We assign each object based on the minimum distance. The Group matrix is shown below.

Scanned by CamScanner

$G^2 =$

A1	A2	A3	B1	B2	B3	C1	C2	
1	0	0	1	0	0	0	1	X1(3,9.5)
0	0	1	0	1	1	0	0	X2(6.5 ,5.25)
0	1	0	0	0	0	1	0	X3(1.5,3.5)

10. Iteration-3, determine centroids

$$X1 = ((2+5+4)/3,\ (10+9+8)/3) = (3.67, 9)$$

$$X2 = ((8+7+6)/3, (4+5+4)/3) = (7, 4.33)$$

$$X3 = ((2+1)/2, (5+2)/2) = (1.5, 3.5)$$

11. Iteration-3, objects-centroids distances

$D^3 =$

A1	A2	A3	B1	B2	B3	C1	C2	
1.95	4.33	6.61	1.66	5.20	5.52	7.49	0.33	X1(3.67,9)
6.01	5.04	1.05	4.17	0.67	1.05	6.44	5.55	X2(7 ,4.33)
6.52	1.58	6.52	5.70	5.70	4.52	1.58	6.04	X3(1.5,3.5)

12. Iteration-3, objects clustering : We assign each object based on the minimum distance. The Group matrix is shown below.

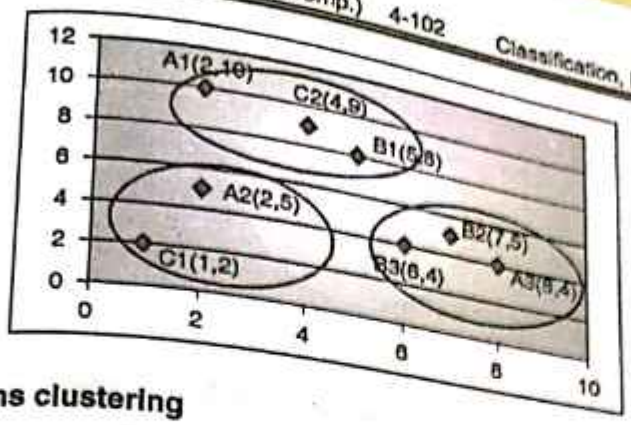
$G^3 =$

A1	A2	A3	B1	B2	B3	C1	C2	
1	0	0	1	0	0	0	1	X1(3.67,9)
0	0	1	0	1	1	0	0	X2(7 ,4.33)
0	1	0	0	0	0	1	0	X3(1.5,3.5)

By comparing $G^3 = G^2$ we see that the objects do not move to new group therefore we can say that K means has reached its stability.

So final clusters are Group 1= { A1, B1, C2 } and Group 2 = { A3, B2 , B3} and Group 3 = {A2, C1}.

Scanned by CamScanner



Strength of K-means clustering

- **Relatively efficient** : O(ikn),
- where n is number of objects,
- k is number of clusters, i is number of iterations.
- Normally, k, i << n.
- K-means often terminates at a **local optimum.**
- Techniques like **deterministic annealing** and **genetic algorithms are used to get the global optimum solution.**

Weakness of K-means clustering

- Applicable only when **mean** is defined.
- Need to specify k, the number of clusters, in advance.
- Unable to handle noisy data and **outliers** (outlier : objects with extremely large values)
- Not suitable to discover clusters with **non-convex shapes.**

Ex. 4.8.7 : Explain K-means clustering algorithm ? Apply K-means algorithms for the following data set with two clusters. Data set = {1,2, 6, 7, 8 8 10, 15, 17,20}.

MU - May 2016, 10 Marks

Soln.:

1. Given : {1,2,6,7,8,10 ,15,17,20}

   Consider initial two centroids for two clusters C1= 6 and C2 = 15

2. Number of cluster = 2, therefore

$$K1 = \{1,2,6,7,8,10\}, C1 = \text{Mean} = 5.67$$

Scanned by CamScanner

$$K2 = \{15,17,20\},\ C1 = \text{Mean} = 17.33$$

3. Re-assign

$$K1 = \{1,2,6,7,8,10\}$$

$$K2 = \{15,17,20\}$$

As no element is moving from cluster , So the final answer is

$$K1 = \{1,2,6,7,8,10\} \quad \text{and} \quad K2 = \{15,17,20\}$$

4.8.2 K-Medoids (Representative Object-based Technique)

Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.

- Also called as Partitioning Around Medoids (PAM).
- Handles outliers well.
- Ordering of input does not impact results.
- Does not scale well.
- Each cluster represented by one item, called the medoid.
- Initial set of K medoids randomly chosen.

In a single partition of data into K clusters where each cluster has a representative point that is centrally located point of the cluster based on some distance measure.

These representative points are called medoids.

Basic K-medoid algorithm

1. Select K points as the initial medoids.
2. Assign all points to the closest medoid.
3. See if any other point is a "better" medoid.
4. Repeat steps 2 and 3 until the medoids don't change.

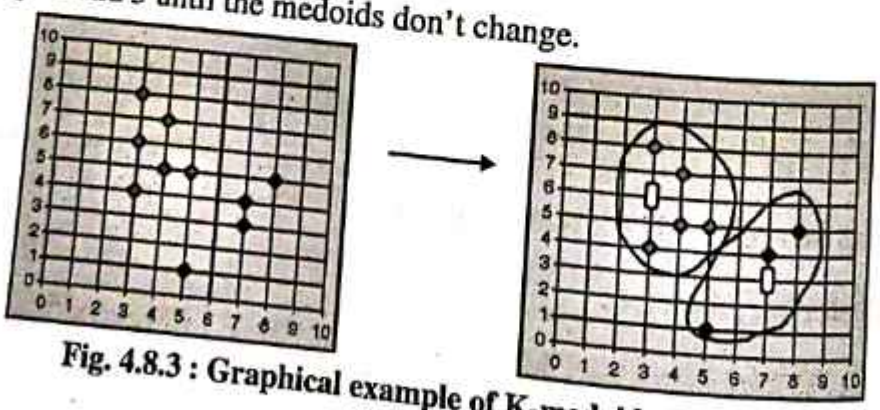


Fig. 4.8.3 : Graphical example of K-medoids clustering

Scanned by CamScanner

- At each step in algorithm, medoids are changed if the overall cost is improved.
- C_{jh} – cost change for an item t_j associated with swapping medoid t_i with non-medoid t_h.

Algorithm given by Margaret H. Dunham

```
Input :
D = {t1, t2,...,tn} // set of elements
A   // Adjacency matrix showing distance between elements.
k   // Number of desired clusters.

Output :
K   // set of clusters.

PAM Algorithm :
Arbitrarily select k medoids from D;

Repeat
  For each th not a medoid do
      For each medoid t, do
          Calculate TCih;
find i,h where TCih is the smallest;
if TCih < 0 then
      replace medoid ti with th ;
until TCih ≤ 0;
  for each ti ∈ D do
      assign ti to Kj where dis(ti, tj) is the smallest over all medoids;
```

Calculation of swapping cost

$$TC_{ih} = \sum_{j=1}^{n} C_{jih}$$

Advantages of PAM (Partitioning Around Medoids)

- PAM works effectively for small data sets, but does not handle large data sets well.
- Complexity is $O(n(n-k)^2)$ for each iteration where n is number of data, k is number of clusters.
- PAM is more robust than k-means in the presence of noise and outliers.

Scanned by CamScanner

Ex. 4.8.8 : Coordinates of objects are given below. Apply K-medoids (PAM). Number of clusters = 2.

Table P. 4.8.8

Number	x co-ordinate	y co-ordinate
1	1.0	4.0
2	5.0	1.0
3	5.0	2.0
4	5.0	4.0
5	10.0	4.0
6	25.0	4.0
7	25.0	6.0
8	25.0	7.0
9	25.0	8.0
10	29.0	7.0

Soln. :

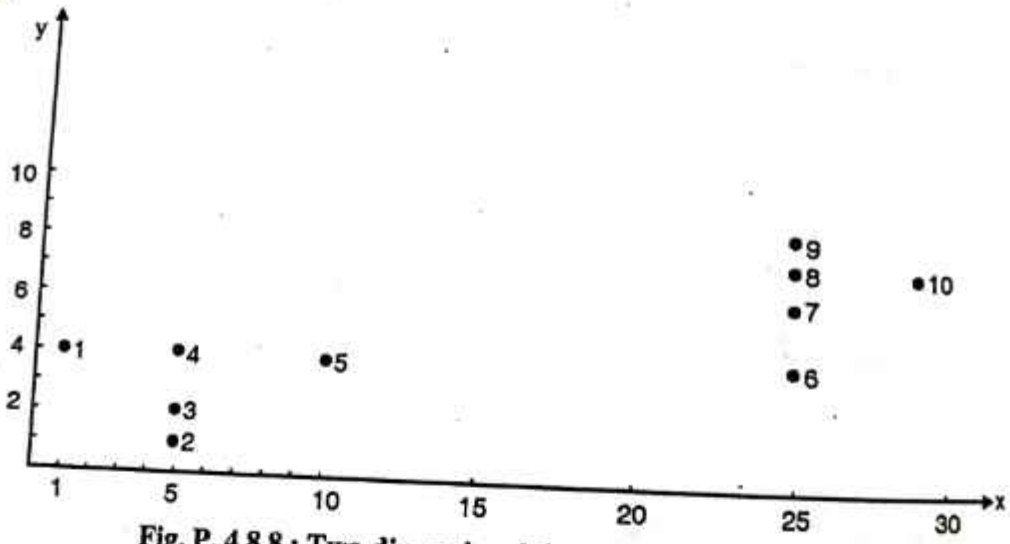


Fig. P. 4.8.8 : Two-dimensional example with 10 objects

Step 1 :

- Objects 1 and 5 are the selected representative objects initially. (Random selection)
- Calculate the distance of every object with respect to selected object 1 and 5.
- Find the closest representative object with respect to the selected object.
- Calculate the average value of minimal dissimilarity.
- Cost = Average value of minimal dissimilarity = 9.37.

Scanned by CamScanner

Table P. 4.8.8(a) : Assignment of objects to two representative objects

Object number	Dissimilarity from object 1	Dissimilarity from object 5	Minimal dissimilarity	Closest representative object
1	0.00	9.00	0.00	1
2	5.00	5.83	5.00	1
3	4.47	5.39	4.47	1
4	4.00	5.00	4.00	1
5	9.00	0.00	0.00	5
6	24.00	15.00	15.00	5
7	24.08	15.13	15.13	5
8	24.19	15.30	15.30	5
9	24.33	15.52	15.52	5
10	28.16	19.24	19.24	5
			Average 9.37	

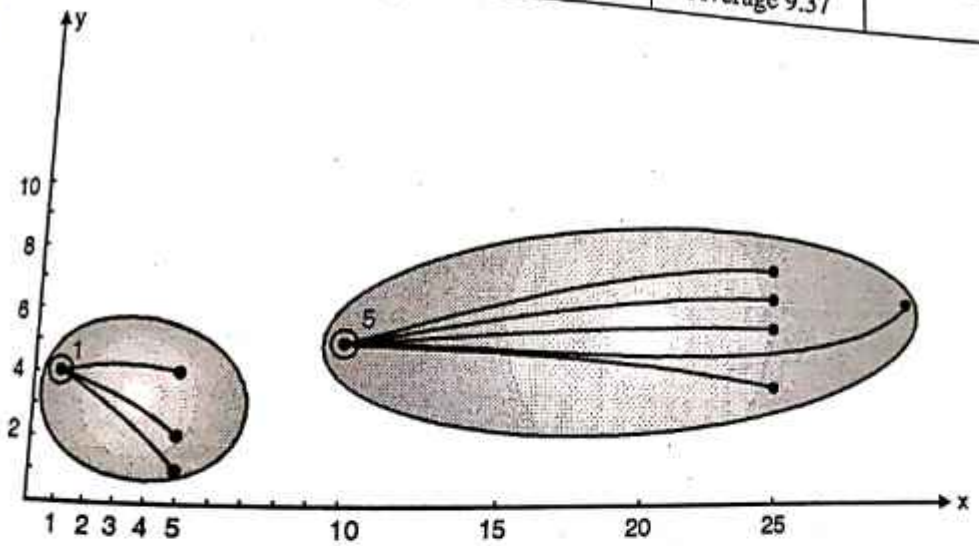


Fig. P. 4.8.8(a)

Step 2 :

- Select two objects randomly.
- Objects 4 and 8 are the selected representative objects.

Scanned by CamScanner

- Repeat the step I.
- Cost = average value of minimal dissimilarity = 2.30.

Table P. 4.8.8(b) : Assignment of objects to two representative objects

Object number	Dissimilarity from object 4	Dissimilarity from object 8	Minimal dissimilarity	Closest representative object
1	4.00	24.19	4.00	4
2	3.00	20.88	3.00	4
3	2.00	20.62	2.00	4
4	0.00	20.22	0.00	4
5	5.00	15.30	5.00	4
6	20.00	3.00	3.00	8
7	20.10	1.00	1.00	8
8	20.22	0.00	0.00	8
9	20.40	1.00	1.00	8
10	24.19	4.00	4.00	8
			Average 2.30	

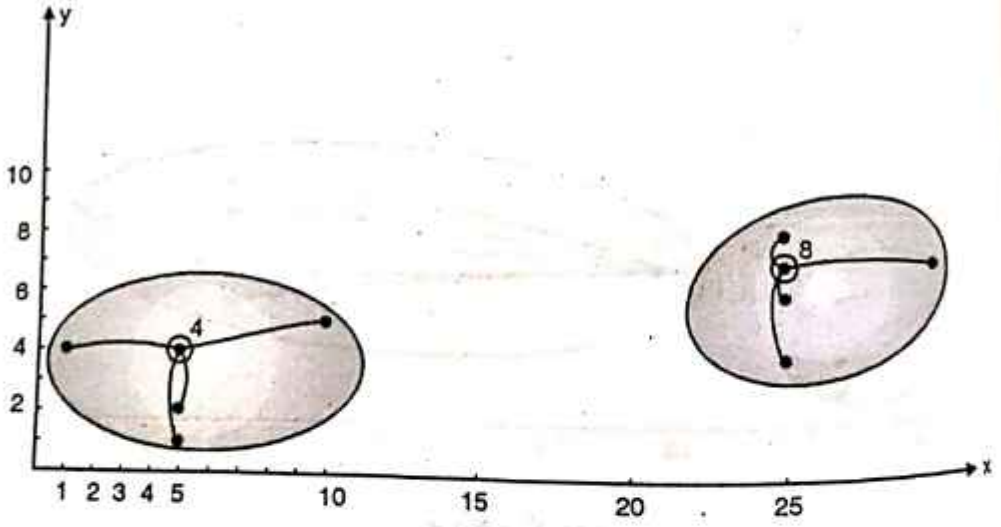


Fig. P. 4.8.8(b) : Clustering corresponding to the selections described in Table P. 4.8.7(a) and (b)

Step 3 :

- Calculation of swapping cost.
- Swapping cost = New cost – Old cost = 2.30 – 9.32 = – 7.02

Scanned by CamScanner

If swapping cost < 0

replace medoid with new selected object.

So new medoids are object 4 and object 8.

Step 4 : Repeat step 2 and 3 until swapping cost <= 0

1&3 Sampling Based Method

1. **CLARA**

- Clustering large Applications.
- Built in statistical analysis packages, such as S+.
- Draw multiple samples of the data set, apply PAM on each sample.
- Performs better than PAM in larger data sets.
- Efficiency depends on the sample size.

2. **CLARANS**

- CLARANS is applicable to large applications which are based upon randomized search.
- CLARANS is more efficient than PAM and CLARA clustering algorithms.
- Search is over the sample of the neighbours of a node.
- It draws a sample of neighbours in each search step.
- It has two main parameters for clustering they are maximum number of neighbours and number of local minima obtained.

Syllabus Topic : Hierarchical Methods - Agglomerative, Divisive

4.9 Hierarchical Clustering

→ **(MU - May 2014, Dec. 2014)**

Various hierarchical clustering algorithms are :

- Single-linkage clustering, nearest-neighbour.
- Complete-linkage, furthest neighbour.
- Average-linkage, Unweighted Pair-Group Method Average (UPGMA).

Scanned by CamScanner

- Weighted-pair group average, UPGMA weighted by cluster sizes.
- Within-groups clustering.
- Ward's method.

Hierarchical clustering technique (Basic algorithm)

1. Compute the proximity matrix (i.e. distance matrix).
2. Let each data point be a cluster.
3. Repeat.
4. Merge the two closest clusters.
5. Update the proximity matrix.
6. Until only a single cluster remains.

Note : Proximity matrix means the matrix which is symmetric, meaning that the numbers on the lower half of the diagonal will be the same as the numbers on the top half of the diagonal.

Different approaches to defining the distance between clusters distinguish the different algorithms i.e.

Single-linkage clustering :

Single Linkage clustering is also called as minimum method, the minimum distance from any object of one cluster to any object of another cluster is considered. In the single linkage method, **D(A,B)** is computed as D(A,B) = Min { d(i,j) : Where object i is in cluster A and object j is in cluster B}

This measure of inter-group distance is illustrated in the Fig. 4.9.1.

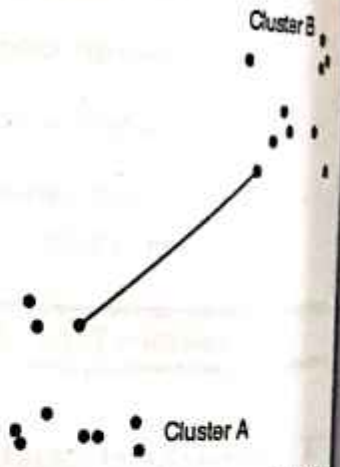


Fig. 4.9.1 : Single-linkage clustering

Complete-linkage clustering

Complete linkage also called as maximum method, the maximum distance between any object of one cluster to any object of another cluster is considered.

In the complete linkage method, **D(A,B)** is computed as D(A,B) = Max { d(i,j) : Where object i is in cluster A and object j is in cluster B}

Scanned by CamScanner

The measure is illustrated in the Fig. 4.9.2.

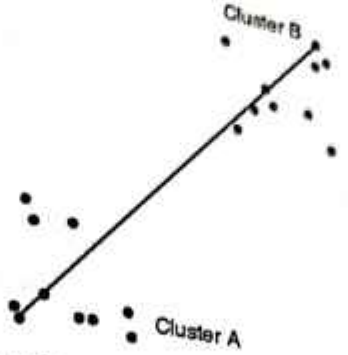


Fig. 4.9.2 : Complete-linkage clustering

Average-linkage clustering

In average-linkage clustering, we consider the distance between any two clusters A and B is taken to be the average of all distances between pairs of objects "i" in A and "j" in B, that is, the mean distance between elements of each cluster.

In the average linkage method, **D(A,B)** is computed as D(A,B) = mean{ d(i,j) : Where object i is in cluster A and object j is in cluster B }

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{p_i \in \text{Cluster}_i} \sum_{p_j \in \text{Cluster}_j} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

Fig. 4.9.3 : Average-linkage clustering

The Fig. 4.9.3 illustrates average linkage clustering.

4.9.1 Agglomerative Hierarchical Clustering

→ **(MU - May 2010)**

In Hierarchical clustering algorithms, either top down or bottom up approach is followed. In Bottom up approach, every object is considered to be a cluster and in subsequent iterations they are merged in to single cluster. Therefore it is also called as Hierarchical Agglomerative Clustering (HAC).

Scanned by CamScanner

An HAC clustering is typically visualized as a *dendrogram* as shown in Fig. 4.9.4 where each merge is represented by a horizontal line.

What is Dendrogram ?

- **Dendrogram :** A tree data structure which illustrates hierarchical clustering techniques.
- Each level shows clusters for that level.
  - Leaf – individual clusters
  - Root – one cluster
- A cluster at level i is the union of its children clusters at level i+1.
- A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

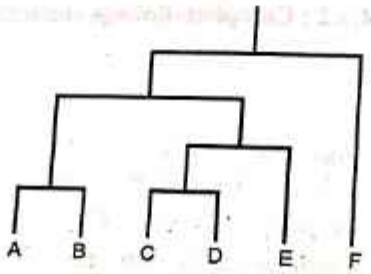


Fig. 4.9.4 : Dendrogram

- The flow chart of agglomerative hierarchical clustering algorithm is shown in Fig. 4.9.5.

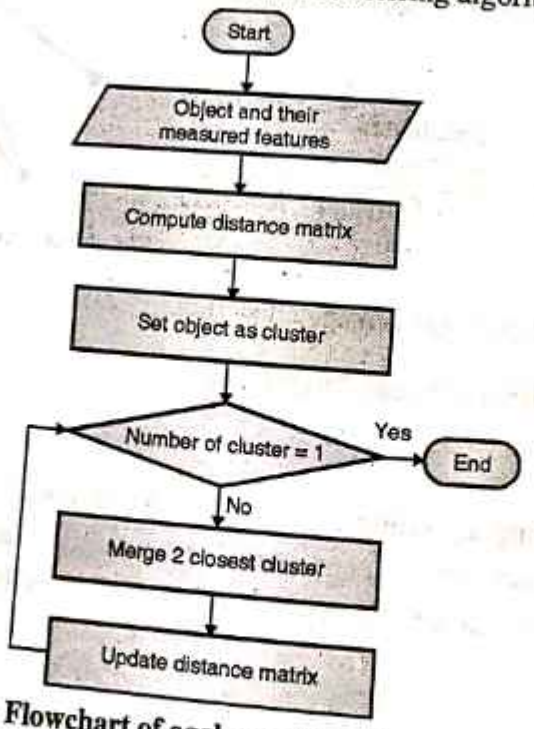


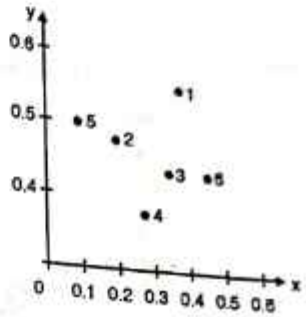
Fig. 4.9.5 : Flowchart of agglomerative hierarchical clustering

Ex. 4.9.1 : Assume that the database D is given by the table below. Follow single link technique to find clusters in D. Use Euclidean distance measure.

		X	Y
	p1	0.40	0.53
	p2	0.22	0.38
D =	p3	0.35	0.32
	p4	0.26	0.19
	p5	0.08	0.41
	p6	0.45	0.30

Soln. :

Step 1 : Plot the objects in n-dimensional space (where n is the number of attributes). In our case we have 2 attributes x and y, so we plot the objects p1, p2, ... p6 in 2-dimensional space :



Step 2 : Calculate the distance from each object (point) to all other points, using Euclidean distance measure, and place the numbers in a distance matrix.

The formula for Euclidean distance between two points i and j is :

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j1}|^2 + \ldots + |x_{ip} - x_{jp}|^2}$$

where x_{i1} is the value of attribute 1 for i and x_{j1} is the value of attribute 1 for j and so on, as many attributes we have ... shown up to p i.e. x_{ip} in the formula.

In our case, we only have 2 attributes. So, the Euclidean distance between our points p1 and p2, which have attributes x and y would be calculated as follows:

$$\begin{aligned} d(p1, p2) &= \sqrt{|xp_1 - xp_1|^2 + |yp_1 - yp_2|^2} \\ &= \sqrt{|0.40 - 0.22|^2 + |0.53 - 0.38|^2} \\ &= \sqrt{|0.18|^2 + |0.15|^2} \\ &= \sqrt{0.0324 + 0.0225} \\ &= \sqrt{0.0549} = 0.2343 \end{aligned}$$

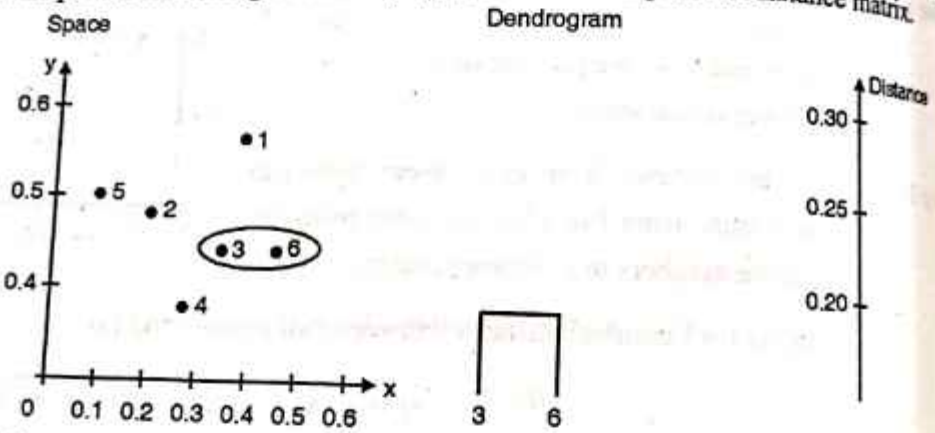
Analogically, we calculate the distance to the remaining points and we will receive the following values :

Scanned by CamScanner

Distance matrix

p1	0					
p2	0.24	0				
p3	0.22	0.15	0			
p4	0.37	0.20	0.15	0		
p5	0.34	0.14	0.28	0.29	0	
p6	0.23	0.25	**0.11**	0.22	0.39	0
	p1	p2	p3	p4	p5	p6

Step 3 : In the above matrix, p6 and p3 are two clusters with shortest distance 0.11, so merge p6 and p3 and make single cluster (p3,p6). Now re-compute the distance matrix.



Distance matrix

p1	0				
p2	0.24	0			
(p3, p6)	**0.22**	**0.15**	0		
p4	0.37	0.20	**0.15**	0	
p5	0.34	0.14	**0.28**	0.29	0
	p1	p2	**(p3, p6)**	p4	p5

To calculate the distance of p1 from (p3,p6) :

$$\text{dist}((p3, p6), p1) = \text{MIN}(\text{dist}(p3, p1), \text{dist}(p6, p1))$$

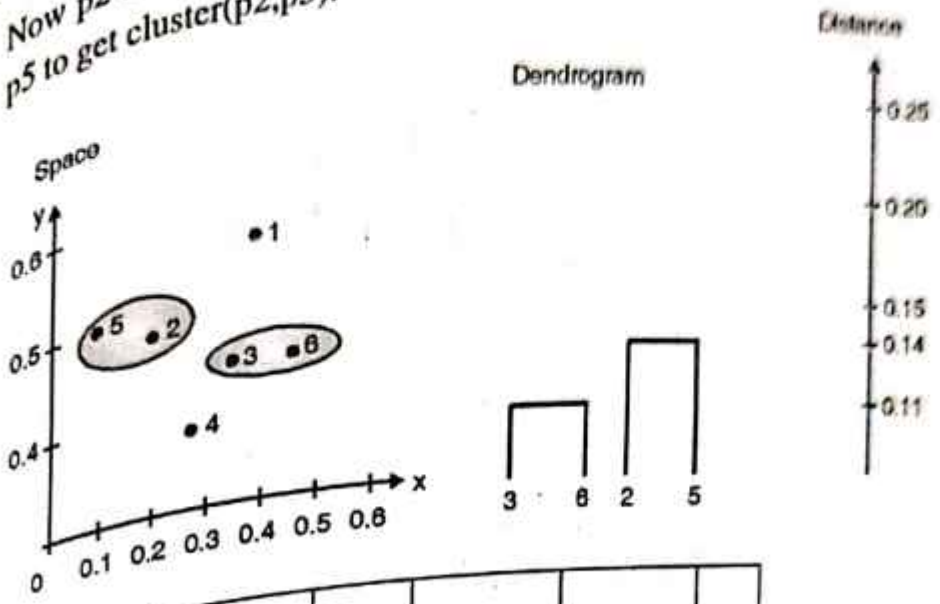
$$= \text{MIN}(0.22, 0.23) \quad \text{//from original matrix}$$

$$= 0.22$$

Scanned by CamScanner

Step 4 : Repeat Step 3 until one single cluster is formed i.e. merge all the clusters.

a. Now p2 and p5 have the smallest distance from above matrix, so merge p2 and p5 to get cluster(p2,p5).



p1	0			
(p2, p5)	**0.24**	0		
(p3, p6)	0.22	**0.15**	0	
p4	0.37	**0.20**	0.15	0
	p1	**(p2, p5)**	(p3, p6)	p4

The distance between (p3, p6) and (p2, p5) is calculated as given below :

dist((p3, p6), (p2, p5)) = MIN (dist(p3, p2) , dist(p6, p2), dist(p3, p5), dist(p6, p5))

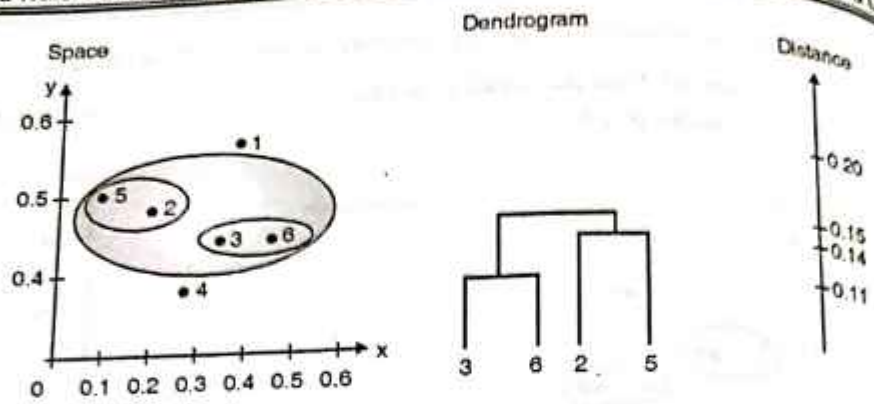
= MIN (0.15, 0.25, 0.28, 0.39) //from original matrix

= 0.15

b. Repeat Step 3.

Merge (p2, p5) and (p3, p6) as having minimum distance i.e.0.15 and again compute distance matrix.

Scanned by CamScanner



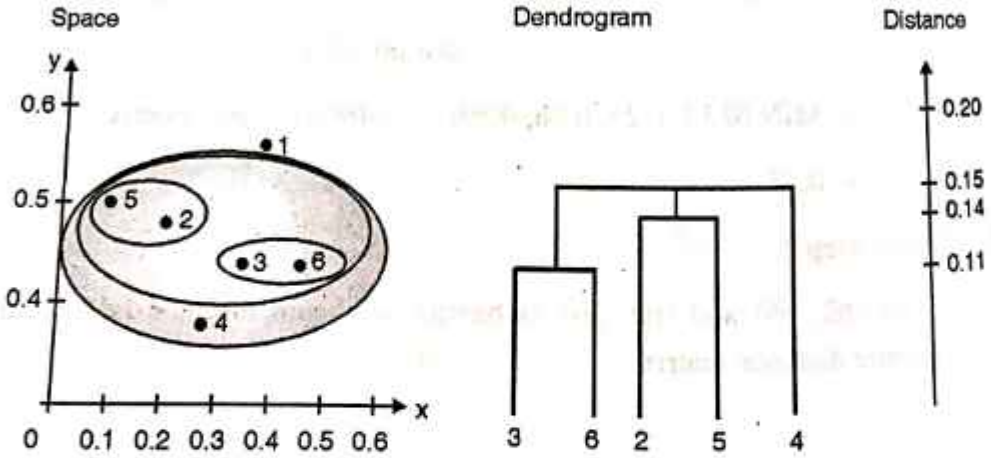
Distance matrix

p1	0		
(p2, p5, p3, p6)	**0.22**	0	
p4	0.37	**0.15**	0
	p1	**(p2, p5, p3, p6)**	p4

c. Since we have more clusters to merge, we continue to repeat Step 3.

So, looking at the last distance matrix above, we see that (p2, p5, p3, p6) and p4 have the smallest distance from all i.e. 0.15. So, we merge those two in a single cluster, and re-compute the distance matrix.

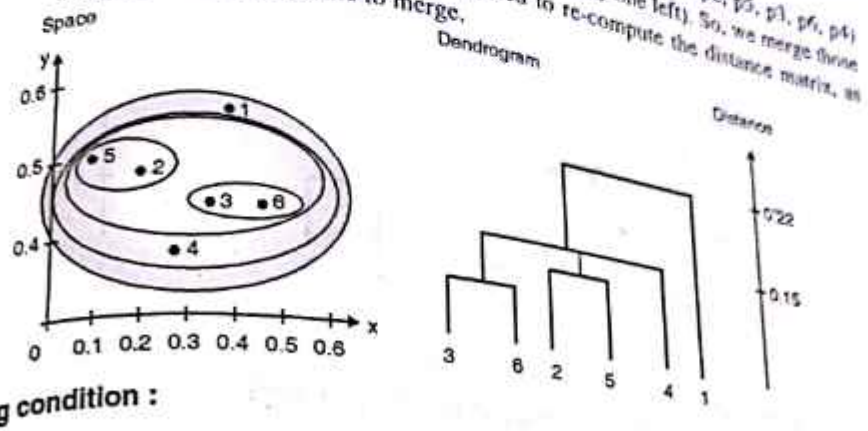
Space dendrogram



Scanned by CamScanner

d. Since we have more clusters to merge, we continue to repeat Step 3.

So, looking at the last distance matrix above, we see that (p2, p5, p1, p6, p4) and p1 have the smallest distance - 0.22 (the only one left). So, we merge those two in a single cluster. There is no need to re-compute the distance matrix, as there are no more clusters to merge.



Stopping condition :

We indicated that "each object is placed in a separate cluster, and at each step we merge the closest pair of clusters, *until certain termination conditions are satisfied*".

To stop clustering either user has to specify the number of clusters he wants or algorithm has to make decision to stop clustering at which level. Through dendrogram, we can notice the merging of clusters at various distances. If merging of clusters is at high distance then we can stop clustering at that level.

Complete link :

Step 1 and step 2 :

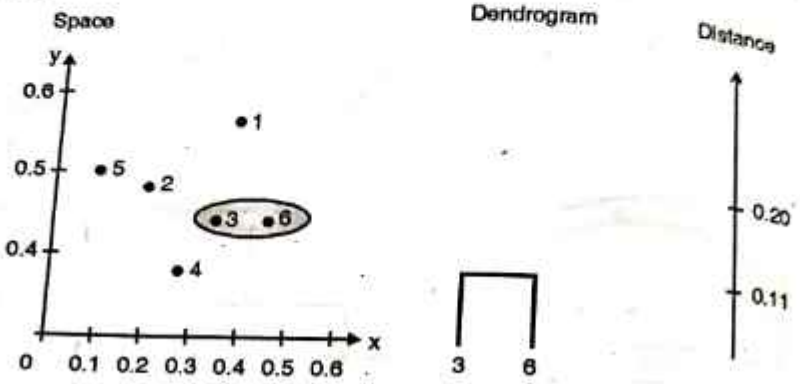
Refer the single link solution (same)

Distance matrix :

	p1	p2	p3	p4	p5	p6
p1	0					
p2	0.24	0				
p3	0.22	0.15	0			
p4	0.37	0.20	0.15	0		
p5	0.34	0.14	0.28	0.29	0	
p6	0.23	0.25	0.11	0.22	0.39	0

Scanned by CamScanner

Step 3 : Identify the two clusters with the shortest distance in the matrix, and merge them together. Re-compute the distance matrix, as those two clusters are now in a single cluster.



By looking at the distance matrix above, we see that p3 and p6 have the smallest distance from all i.e. 0.11 So, we merge those two in a single cluster and re-compute the distance matrix.

$$dist((p3, p6), p1) = MAX(dist(p3, p1), dist(p6, p1))$$
$$= MAX(0.22, 0.23) \quad \text{//from original matrix}$$
$$= 0.23$$

$$dist((p3, p6), p2) = MAX(dist(p3, p2), dist(p6, p2))$$
$$= MAX(0.15, 0.25) \quad \text{//from original matrix}$$
$$= 0.25$$

$$dist((p3, p6), p4) = MAX(dist(p3, p4), dist(p6, p4))$$
$$= MAX(0.15, 0.22) \quad \text{//from original matrix}$$
$$= 0.22$$

$$dist((p3, p6), p5) = MAX(dist(p3, p5), dist(p6, p5))$$
$$= MAX(0.28, 0.39) \quad \text{//from original matrix}$$
$$= 0.39$$

Scanned by CamScanner

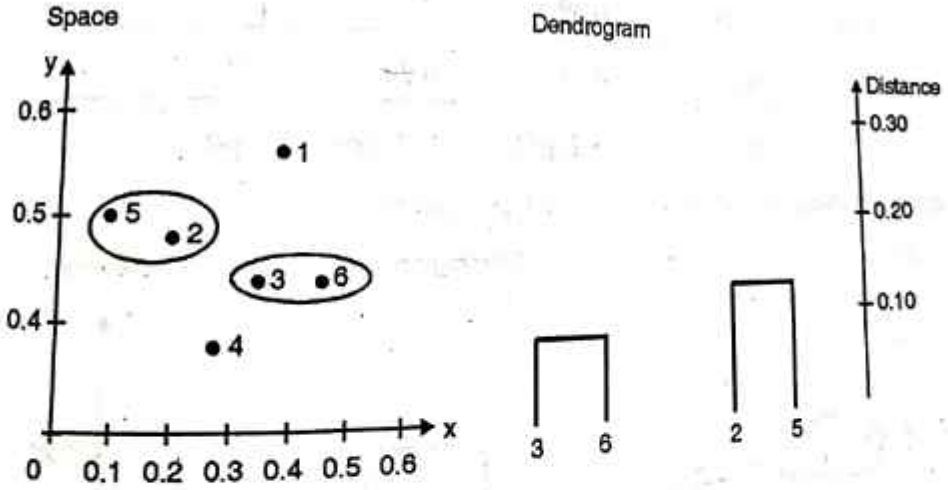
New distance matrix

	p1	p2	(p3, p6)	p4	p5
p1	0				
p2	0.24	0			
(p3, p6)	**0.23**	**0.25**	0		
p4	0.37	0.20	**0.22**	0	
p5	0.34	0.14	**0.39**	0.29	0

Step 4 : Consider the following distance matrix

	p1	p2	(p3, p6)	p4	p5
p1	0				
p2	0.24	0			
(p3, p6)	**0.23**	**0.25**	0		
p4	0.37	0.20	**0.22**	0	
p5	0.34	0.14	**0.39**	0.29	0

So, looking at the above distance matrix, we see that p2 and p5 have the smallest distance from all – 0.14. So, we merge those two in a single cluster, and re-compute the distance matrix using the following calculations.



$$dist((p2, p5), p1) = MAX(dist(p2, p1), dist(p5, p1))$$

$$= MAX(0.24, 0.34) \quad \text{//from original matrix}$$

$$= 0.34$$

Scanned by CamScanner

dist((p2, p5), (p3,p6)) = MAX (*dist*(p2, p3) , *dist*(p2, p6), *dist*(p5, p3), *dist*(p5, p6))

= MAX (0.15, 0.25, 0.28, 0.39) //from original matrix

= 0.39

dist((p2, p5), p4) = MAX (*dist*(p2, p4) , *dist*(p5, p4))

= MAX (0.20, 0.29) //from original matrix

= 0.29

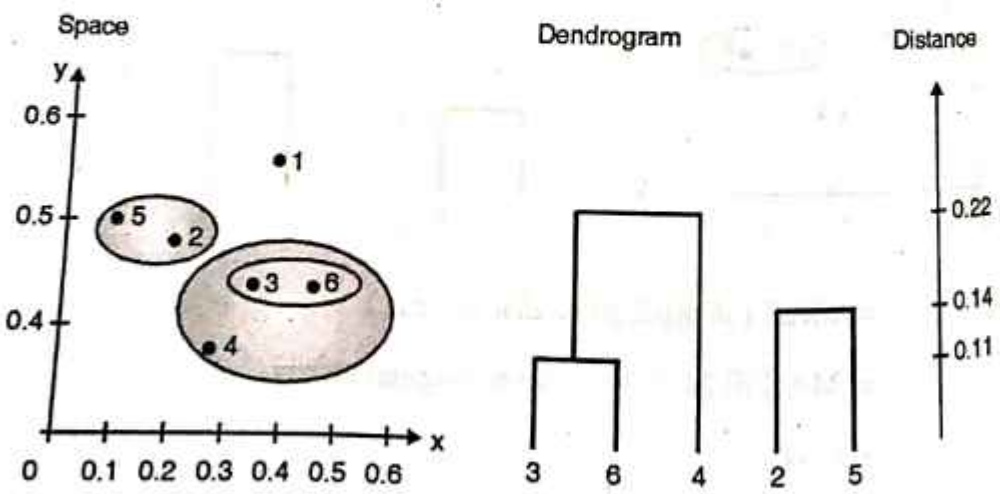
Therefore new distance matrix is :

p1	0			
(p2, p5)	**0.34**	0		
(p3, p6)	0.23	**0.39**	0	
p4	0.37	**0.29**	0.22	0
	p1	**(p2, p5)**	(p3, p6)	p4

Step 5 : Consider the following matrix

p1	0			
(p2, p5)	**0.34**	0		
(p3, p6)	0.23	**0.39**	0	
p4	0.37	**0.29**	0.22	0
	p1	**(p2, p5)**	(p3, p6)	p4

The minimum distance is 0.22, so merge (p3, p6) and p4



Scanned by CamScanner

$$dist((p3, p6, p4), p1) = MAX(dist(p3, p1), dist(p6, p1), dist(p4, p1))$$
$$= MAX(0.22, 0.23, 0.37) \text{ //from original matrix}$$
$$= 0.37$$

$$dist((p3, p6, p4), (p2, p5)) = MAX(dist(p3, p2), dist(p3, p5), dist(p6, p2), dist(p6, p5), dist(p4, p2), dist(p4, p5))$$
$$= MAX(0.15, 0.28, 0.25, 0.39, 0.20, 0.29)$$
$$= 0.39$$

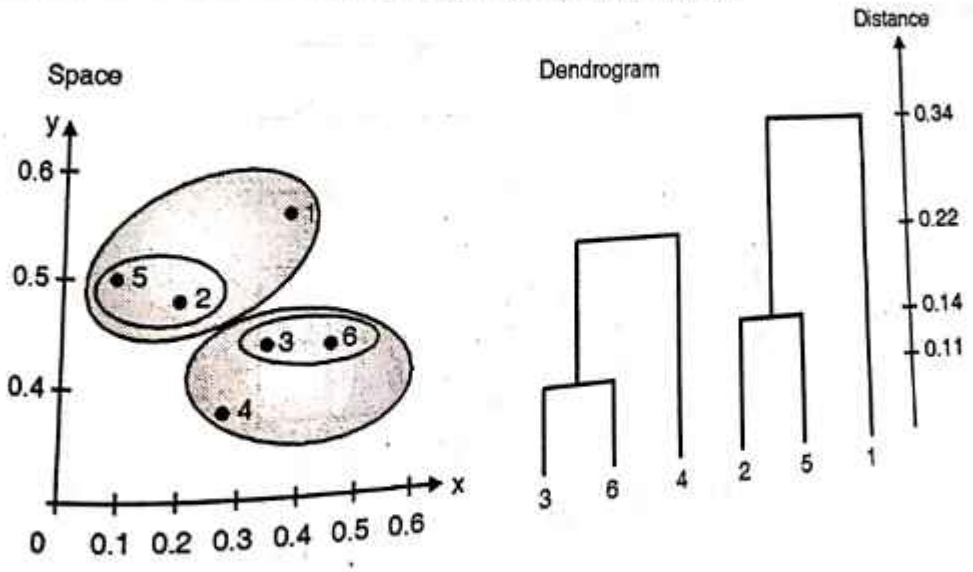
Therefore new distance matrix

p1	0		
(p2, p5)	**0.34**	0	
(p3, p6, p4)	0.37	**0.39**	0
	p1	**(p2, p5)**	(p3, p6, P4)

Step 6 : Now consider the following distance matrix

p1	0		
(p2, p5)	**0.34**	0	
(p3, p6, p4)	0.37	**0.39**	0
	p1	**(p2, p5)**	(p3, p6, P4)

Since the minimum distance is 0.34, merge (p2, p5) with p1.



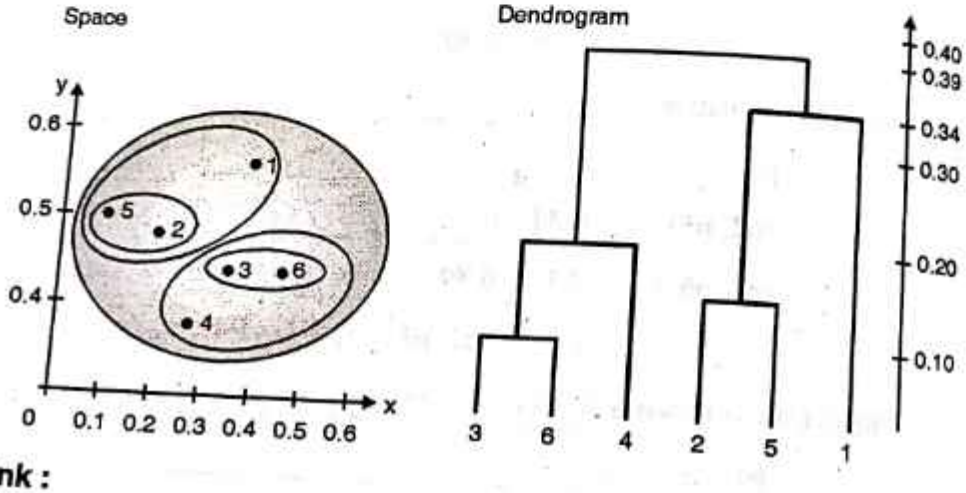
Scanned by CamScanner

dist ((p2, p5, p1), (p3, p6, p4) = 0.39

Therefore new distance matrix

	(p2, p5,p1)	(p3, p6, P4)
(p2, p5,p1)	0	
(p3, p6, p4)	0.39	0

Finally, merge the cluster (p2, p5,p1) and (p3,p6,p4)



Average link :

Step 1 and Step 2 :

Refer the single link solution

Distance matrix :

	p1	p2	p3	p4	p5	p6
p1	0					
p2	0.24	0				
p3	0.22	0.15	0			
p4	0.37	0.20	0.15	0		
p5	0.34	0.14	0.28	0.29	0	
p6	0.23	0.25	0.11	0.22	0.39	0

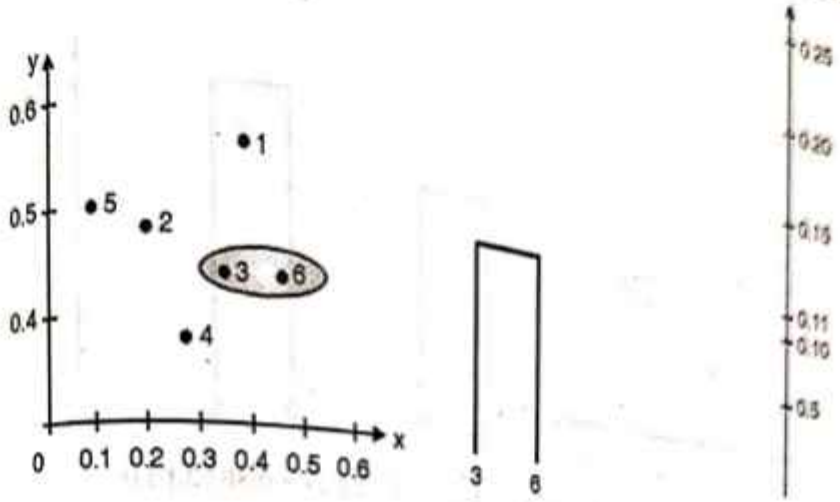
Scanned by CamScanner

Step 3: Identify the two clusters with the shortest distance in the matrix, and merge them together. Re-compute the distance matrix, as those two clusters are now in a single cluster.

By looking at the distance matrix above, we see that p3 and p6 have the smallest distance from all – 0.11 So, we merge those two in a single cluster, and re-compute the distance matrix.

Space

Dendrogram



dist((p3 p6), p1) = 1/2 (dist(p3,p1) + dist(p6,p1)) = 0.5 × (0.22 + 0.23) = 0.23

dist((p3 p6), p2) = 1/2 (dist(p3,p2) + dist(p6,p2)) = 0.5 × (0.15 + 0.25) = 0.2

dist((p3 p6), p4) = 1/2 (dist(p3,p4) + dist(p6,p4)) = 0.5 × (0.15 + 0.22) = 0.19

dist((p3 p6), p5) = 1/2 (dist(p3,p5) + dist(p6,p5)) = 0.5 × (0.28 + 0.39) = 0.34

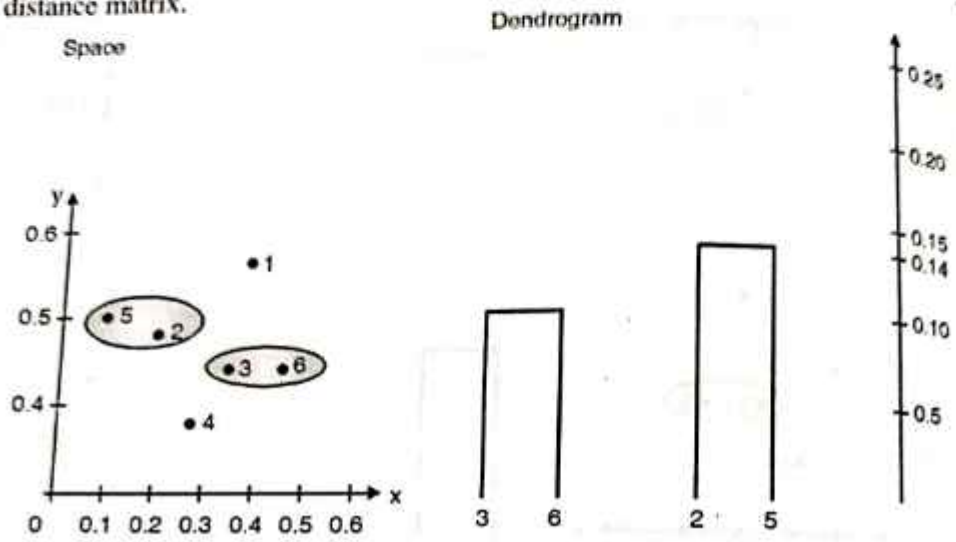
Distance matrix :

p1	0				
p2	0.24	0			
(p3, p6)	**0.23**	**0.2**	0		
p4	0.37	0.20	**0.19**	0	
p5	0.34	0.14	**0.34**	0.29	0
	p1	p2	**(p3, p6)**	p4	p5

Scanned by CamScanner

Step 4 :

So, looking at the above distance matrix above, we see that p2 and p5 have the smallest distance from all – 0.14. So, we merge those two in a single cluster, and re-compute the distance matrix.



$$\text{dist}((p2, p5), p1) = 1/2\,(\text{dist}(p2,p1) + \text{dist}(p5,p1))$$

$$= 0.5 \times (0.24 + 0.34) = 0.29$$

$$\text{dist}((p2, p5), (p3,p6)) = 1/4\,(\text{dist}(p2,p3) + \text{dist}(p2,p6) + \text{dist}(p5,p3) + \text{dist}(p5,p6))$$

$$= 1/4 \times (0.15 + 0.25 + 0.28 + 0.39) = 0.27$$

$$\text{dist}((p2, p5), p4) = 1/2\,(\text{dist}(p2,p4) + \text{dist}(p5,p4))$$

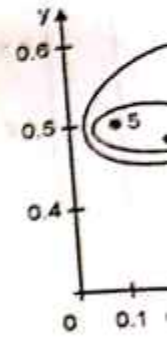
$$= 0.5 \times (0.14 + 0.29) = 0.22$$

Distance matrix :

p1	0			
(p2, p5)	**0.29**	0		
(p3, p6)	0.22	**0.27**	0	
p4	0.37	**0.22**	0.15	0
	p1	**(p2, p5)**	(p3, p6)	p4

Since, we have merged (p2, p5) together in a cluster, we now have one entry for (p2, p5) in the table, and no longer have p2 or p5 separately.

Step 5 : Now the closest
looking at the ma



dist((p3, p6, p4),

dist((p3, p6

Distance matrix :

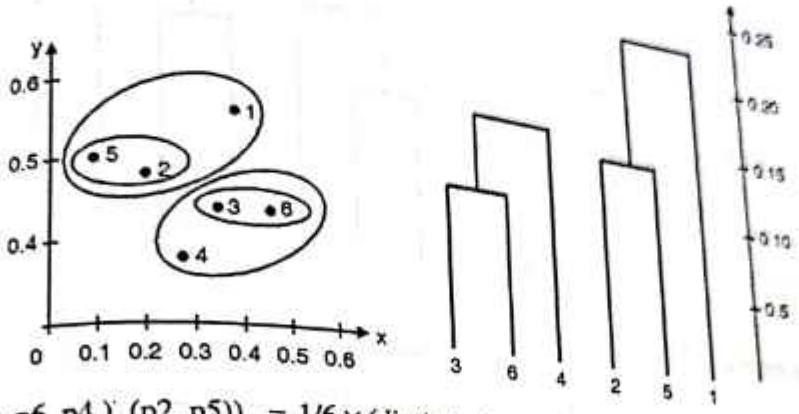
Step 6 : So merge t

dist((p3 p6 p4

Scanned by CamScanner

Step 5 : Now the closest clusters are merged, where distance is the smallest measure by looking at the maximum distance between any two points.

Space Dendrogram



$$\text{dist}((p3, p6, p4), (p2, p5)) = 1/6 \times (\text{dist}(p3,p2) + \text{dist}(p3,p5) + \text{dist}(p6,p2) + \text{dist}(p6,p5) + \text{dist}(p4,p2) + \text{dist}(p4,p5))$$

$$= 1/6 \times (0.15 + 0.28 + 0.25 + 0.39 + 0.20 + 0.29) = 0.26$$

$$\text{dist}((p3, p6, p4), p1) = 1/3 \times (\text{dist}(p3,p1) + \text{dist}(p6,p1) + \text{dist}(p4,p1))$$

$$= 1/3 \times (0.22 + 0.23 + 0.37) = 0.27$$

Distance matrix :

p1	0		
(p2, p5)	**0.24**	0	
(p3, p6, p4)	0.27	**0.26**	0
	p1	**(p2, p5)**	(p3, p6, P4)

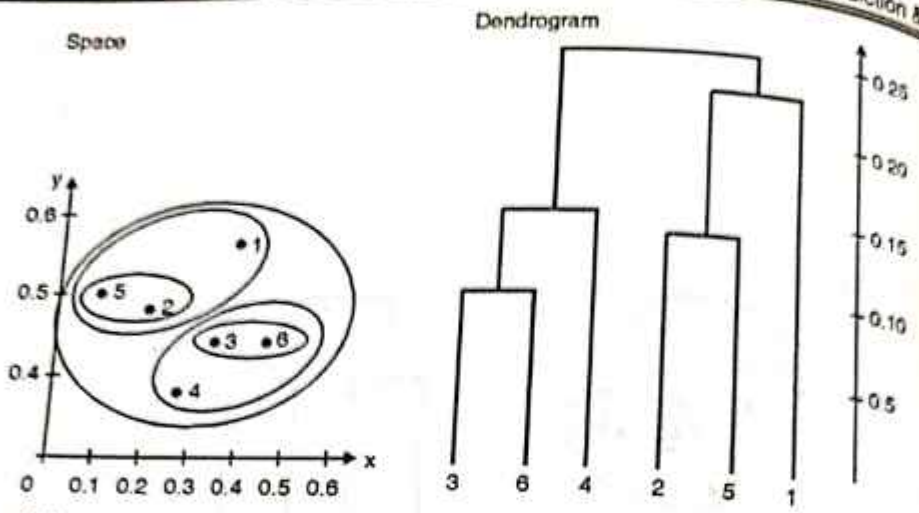
Step 6 : So merge the cluster (p2,p5) and p1.

$$\text{dist}((p3\ p6\ p4), (p2,p5,p1)) = 1/9 \times (\text{dist}(p3,p2) + \text{dist}(p3,p5) + \text{dist}(p3,p1) + \text{dist}(p6,p2) + \text{dist}(p6,p5) + \text{dist}(p6,p1) + \text{dist}(p4,p2) + \text{dist}(p4,p5) + \text{dist}(p4+p1))$$

$$= 1/9 \times (0.15 + 0.28 + 0.22 + 0.25 + 0.39 + 0.23 + 0.20 + 0.29 + 0.37)$$

$$= 0.26$$

Scanned by CamScanner

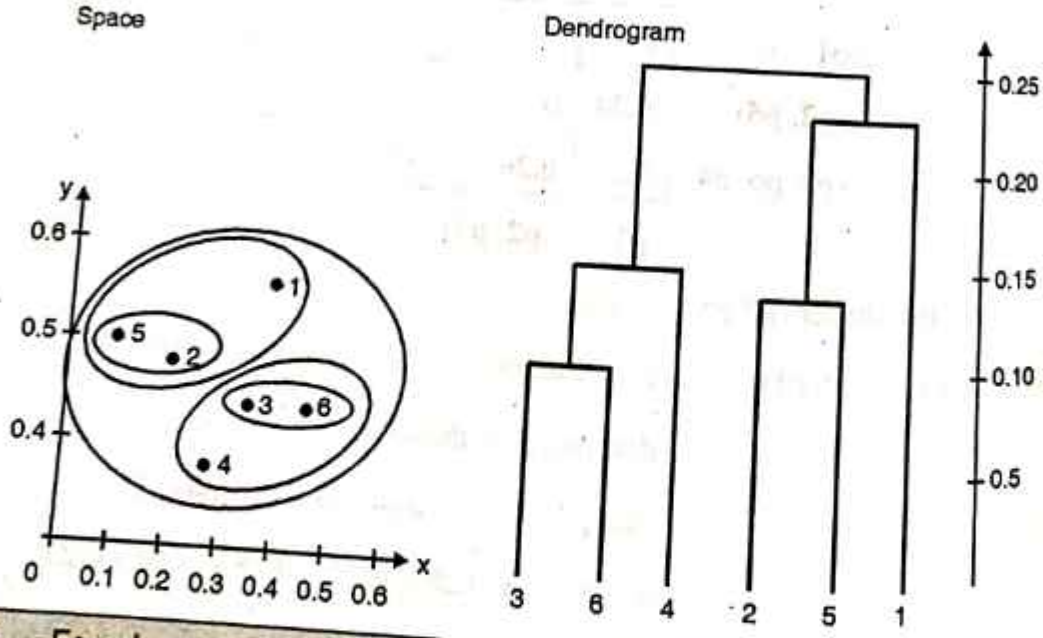


Distance matrix :

(p2, p5,p1)	0	
(p3, p6, p4)	0.26	0
	(p2, p5,p1)	(p3, p6, P4)

We need to re-compute the distance from all other points / clusters to our new cluster - (p2, p5,p1) to (p3,p6,p4) and enter the maximum distance in the above matrix (use original distance matrix).

Finally merge the cluster (p2, p5,p1) and (p3,p6,p4).



Ex. 4.9.2 : For given distance matrix, draw single link, complete link and average link dendrogram.

MU - May 2013, 10 Marks

Scanned by CamScanner

Soln. :

(i) **Single link**

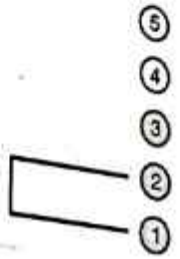
Step 1 :

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array}$$

$d_{(1,2)3} = \min\{d_{1,3}, d_{2,3}\} = \min\{6, 3\} = 3$

$d_{(1,2)4} = \min\{d_{1,4}, d_{2,4}\} = \min\{10, 9\} = 9$

$d_{(1,2)5} = \min\{d_{1,5}, d_{2,5}\} = \min\{9, 8\} = 8$

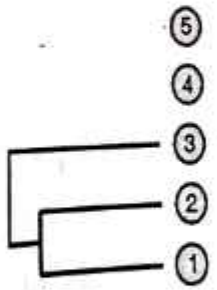


Step 2 :

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2,3) & 4 & 5 \\ \hline (1,2,3) & 0 & & \\ 4 & 7 & 0 & \\ 5 & 5 & 4 & 0 \end{array}$$

$d_{(1,2,3),4} = \min\{d_{(1,2),4}, d_{3,4}\} = \min\{9, 7\} = 7$

$d_{(1,2,3)5} = \min\{d_{(1,2),5}, d_{3,5}\} = \min\{8, 5\} = 5$



Scanned by CamScanner

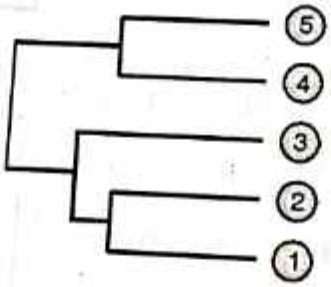
Step 3 :

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2,3) & 4 & 5 \\ \hline (1,2,3) & 0 & & \\ 4 & 7 & 0 & \\ 5 & 5 & 4 & 0 \end{array}$$

$$\Downarrow$$

$$\begin{array}{c|cc} & (1,2,3) & 4,5 \\ \hline (1,2,3) & 0 & \\ (4,5) & 5 & 0 \end{array}$$

$$d_{(1,2,3),(4,5)} = \min\{d_{(1,2,3),4}, d_{(1,2,3),5}\} = \min\{7, 5\} = 5$$



(ii) Complete link

Step 1 :

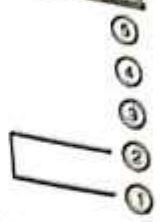
$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 6 & 0 & & \\ 4 & 10 & 7 & 0 & \\ 5 & 9 & 5 & 4 & 0 \end{array}$$

Scanned by CamScanner

$d_{(1,2),3} = \max\{d_{1,3}, d_{2,3}\} = \max\{6, 3\} = 6$

$d_{(1,2)4} = \max\{d_{,1,4}, d_{2,4}\} = \max\{10, 9\} = 10$

$d_{(1,2)5} = \max\{d_{,1,5}, d_{2,5}\} = \max\{9, 8\} = 9$

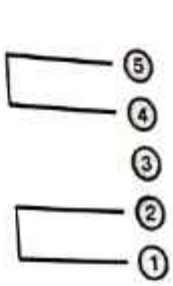


Step 2 :

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 6 & 0 & & \\ 4 & 10 & 7 & 0 & \\ 5 & 9 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2) & 3 & (4,5) \\ \hline (1,2) & 0 & & \\ 3 & 6 & 0 & \\ (4,5) & 10 & 7 & 0 \end{array}$$

$d_{(1,2),(4,5)} = \max\{d_{(1,2),4}, d_{(1,2),5}\} = \max\{10, 9\} = 10$

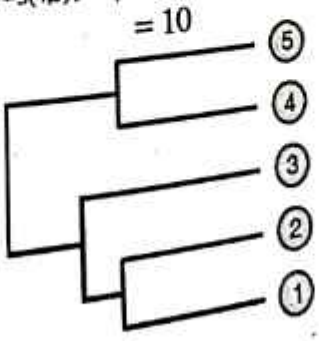
$d_{3,(4,5)} = \max\{d_{3,4}, d_{3,5}\} = \max\{7, 5\} = 7$



Step 3 :

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 6 & 0 & & \\ 4 & 10 & 7 & 0 & \\ 5 & 9 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2) & 3 & (4,5) \\ \hline (1,2) & 0 & & \\ 3 & 6 & 0 & \\ (4,5) & 10 & 7 & 0 \end{array}$$

$d_{(1,2,3),(4,5)} = \max\{d_{(1,2),(4,5)}, d_{3(4,5)}\} = \max\{10, 7\}$
$= 10$



$$\begin{array}{c|cc} & (1,2,3) & (4,5) \\ \hline (1,2,3) & 0 & \\ (4,5) & 10 & 0 \end{array}$$

Scanned by CamScanner

(iii) Average link

Step 1 :

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 4.5 & 0 & & \\ 4 & 9.5 & 7 & 0 & \\ 5 & 8.5 & 5 & 4 & 0 \end{array}$$

$$d_{(1,2),3} = \frac{1}{2}(d_{1,3} + d_{2,3}) = \frac{6+3}{2} = 4.5$$

$$d_{(1,2),4} = \frac{1}{2}(d_{1,4} + d_{2,4}) = \frac{10+9}{2} = 9.5$$

$$d_{(1,2),5} = \frac{1}{2}(d_{1,5} + d_{2,5}) = \frac{9+8}{2} = 8.5$$

Step 2 :

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 4.5 & 0 & & \\ 4 & 9.5 & 7 & 0 & \\ 5 & 8.5 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2) & 3 & (4,5) \\ \hline (1,2) & 0 & & \\ 3 & 4.5 & 0 & \\ (4,5) & 9 & 6 & 0 \end{array}$$

$$d_{(1,2),(4,5)} = \frac{1}{4}\{d_{1,4} + d_{1,5} + d_{2,4} + d_{2,5}\}$$

$$= 1/4\ \{10 + 9 + 9 + 8\}$$

$$= 9$$

$$d_{3,(4,5)} = \frac{1}{2}\{d_{3,4} + d_{3,5}\} = 6$$

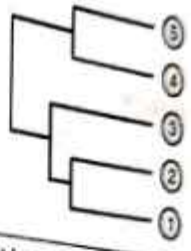
Scanned by CamScanner

Step 3 :

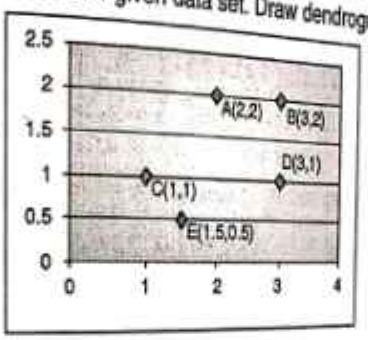
$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 4.5 & 0 & & \\ 4 & 9.5 & 7 & 0 & \\ 5 & 8.5 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2) & 3 & (4,5) \\ \hline (1,2) & 0 & & \\ 3 & 4.5 & 0 & \\ (4,5) & 9 & 6 & 0 \end{array}$$

$$d_{(1,2,3),(4,5)} = \frac{1}{6}(d_{1,4} + d_{1,5} + d_{2,4} + d_{2,5} + d_{3,4} + d_{3,5}) = 1/6\,[10+9+9+8+7+5] = 8$$

$$\begin{array}{c|cc} & (1,2,3) & (4,5) \\ \hline (1,2,3) & 0 & \\ (4,5) & 8 & 0 \end{array}$$



Ex. 4.9.3 : Use the data given below. Create adjacency matrix. Use single link or complete link algorithm to cluster given data set. Draw dendrogram.



Soln. :

Object	Attribute 1 (X)	Attribute 2 (Y)
A	2	2
B	3	2
C	1	1
D	3	1
E	1.5	0.5

Scanned by CamScanner

For simplicity we can find the adjacency matrix which gives distances of all object from each other. Using Euclidean distance we have

$$D(i,j) = \sqrt{|x_2 - x_1|^2 + |y_2 - y_1|^2}$$

$$D(A,B) = \sqrt{(2-3)^2 + (2-2)^2} = 1,$$

Similarly we can compute for the rest.

	A	B	C	D	E
A	0				
B	1	0			
C	1.41	2.24	0		
D	1.41	1	2	0	
E	1.58	2.12	0.71	1.58	0

(i) Singlelink

Step 1 : Since C, E is minimum we can combine clusters C, E

	A	B	(C,E)	D
A	0			
B	1	0		
(C,E)	1.41	2.12	0	
D	1.41	1	1.58	0

Step 2 : Now A and B is having minimum value therefore we merge these two clusters.

	(A, B)	(C,E)	D
(A,B)	0		
(C, E)	1.41	0	
D	1	1.58	0

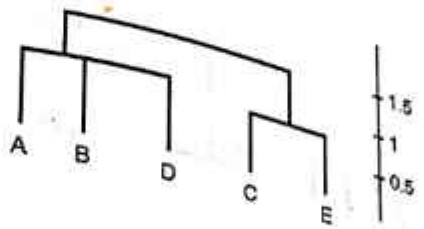
Step 3 : Cluster (A,B) and D can be merged together as they are having minimum distance value.

	(A, B,D)	(C,E)
(A,B,D)	0	
(C, E)	1.41	0

Scanned by CamScanner

Step 4 : In the last step there are only two clusters to be combined they are, (A,B,D) and (C,E).

Now the final dendrogram is



(II) Complete link

Step 1 : Closest clusters are merged where the distance is the smallest measured by looking at the maximum distance between any two point.

Since C, E is minimum we can combine clusters C, E.

	A	B	(C, E)	D
A	0			
B	1	0		
(C,E)	1.58	2.24	0	
D	1.41	1	2	0

Step 2 : Now A and B is having minimum closest measure value therefore we merge these two clusters.

	(A, B)	(C,E)	D
(A,B)	0		
(C, E)	2.24	0	
D	1.41	2	0

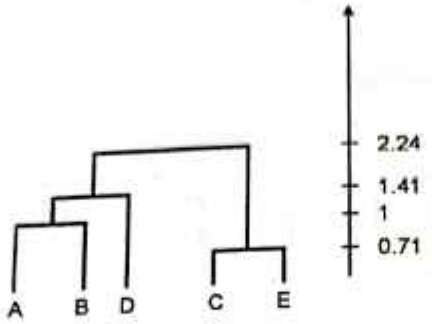
Step 3 : Cluster (A,B) and D can be merged together as they are having minimum distance value.

	(A, B,D)	(C,E)
(A,B,D)	0	
(C, E)	2.24	0

Step 4 : In the last step there are only two clusters to be combined they are, (A,B,D) and (C,E).

Scanned by CamScanner

Final dendrogram



Ex. 4.9.4 : Discuss the agglomerative algorithm using following data and plot a dendrogram using single link approach. The following figure contains sample data items indicating the distance between the elements. **MU - May 2010, 10 Marks**

Item	E	A	C	B	D
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

Soln. :

Given :

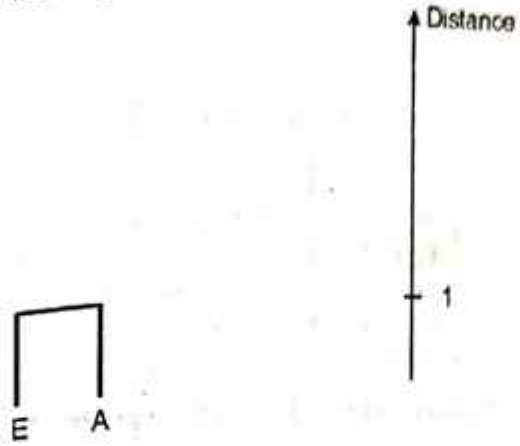
Distance matrix

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

Step 1 : From above given distance matrix, E and A clusters are having minimum distance, so merge them together to form cluster (E, A).

Scanned by CamScanner

Dendrogram



Distance matrix

dist((E A), C) = MIN (dist(E,C), dist(A,C)) = MIN (2,2) = 2

dist((E A), B) = MIN (dist(E,B), dist(A,B)) = MIN (2,5) = 2

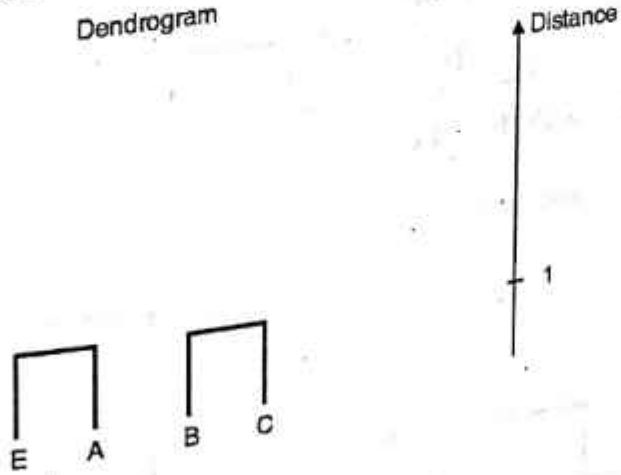
dist((E A), D) = MIN (dist(E,D), dist(A,D)) = MIN (3,3) = 3

	E, A	C	B	D
E, A	0			
C	2	0		
B	2	1	0	
D	3	6	3	0

Step 2 : Consider the distance matrix obtained in step 1 (given above)

Since B,C distance is minimum, we combine B and C.

Dendrogram



Distance matrix :

dist((B C), (E A)) = MIN (dist(B E), dist(B A), dist(C E), dist(C A))

= MIN (2, 5, 2, 2) = 2

Scanned by CamScanner

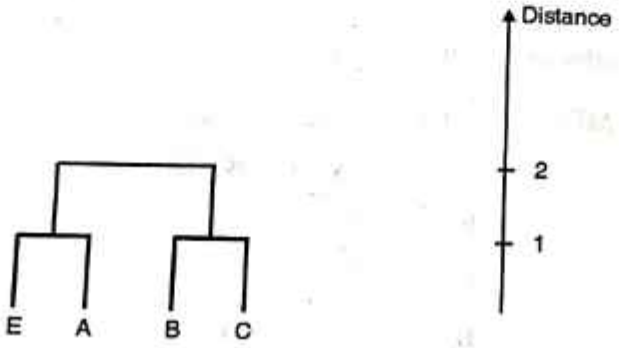
$$dist((B\ C),\ D) = MIN\ (dist(B,D),\ dist(C\ D))$$
$$= MIN\ (3, 6) = 3$$

	E, A	B, C	D
E, A	0		
B, C	2	0	
D	3	3	0

Step 3 : Consider the distance matrix obtained in step 2 (given above)

Since (E,A) and (B,C) distance is minimum, we combine them

Dendrogram



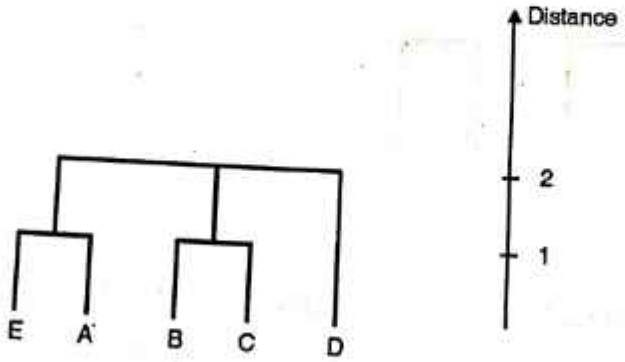
$$dist((E\ A),\ (B\ C)) = MIN\ (dist(E,B),\ dist(E\ C),\ dist(A,B),\ dist(A\ C))$$
$$= MIN\ (2, 2, 2, 5, 2) = 2$$

	E,A, B, C	D
E, A, B,C	0	
D	2	0

Step 4 : Finally combine D with (E A B C)

Final dendrogram

Dendrogram



Scanned by CamScanner

Ex. 4.9.5 : Following table gives fat and proteins content of items. Apply single linkage clustering and construct dendrogram.

Food Item	Protein	Fat
1	1.1	60
2	8.2	20
3	4.2	35
4	1.5	21
5	7.6	15
6	2.0	55
7	3.9	39

Soln. :

Step 1 : Calculate the distance matrix using Euclidian distance formula :

Cluster number	C1	C2	C3	C4	C5	C6	C7
C1	0	40.62	25.19	39.00	45.46	5.08	21.18
C2		0	15.52	6.77	5.03	35.54	19.48
C3			0	14.25	20.28	20.12	4.01
C4				0	8.55	34.00	18.19
C5					0	40.39	24.28
C6						0	16.11

From the above table, the minimum distance between any two points is 4.01 and this distance is between C3 and C7. So, these two points can be merged into a single point (cluster) and is called the C37.

Step 2 : Calculate the new distance matrix with C37 using single linkage clustering.

Therefore,

dis(C37,4) = min(dis(3,4), dis(7,4)) = min(14.25, 18.19) = 14.25

Similarly, calculate the other distances to get the distance matrix.

Cluster number	C1	C2	C37	C4	C5	C6
C1	0	40.62	21.18	39.00	45.46	5.08
C2		0	15.52	6.77	5.03	35.54
C37			0	14.25	20.28	16.11
C4				0	8.55	34.00
C5					0	40.39
C6						0

Scanned by CamScanner

In the above matrix distance between points 2 and 5 is minimum i.e. 5.03. So combine the cluster as C25.

Step 3 : Calculate the new distance matrix with C25 using single linkage clustering.

Cluster number	C1	C25	C37	C4	C6
C1	0	40.62	21.18	39.00	5.08
C25		0	15.52	6.77	35.54
C37			0	14.25	16.11
C4				0	34.00
C6					0

In the above matrix distance between C1 and C6 in minimum which is 5.08. So newly formed new cluster is C16.

Step 4 : Calculate the new distance matrix with cluster C16.

Cluster number	C16	C25	C37	C4
C16	0	35.54	16.11	34.00
C25		0	15.52	6.77
C37			0	14.25
C4				0

The minimum distance is 6.77. So combine clusters C25 and C4

Step 5 : Calculate new distance matrix.

Cluster number	C16	C254	C37
C16	0	34.00	16.11
C254		0	14.25
C37			0

Combine the clusters C254 and C37 which has minimum distance 14.25.

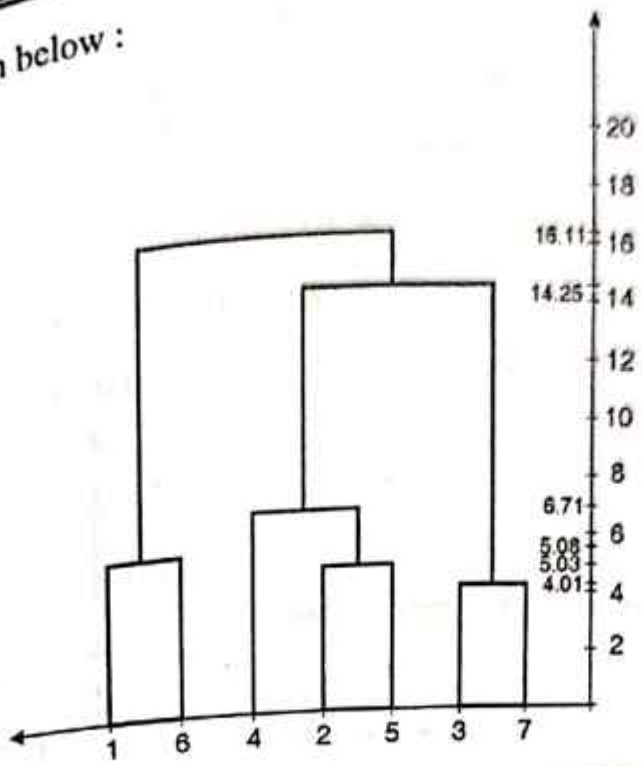
Step 6 : New distance matrix is

Cluster number	C16	C25437
C16	0	16.11
C25437		0

So finally combine the clusters C25437 and C16.

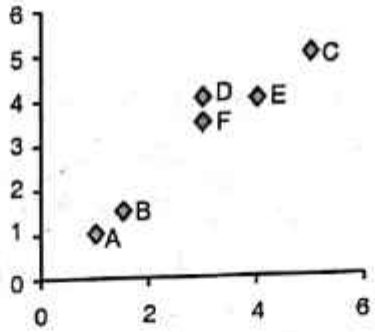
Scanned by CamScanner

Dendrogram is given below :



Ex. 4.9.6 : Suppose we have 6 objects (with name A, B, C, D, E and F) and each object have two measured features (X1 and X2).

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5



Apply Single linkage clustering and draw Dendrogram.

Soln. : We have given an input distance matrix of size 6 by 6.

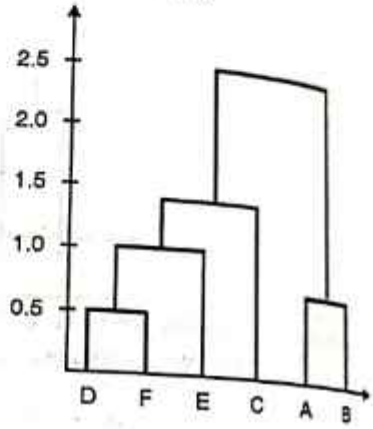
Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Scanned by CamScanner

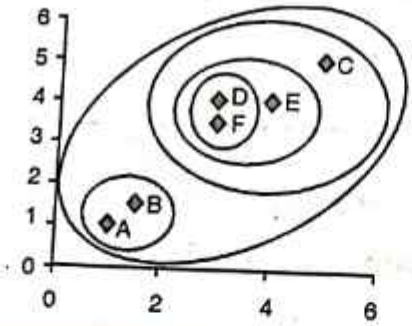
We summarized the results of computation as follow :

1. In the beginning we have 6 clusters: A, B, C, D, E and F.
2. We merge cluster D and F into cluster (D, F) at distance 0.50.
3. We merge cluster A and cluster B into (A, B) at distance 0.71.
4. We merge cluster E and (D, F) into ((D, F), E) at distance 1.00.
5. We merge cluster ((D, F), E) and C into (((D, F), E), C) at distance 1.41.
6. We merge cluster (((D, F), E), C) and (A, B) into ((((D, F), E), C), (A, B)) at distance 2.50.
7. The last cluster contain all the objects, thus conclude the computation.

Using this information, we can now draw the final results of a dendrogram.



We can also plot the clustering hierarchy into XY space.



	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

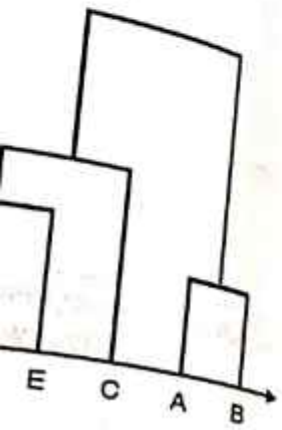
Ex. 4.9.7 : Consider the data set given. Create the adjacency matrix. Use single link agglomerative technique to cluster the given data. Draw the dendrogram.

MU - May 2011, 10 Marks

Object	X	Y
A	2	1
B	3	1
C	3	2
D	4	4
E	5	5

Scanned by CamScanner

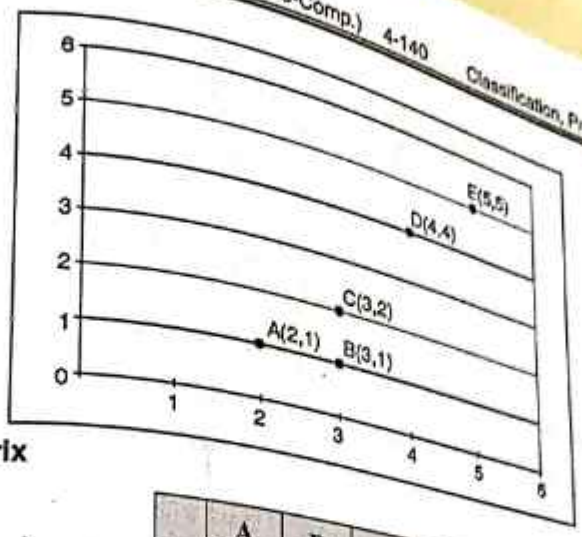
e 1.41.



. Use single link
ndrogram.

2011, 10 Marks

Soln. :



Adjacency Matrix

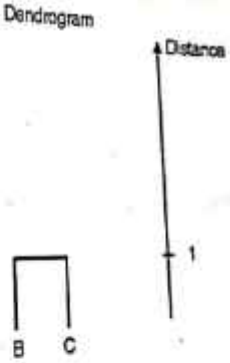
	A	B	C	D	E
A	0				
B	2	0			
C	1.41	1	0		
D	3.60	3.16	2.24	0	
E	5	4.47	3.60	1.41	0

Using Euclidean Distance we have

$$D(i,j) = \sqrt{|x2 - x1|^2 + |y2 - y1|^2}$$

$$D(A,B) = 2, \text{ similarly we can compute for the rest.}$$

Dendrogram



Distance matrix :

Step 1 : Identify the two clusters with the shortest distance in the matrix, and merge them together. Re-compute the distance matrix, as those two clusters are now in a single cluster, (no longer exist by themselves).

By looking at the distance matrix above, we see that C and B have the smallest distance from all - 1 So, we merge those two in a single cluster, and re-compute the distance matrix.

Distance matrix

$$dist((B,C), A) = MIN(dist(B,A), dist(C,A)) = 1.41$$

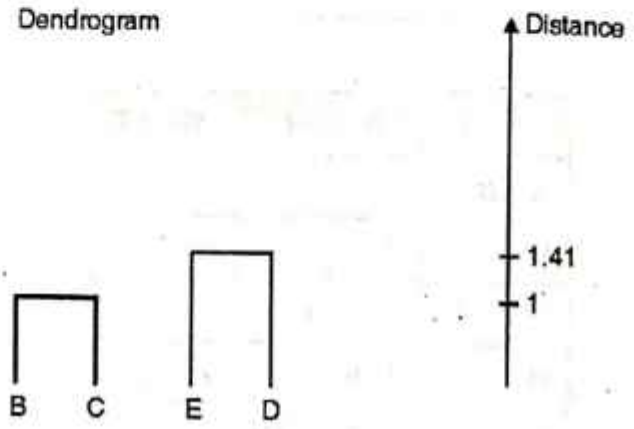
Scanned by CamScanner

Similarly calculate for other objects.

	A	B,C	D	E
A	0			
B,C	1.41	0		
D	3.60	2.24	0	
E	5	3.60	1.41	0

Step 2 : Consider the distance matrix obtained in step 1 (given above)

Since E,D distance is minimum, we combine E and D

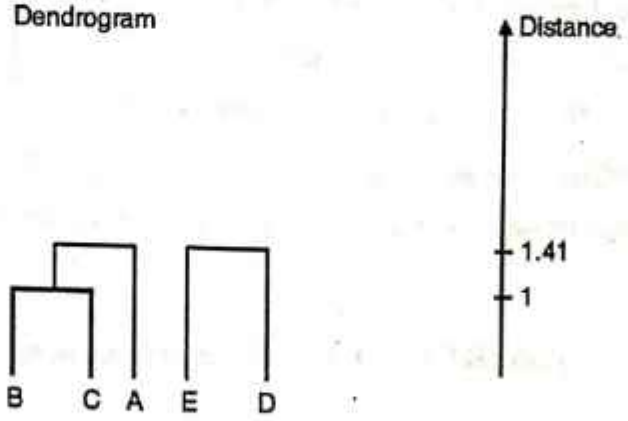


New Distance matrix :

	A	B,C	D,E
A	0		
B,C	1.41	0	
D,E	3.60	2.24	0

Step 3 : Consider the distance matrix obtained in step 2

Since (B,C) and (A) distance is minimum, we combine them



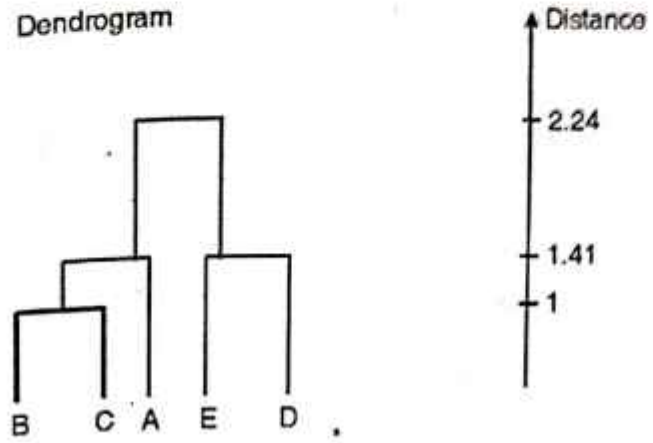
Scanned by CamScanner

New distance matrix is

	A,B,C	D,E
A,B,C	0	
D,E	2.24	0

Step 4 : Finally combine A,B,C and D,E

Final Dendrogram

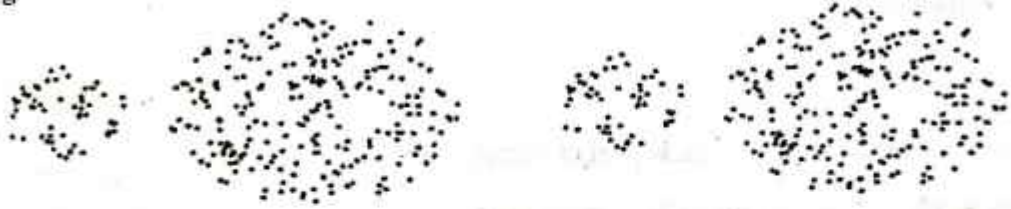


Comparison of the three methods (based on distance formula)

1. **Single-link :**

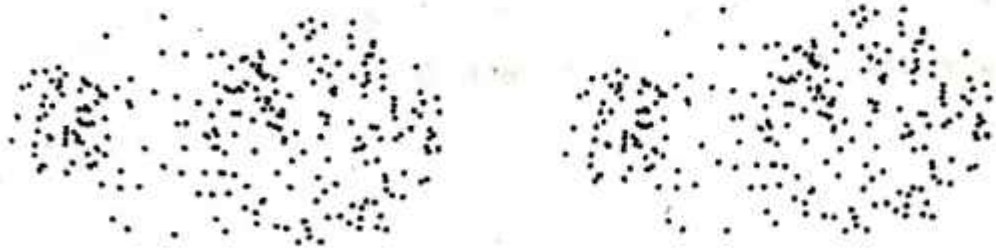
"Loose" clusters.

Strength : Can handle non-elliptical shapes.



Original points Two clusters

Limitations : Sensitive to noise and outliers.



Original points Two clusters

Scanned by CamScanner

2. Complete-link :

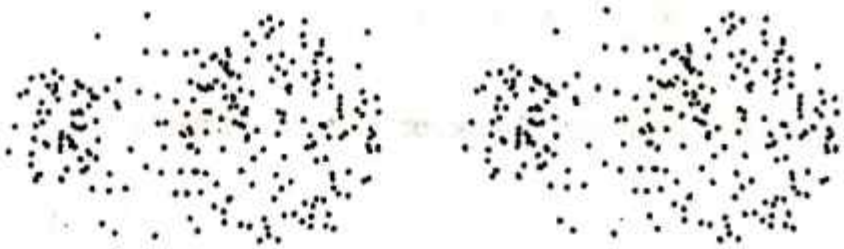
"Tight" clusters.

Strength : Less susceptible to noise and outliers.

Original points Two clusters

Limitations :

- Tends to break large clusters.
- Biased towards globular clusters.



Original points Two clusters

3. Average-link :

- "In between".
- Compromise between Single and complete link.
- **Strengths :** Less susceptible to noise and outliers.
- **Limitations :** Biased towards globular clusters.

Note : Which one is the best? Depends on what you need!

Agglomerative algorithm given by Margaret H. Dunham :

Input :

$D = \{t_1, t_2, \ldots, t_n\}$ // set of elements

A // Adjacency matrix showing distance between elements

Scanned by CamScanner

Output :

DE // Dendrogram represented as a set of ordered triples

Agglomerative algorithm:

```
D = 0;
k = n ;
K = {{t1},.....,{tn}}
DE = <d, k, K >; // Initially dendrogram contains each element in its own cluster.
Repeat
    Oldk = k;
    d = d + 1 ;
    for each pair of Ki,Kj ∈ K do
        ave = average distance between all ti ∈ Ki and tj ∈ Kd,k,Kj
        if ave <= d, then
            K = K – {Ki} – {Kj} U { Ki U Kj}
            K = oldk – 1
            DE = DE U < d,k, K> // New set of clusters added to dendrogram.
Until k = 1
```

4.9.2 Divisive Hierarchical Clustering

- In this data objects are grouped in a top down manner.
- Initially all objects are in one cluster.
- Then the cluster is subdivided into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions as the desired numbers of clusters are obtained.
- Divisive methods are not generally available, and rarely have been applied.

AGNES (AGglomerative NESting) and DIANA (Divisive ANAlysis)

- Agglomerative and divisive hierarchical clustering on data objects {1,2,3,4,5}.

Scanned by CamScanner

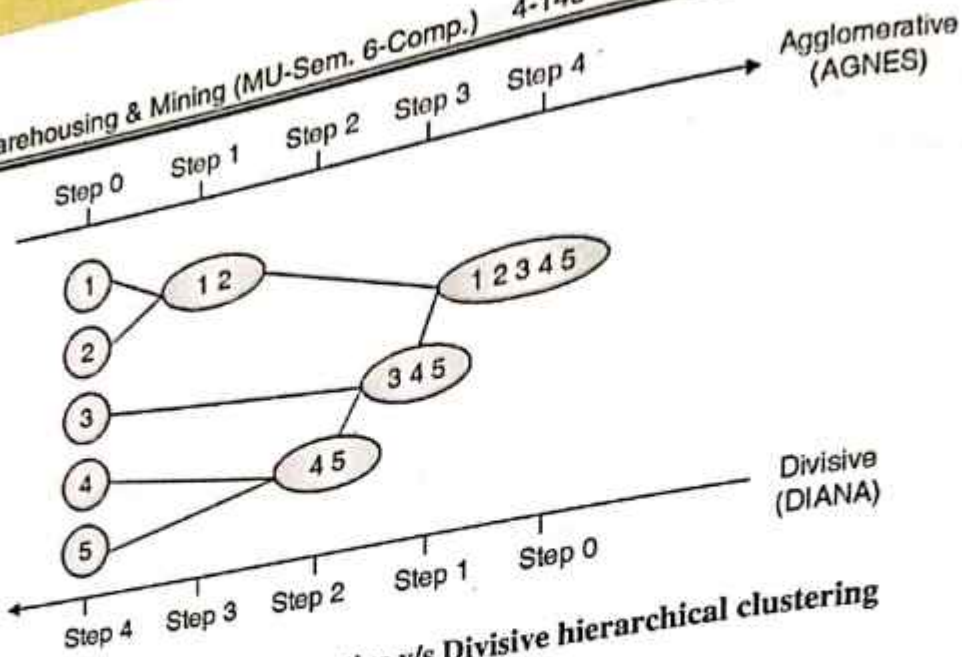


Fig. 4.9.6 : Agglomerative v/s Divisive hierarchical clustering

Advantages

- Is simple and represents the output as a hierarchy, a structure that is more informative.
- It does not require us to pre-specify the number of clusters.

Disadvantages

- Selection of merge or split points is critical as once a group of objects is merged or split, it will operate on the newly generated clusters and will not undo what was done previously.
- Thus merge or split decisions if not well chosen may lead to low-quality clusters.

Difference between agglomerative and divisive

Sr. No.	Agglomerative	Divisive
1.	Start with the points as individual clusters.	Start with one, all-inclusive cluster.
2.	At each step, merge the closest pair of clusters until only one cluster (or k clusters) left.	At each step, split a cluster until each cluster contains a point (or there are k clusters).

4.9.3 BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

- It is a scalable clustering method.
- BIRCH technique place significant emphasis on scalability of very large data sets.
- This technique is efficient for data where average make sense.

Scanned by CamScanner

- It is based on the notation of CF (Clustering Feature).
- CF tree is a height-balanced tree that stores the clustering features for Hierarchical clustering.
- Cluster of data points (vector) is represented by a triple numbers (N,LS,SS).

where,

N = Number of items in the sub cluster,

LS = Linear sum of the points,

SS = Sum of the square of points

- A CF Tree is a height-balanced tree that stores the clustering features in a hierarchy.
- Internal nodes store the sums of their descendants.

A CF tree structure is given as below :

- Each non-leaf node has at most B entries.
- Each leaf node has at most L CF entries which each satisfy threshold T, a maximum diameter or radius.
- P (page size in bytes) is the maximum size of a node.
- **Compact :** Each leaf node is a sub cluster, not a data point.

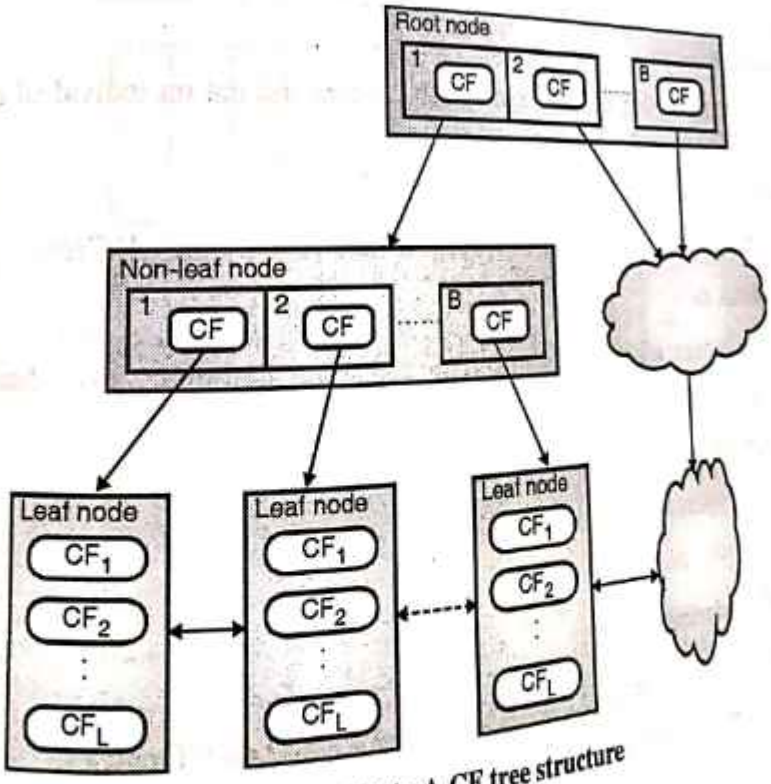


Fig. 4.9.7 : A CF tree structure

Scanned by CamScanner

Detail algorithm as a flow-graph

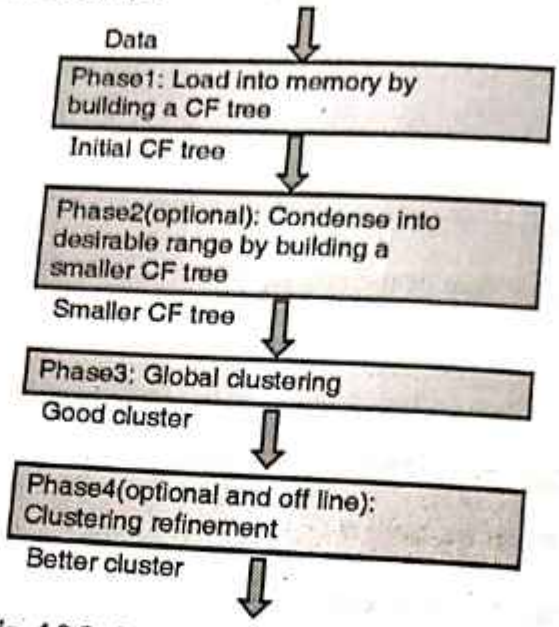


Fig. 4.9.8 : Flow chart of BIRCH algorithm

Algorithm

Step 1 : The data is loaded into the memory.

- In one scan a CF tree in memory is built with the data.
- Subsequent phases are :
  - Fast (The processing is carried out on sub clusters and not on individual data points, no more I/O needed)
  - Accurate (outliers are separated)
  - Less order Sensitive (As initial ordering of data is done by the CF-Tree)

Step 2 : Condense data

- Data resizing is done, which helps step 3 to be executed on optimally sized data.
- With a Larger T, CF tree is rebuilt.
- More outliers are removed.
- Crowded sub clusters are grouped together.
- Condensing is optional.

Step 3 : Global clustering

- Use clustering algorithm (e.g., HC, KMEANS, CLARANS) on CF entries.
- The problem is fixed where natural clusters span nodes.

Scanned by CamScanner

Step 4 : Cluster refining

- Extra passes over the dataset are carried out and the data points are reassigned to the closest centroid from step 3.
- Refining is optional.
- The problem with CF trees is fixed where different leaf entries are assigned the same valued data points.
- Always converges to a minimum.
- Allows to discard more outliers.

Example :

- Clustering feature :

$$CF = (N, LS, SS)$$

N : Number of data points

$$LS : \sum_{i=1}^{N} = X_i$$

$$SS : \sum_{i=1}^{N} = X_i^2$$

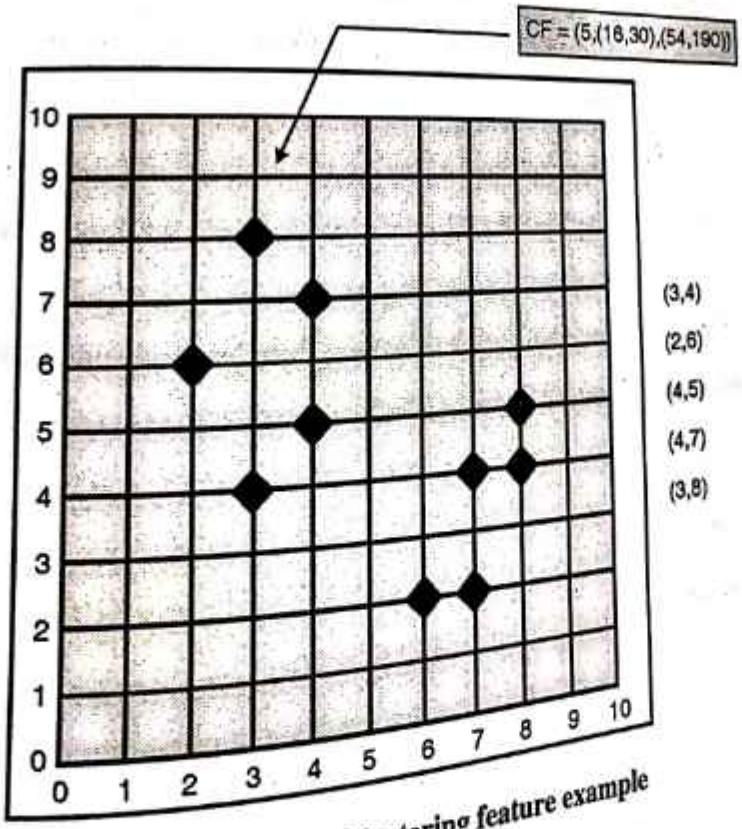


Fig. 4.9.9 : Clustering feature example

Scanned by CamScanner

$N = 5$

$NS = (16,30)$ i.e. $3 + 2 + 4 + 4 + 3 = 16$ and $4 + 6 + 5 + 7 + 8 = 30$

$SS = (54,190)$ i.e. $3^2 + 2^2 + 4^2 + 4^2 + 3^2 = 54$ and $4^2 + 6^2 + 5^2 + 7^2 + 8^2 = 190$

- **Advantage :** Finds a good clustering with a single scan and improves the quality with a few additional scans.
- **Weakness :** Handles only numeric data, and sensitive to the order of the data record.
- **Complexity :** Algorithm complexity O(n) where n is number of objects to be clustered.

Practical use of BIRCH

1. Pixel classification in images :
   - From top to bottom
   - BIRCH classification
   - Visible wavelength band
   - Near-infrared band

Fig. 4.9.10

2. Image compression using vector quantization :
   - Generate codebook for frequently occurring patterns.
   - BIRCH performs faster than CLARANS or LBG, while getting better compression and nearly as good quality.
3. BIRCH works with very large data sets.
4. Explicitly bounded by computational resource i.e. it runs with specified amount of memory.
5. Superior to CLARANS and KMEANS in terms of quality, speed, stability and scalability.

4.9.4 Advantages and Disadvantages of Hierarchical Clustering

Advantages

- It is simple and outputs a hierarchy, a structure that is more informative.
- It does not require pre-specifying the number of clusters.

Scanned by CamScanner

Disadvantages

- Combined clusters cannot be split.
- It doesn't give proper clustering if data is noisy as sensitive to noise and outliers.
- Breaking of big cluster into smaller pieces is one of the problems.

4.10 University Questions and Answers

May 2010

Q.1 What is K-means clustering ? Confer the K-means algorithm with the following data for two clusters. Data set {10, 4, 2, 12, 3, 20, 30, 11, 25, 31} *(Ans. : Refer section 4.8.1 and Ex. 4.8.3)* **(10 Marks)**

Q.2 What is Clustering Technique ? Discuss the Agglomerative algorithm using following data and plot a Dendrogram using single link approach. The following figure contains sample data items indicating the distance between the elements :

Item	E	A	C	B	D
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

(Ans. : Refer section 4.5.2, 4.9.1 and Ex. 4.9.4) **(10 Marks)**

Dec. 2010

Q.3 What is Clustering ? Explain K-means clustering algorithm. Suppose the data for clustering is {2, 4, 10, 12, 3, 20, 30, 11, 25} consider K = 2, cluster the given data using above algorithm. *(Ans. : Refer section 4.5.1 and Ex. 4.8.1)* **(10 Marks)**

Q.4 Explain Partitioning Method for Clustering. *(Ans. : Refer section 4.8)* **(10 Marks)**

May 2011

Q.5 Consider the data set given. Create the adjacency matrix. Use single link agglomerative technique to cluster the given data. Draw the dendrogram.

Scanned by CamScanner

Given :

Object	X	Y
A	2	1
B	3	1
C	3	2
D	4	4
E	5	5

(Ans. : Refer Ex. 4.9.7)

Dec. 2011 **(10 Marks)**

Q. 6 Write short notes on Similarity and distance measures in data mining.
(Ans. : Refer section 4.7)

May 2012 **(10 Marks)**

Q. 7 Explain what is meant by clustering. State and explain the various types with suitable example for each. *(Ans. : Refer sections 4.5.1 and 4.5.2)*

Dec. 2012 **(10 Marks)**

Q. 8 Give five examples of applications that can use Clustering. Describe any one clustering algorithm with the help of an example.
(Ans. : Refer sections 4.5.1 and 4.8.1)

May 2013 **(10 Marks)**

Q. 9 Illustrate how the supermarket can use clustering methods to improve sales. **(5 Marks)**

Ans. :

- Identifies the key store characteristics that should directly impact a store's range, merchandising, pricing, advertising and promotions.
- To provide economies of scale to this process for a chain or supplier undertaking, stores with the same mix of characteristics can be grouped into 'store clusters'. This enables more targeted space management, merchandising, promotional and advertising propositions to be developed cost effectively.
- The considered use of store profiles is the opposite of the 'one size fits all' or cookie cutter approach. It enables stores to specifically target key shopper groups and competitors to make the best use of floor and shelf space and to employ the most relevant and effective marketing tactics.

Scanned by CamScanner

- This laser focus cuts down on waste - wasted investment in holding stock, in allocating shelf space to slow selling lines, in discounting lines that have limited appeal to the store's shoppers, in advertising and promotions that do not motivate shoppers to shop more often or to spend more.
- The demographic profile of a store enables retailers and suppliers to identify and deliver targeted product offerings needed for shoppers of specific religious, cultural and ethnic backgrounds, as well as different life stages and incomes.
- Mission profiles reveal details about when shoppers shop – monthly shopping, after work pop in, lunch time shopping or weekend shopping expeditions.
- An example of the way this information could be used would be if a significant number of a store's shoppers are popping in for bread and milk, one could argue that these products should be located near to the front of the store to improve the shopping experience and discourage these shoppers from choosing a convenience store instead. Surrounding the bread and milk with related impulse lines would also help increase the basket and the margin on these shopper missions.

Q.10 Apply Agglomerative Hierarchical Clustering and draw single link and average lank dendrogram for the following distance matrix.

	A	B	C	D	E
A	0	2	6	10	9
B	2	0	3	9	8
C	6	3	0	7	5
D	10	9	7	0	4
E	9	8	5	4	0

(Ans. : Refer Ex. 4.9.2) **(10 Marks)**

Q.11 Write detailed notes on K-Means Clustering. *(Ans. : Refer section 4.8.1)* **(10 marks)**

Dec. 2013

Q.12 Give five examples of application that can use clustering. Describe any one clustering algorithm with an example. *(Ans. : Refer sections 4.5.1 and 4.8.1)* **(10 Marks)**

Scanned by CamScanner

May 2014

Q. 13 Describe the working of the K-Means clustering algorithm with the help of the a sample dataset. *(Ans. : Refer section 4.8.1 and Ex. 4.8.1)* **(10 Marks)**

Q. 14 Write detailed notes on Hierarchical clustering methods. *(Ans. : Refer section 4.9)* **(10 Marks)**

Dec. 2014

Q. 15 Illustrate how the supermarket can use clustering methods to improve sales. *(Ans. : Refer Q. 10 of May 2013)* **(5 Marks)**

Q. 16 Explain hierarchical clustering methods. *(Ans. : Refer section 4.9)* **(10 Marks)**

Q. 17 Write detailed notes on K-Means clustering. *(Ans. : Refer section 4.8.1)* **(10 Marks)**

May 2016

Q. 18 Explain K-means clustering algorithm ? Apply K-means algorithms for the following data set with two clusters. Data set = {1,2, 6, 7, 8 8 10, 15, 17,20}. *(Ans. : Refer section 4.8.1 and Ex. 4.8.7)* **(10 Marks)**

❑❑❑

Chapter Ends

Scanned by CamScanner

the help of the a sample
(10 Marks)

(10 Marks)

ove sales.
(5 Marks)

(10 Marks)

9.1) (10 Marks)

for the following

(10 Marks)

□□□

Chapter Ends

CHAPTER 5

Module 5

Mining Frequent Patterns and Association Rules

Syllabus :

Market Basket Analysis, Frequent Item sets, Closed Item sets, and Association Rule, Frequent Pattern Mining, Efficient and Scalable Frequent Item set Mining Methods : Apriori Algorithm, Association Rule Generation, Improving the Efficiency of Apriori, FP growth, Mining frequent Itemsets using Vertical Data Format, Introduction to Mining Multilevel Association Rules and Multidimensional Association Rules.

Syllabus Topic : Market Basket Analysis

5.1 Market Basket Analysis

→ **(MU - May 2012, Dec. 2013)**

5.1.1 What is Market Basket Analysis?

- Market basket analysis is a modelling technique which is also called as affinity analysis, it helps identifying which items are likely to be purchased together.
- The market-basket problem assumes we have some large number of items, e.g., "bread", "milk.", etc. Customers buy the subset of items as per their need and marketer gets the information that which things customers have taken together. So the marketers use this information to put the items on different position.
- **For Example :** If someone buys a packet of milk also tends to buy a bread at the same time

Milk => Bread

Scanned by CamScanner

- Market basket analysis algorithms are straightforward; difficulties arise mainly in dealing with large amounts of transactional data, where after applying algorithm it may give rise to large number of rules which may be trivial in nature.

5.1.2 How is it Used ?

- Market basket analysis is used in deciding the location of items inside a store, for e.g. if a customer buys a packet of bread he is more likely to buy a packet of butter too, keeping the bread and butter next to each other in a store would result in customers getting tempted to buy one item with the other.
- The problem of large volume of trivial results can be overcome with the help of differential market basket analysis which enables in finding interesting results and eliminates the large volume.
- Using differential analysis it is possible to compare results between various stores, between customers in various demographic groups.
- Some special observations among the rules for e.g. if there is a rule which holds in one store but not in any other (or vice versa) then it may be really interesting to note that there is something special about that store in the way it has organized its items inside the store may be in a more lucrative way. These types of insights will improve company sales.
- Identification of sets of items purchases or events occurring in a sequence , something that may be of interest to direct marketers, criminologists and many others, this approach may be termed as Predictive market basket analysis.

5.1.3 Applications of Market Basket Analysis

- Credit card transactions done by a customer may be analysed.
- Phone calling patterns may be analysed.
- Fraudulent Medical insurance claims can be identified.
- For a financial services company :
  - Analysis of credit and debit card purchases.
  - Analysis of cheque payments made.
  - Analysis of services/products taken e.g. a customer who has taken executive credit card is also likely to take personal loan of $5,000 or less.
- For a telecom operator :
  - Analysis of telephone calling patterns.

Scanned by CamScanner

ties arise mainly in dealing
algorithm it may give rise

nside a store, for e.g. if a
et of butter too, keeping
stomers getting tempted

ome with the help of
interesting results and

tween various stores,

e which holds in one
ing to note that there
tems inside the store
company sales.

nce , something that
this approach may

xecutive credit

- Analysis of value-added services taken together. Rather than considering services taken together at a point in time, it could be services taken over a period of, let's say, six months.

Various ways can be used to apply market basket analysis :

- Special combo offers may be offered to the customers on the products sold together.
- Placement of items nearby inside a store which may result in customers getting tempted to buy one product with the other.
- The layout of catalogue of an ecommerce site may be defined.
- Inventory may be managed based on product demands.

Syllabus Topic : Frequent Item Sets, Closed Item Sets and Association Rule

5.2 Frequent Item Sets, Closed Item Sets and Association Rule

5.2.1 Frequent Itemsets

- An itemset X is *frequent* if X's support is no less than a *minimum support* threshold.
- A frequent itemset is a set of items that appears at least in a pre-specified number of transactions. Frequent itemsets are typically used to generate association rules.
- Consider a data set S, frequent itemset in S are those items that appear in at least a fraction s of the basket, where s is a chosen constant with a value of 0.01 or 1%.
- To find frequent itemsets one can use the monotonicity principle or a-priori trick which is given as,

  If a set of items say S is frequent then all its subsets are also frequent.
- The procedure to find frequent itemsets :
  - A level wise search may be conducted to find the frequent-1 items(set of size 1), then proceed to find frequent -2 items and so on.
  - Next search for all maximal frequent itemsets.

5.2.2 Closed Itemsets

- An itemset is closed if none of its immediate supersets has the same support as the itemset.

Scanned by CamScanner

- Consider two itemsets X and Y, if every item of X is in Y but there is at least one item of Y, which is not in X, then Y is not a proper super-itemset of X. In this case, itemset X is closed.
- If X is both closed and frequent, it is known as closed frequent itemset.
- An itemset is maximal frequent if none of its immediate supersets is frequent.
- An itemset X is *maximal frequent itemset or max-itemset* if X is frequent and there exist no super itemset Y such that X is subset of Y and Y is frequent.

Example :

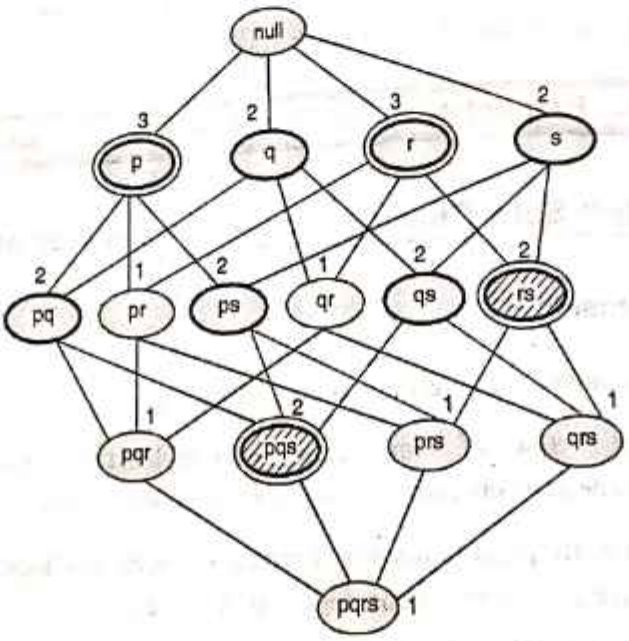


Fig. 5.2.1 : Lattice diagram for maximal, closed and frequent itemsets

Let us consider minimum support = 2.

- The itemsets that are circled with thick lines are the frequent itemsets as they satisfy the minimum support. Fig. 5.2.1, Frequent itemsets are { p,q,r,s,pq,ps,qs,rs,pqs}
- The itemsets that are circled with the double lines are closed frequent itemsets. Fig. 5.2.1, closed frequent itemsets are { p,r,rs,pqs}. For example {rs} is closed frequent itemset as all of its superset {prs ,qrs} have support less than 2.
- The itemsets that are circled with the double lines and shaded are maximal frequent itemsets. Fig. 5.2.1, maximal frequent itemsets are {rs,pqs}. For example {rs} is maximal frequent itemset as none of its immediate supersets like {prs, qrs} is frequent.

Scanned by CamScanner

5.2.3 Association Rules

→ **(MU - Dec. 2011, May 2012, Dec. 2013)**

- The items or objects in Relational databases, transactional databases or other information repositories are considered for finding frequent patterns, associations, correlations, or causal structures.
- It searches for interesting relationships among items in a given data set by examining transactions, or shop carts, we can find which items are commonly purchased together. This knowledge can be used in advertising or in goods placement in stores.
- Association rules have the general form

$$I_1 \rightarrow I_2 \text{ (where } I_1 \cap I_2 = 0)$$

Where, I_n are sets of items, for example can be purchased in a store.

- The rule should be read as "Given that someone has bought the items in the set I_1 they are likely to also buy the items in the set I_2".

Large Itemsets

- An itemset is a set of single items from the database of transactions.
- If some items often occur together they can form an association rule.

Support

- The support of an itemset is the count of that itemset in the total number of transactions, or in other words it is the percentage of the transactions in which the items appear.

$$\text{If } A \Rightarrow B$$

$$\text{Support } (A \Rightarrow B) = \frac{\#_\text{tuples_containing_both_A_and_B}}{\text{total_\#_of_tuples}}$$

- The support(s) for an association rule $X \Rightarrow Y$ is the percentage of transactions in the database that contains $X \cup Y$ i.e. (X and Y together).
- An itemset is considered to be a *large itemset* if its support is above some threshold.

Confidence

- The confidence or strength for an association rule $A \Rightarrow B$ is the ratio of the number of transactions that contain $A \cup B$ to the number of transactions that contain A.
- Consider a rule $A \Rightarrow B$, it is measure of ratio of the number of tuples containing both A and B to the number of tuples containing A.

Scanned by CamScanner

$$\text{Confidence } (A \Rightarrow B) = \frac{\#_\text{tuples_containing_both_A_and_B}}{\#_\text{tuples_containing_A}}$$

Finding the large Itemsets

1. **The Brute Force approach**
   - Find all the possible association rules.
   - Calculate the support and confidence for each rule generated in the above step.
   - The Rules that fail the minsup and minconf are prunned from the above list.
   - The above steps would be a time consuming process, we can have a better approach as given below.
2. **A better approach**

   The Apriori Algorithm.

Syllabus Topic : Frequent Pattern Mining

5.3 Frequent Pattern Mining

Frequent pattern mining is classified in the various ways based on following criteria :

1. **Completeness of the pattern to be mined :** Here we can mine the complete set of frequent itemset, closed frequent itemset, constrained frequent itemsets.
2. **Levels of abstraction involved in the rule set :** Here we use multilevel association rules based on the levels of abstraction of data.
3. **Number of data dimensions involved in the rule :** Here we use single dimensional association rule, there is only one dimension or multidimensional association rule if there is more than one dimension.
4. **Types of the values handled in the rule :** Here we use Boolean and quantitative association rules.
5. **Kinds of the rules to be mined :** Here we use association rules and correlation rules based on the kinds of the rules to be mined.
6. **Kinds of pattern to be mined :** Here we use frequent itemset mining, sequential pattern mining and structured pattern mining.

Scanned by CamScanner

Syllabus Topic : Efficient and Scalable Frequent Itemset Mining Methods

5.4 Efficient and Scalable Frequent Itemset Mining Method

1. Apriori Algorithm
2. FP Tree

Syllabus Topic : Apriori Algorithm

5.4.1 Apriori Algorithm

Apriori algorithm finding Frequent Itemsets using Candidate Generation → (MU - Dec. 2011)

- The Apriori Algorithm solves the frequent item sets problem.
- The algorithm analyzes a data set to determine which combinations of items occur together frequently.
- The Apriori algorithm is at the core of various algorithms for data mining problems. The best known problem is finding the association rules that hold in a basket - item relation.

Basic Idea

- An itemset can only be a large itemset if all its subsets are large itemsets.
- Frequent itemsets: The sets of items that have minimum support.
- All the subsets of a frequent itemset must be frequent for e.g. {PQ} is a frequent itemset {P} and {Q} must also be frequent.
- Find frequent itemsets frequently with cardinality 1 to k(k-itemset).
- Generate association rules from frequent itemsets.

Apriori Algorithm given by Jiawei Han et al.

Input :

D: a database of transactions;

min_sup : the minimum support count threshold.

Output : L : frequent itemsets in D.

Method :

(1) L_1 = find_frequent_1-itemsets(D);

Scanned by CamScanner

```
(2)  for (k = 2; L_{k-1} ≠ ∅; k++){
(3)  C_k = apriori_gen (L_{k-1});
(4)  for each transaction t ∈ D { // scan D for counts
(5)          C_t = subset (C_k, t); // get the subsets of t that are candidates
(6)          for each candidate c ∈ C_t
(7)                  c.count ++ ;
(8)  }
(9)  L_k = {c ∈ C_k | c.count ≥ min_sup}
(10) }
(11) return L = U_k L_k;
```

```
Procedure apriori_gen (L_{k-1}: frequent (k - 1) - itemsets)
(1)  for each itemset l_1 ∈ L_{k-1}
(2)  for each itemset l_2 ∈ L_{k-1}
(3)          if (l_1 (1) = l_2 (l_1) ∧ (l_1 (2) = l_2 (2))
             ∧ ... ∧ (l_1 [k - 2] = l_2 [k - 2]) ∧ (l_1 [k - 1] < l_2 [k - 1]) then {
(4)                  c = l_1 ⋈ l_2; // join step:generate candidates
(5)                  if has_infrequent_subset(c, L_{k-1}) then
(6)                          delete c; // prune step: remove unfruitful candidate
(7)                  else add c to C_k
(8)          }
(9)  return C_k;
```

```
Procedure has_infrequent_subset(c: candidate k-intemset;
        L_{k-1}: frequent (k - 1) -itemsets); // use prior knowledge
(1)  for each (k - 1)-subset s of c
(2)          if s ∉ L_{k-1} then
(3)                  return TRUE;
(4)  return FALSE;
```

5.4.2 Advantages and Disadvantages of Apriori Algorithm

Some of the advantages and disadvantages of Apriori Algorithm are listed as :

Scanned by CamScanner

ntages

The algorithm makes use of large itemset property.

The method can be easily parallelized.

The algorithm is easy from implementation point of view.

dvantages

Although the algorithm is easy to implement it needs many database scans which reduces the overall performance.

Due to Database scans, the algorithm assumes transaction database is memory resident.

.3 Solved Examples on Apriori Algorithm

.5.4.1 : Given the following data, apply the Apriori algorithm. Min support = 50 % Database D.

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

ln :

tep 1 : Scan D for count of each candidate. The candidate list is {1, 2, 3,4, 5} and find the support.

C_1 =

Itemset	Sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

Scanned by CamScanner

Step 2 : Compare candidate support count with minimum support count (i.e. 50%)

L_1=

Itemset	Sup.
{1}	2
{2}	3
{3}	3
{5}	3

Step 3 : Generate candidate C_2 from L_1

C_2 =

Itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Step 4 : Scan D for count of each candidate in C_2 and find the support

C_2 =

Itemset	Sup.
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

Scanned by CamScanner

n support count (i.e. 50%)

Step 5: Compare candidate (C_2) support count with the minimum support count

L_2=

Itemset	Sup.
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

Step 6: Generate candidate C_3 from L_2

C_3 =

Itemset
{1,3,5}
{2,3,5}
{1,2,3}

Step 7: Scan D for count of each candidate in C_3

C_3 =

Itemset	sup
{1,3,5}	1
{2,3,5}	2
{1,2,3}	1

Step 8: Compare candidate (C_3) support count with the minimum support count

L_3 =

Itemset	sup
{2,3,5}	2

Step 9: So data contain the frequent itemset(2,3,5)

Therefore the association rule that can be generated from L_3 are as shown below with the support and confidence.

Scanned by CamScanner

Association Rule	Support	Confidence	Confidence %
2^3=>5	2	2/2=1	100%
3^5=>2	2	2/2=1	100%
2^5=>3	2	2/3=0.66	66%
2=>3^5	2	2/3=0.66	66%
3=>2^5	2	2/3=0.66	66%
5=>2^3	2	2/3=0.66	66%

If the minimum confidence threshold is 70% (Given), then only the first and second rules above are output, since these are the only ones generated that are strong.

Final rules are :

Rule 1: 2^3=>5 and Rule 2 : 3^5=>2

Ex. 5.4.2 : Find the frequent item sets in the following database of nine transactions , with a minimum support 50% and confidence 50%

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Soln. :

Step 1 : Scan D for count of each candidate. The candidate list is {A,B,C,D,E,F} and find the support

C_1=

Items	Sup.
{A}	3
{B}	2
{C}	2
{D}	1
{E}	1
{F}	1

Scanned by CamScanner

Idence %
00%
00%
5%
%
%
%

first and second rules

transactions , with a

,F} and find the

Step 2: Compare candidate support count with minimum support count (50%)

L_1=

Items	Sup.
{A}	3
{B}	2
{C}	2

Step 3: Generate candidate C_2 from L_1

C_2=

Items
{A,B}
{A,C}
{B,C}

Step 4: Scan D for count of each candidate in C_2 and find the support

C_2=

Items	Sup.
{A,B}	1
{A,C}	2
{B,C}	1

Step 5: Compare candidate (C_2) support count with the minimum support count

L_2=

Items	Sup.
{A,C}	2

Step 6: So data contain the frequent item l(A,C)

Therefore the association rule that can be generated from L are as shown below with the support and confidence

Association Rule	Support	Confidence	Confidence %
A -> C	2	2/3 = 0.66	66 %
C-> A	2	2/2 = 1	100 %

Scanned by CamScanner

Minimum confidence threshold is 50% (Given), then both the rules are output as the confidence is above 50 %.

So final rules are :

Rule 1 : A - > C

Rule 2 : C- > A

Ex. 5.4.3 : Consider the transaction database given below. Use Apriori algorithm with minimum support count 2. Generate the association rules along with its confidence. **MU - Dec. 2010, May 2016, 10 Marks, May 2011, 8 Marks**

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Soln. :

Step 1 : Scan the transaction Database D and find the count for item-1 set which is the candidate. The candidate list is {I1, I2, I3, I4, I5} and find each candidates support.

C_1=

1-Itemsets	Sup-count
I1	6
I2	7
I3	6
I4	2
I5	2

Scanned by CamScanner

Step 2: Find out whether each candidate item is present in at least two transactions (As support count given is 2).

L_1=

1-Itemsets	Sup-count
1	6
2	7
3	6
4	2
5	2

Step 3: Generate candidate C_2 from L_1 and find the support of 2-itemsets.

C_2=

2-Itemsets	Sup-count
1,2	4
1,3	4
1,4	1
1,5	2
2,3	4
2,4	2
2,5	2
3,4	0
3,5	1
4,5	0

Step 4: Compare candidate (C_2) generated in step 3 with the support count, and prune those itemsets which do not satisfy the minimum support count.

L_2 =

Frequent 2-Itemsets	Sup-count
1,2	4
1,3	4
1,5	2
2,3	4
2,4	2
2,5	2

Scanned by CamScanner

Step 5 : Generate candidate C_3 from L_2

$C_3 =$

Frequent 3-Itemset
1,2,3
1,2,5
1,2,4

Step 6 : Scan D for count of each candidate in C_3 and find their support count.

$C_3 =$

Frequent 3-Itemset	Sup-count
1,2,3	2
1,2,5	2
1,2,4	1

Step 7 : Compare candidate (C_3) support count with the minimum support count and prune those itemsets which do not satisfy the minimum support count.

$L_3 =$

Frequent 3-Itemset	Sup-count
1,2,3	2
1,2,5	2

Step 8 : Frequent itemsets are {I1,I2,I3} and {I1,I2,I5}

Let us consider the frequent itemsets={I1, I2, I5}. Following are the Association rules that can be generated shown below with the support and confidence.

Association Rule	Support	Confidence	Confidence %
I1 ^ I2 => I5	2	2/4	50%
I1 ^ I5 => I2	2	2/2	100%
I2 ^ I5 => I1	2	2/2	100%
I1 => I2 ^ I5	2	2/6	33%
I2 => I1 ^ I5	2	2/7	29%
I5 => I1 ^ I2	2	2/2	100%

Scanned by CamScanner

Suppose if the minimum confidence threshold is 75% then only the following rules will be considered as output, as they are strong rules.

Rules	Confidence
I1^I5=>I2	100%
I2^I5=>I1	100%
I5=>I1^I2	100%

Q.44: Consider the following transactions :

TID	Items
01	1, 3, 4, 6
02	2,3,5,7
03	1,2,3,5,8
04	2,5,9,10
05	1,4

Apply the Apriori with minimum support of 30% and minimum confidence of 75% and find large item set L. **MU - May 2010, May 2012, Dec. 2013, 10 Marks**

Sol:

Step 1: Scan the transaction Database D and find the count for item-1 set which is the candidate. The candidate list is {1,2,3,4,5,6,7,8,9,10} and find the support.

C_1 =

Itemset	Sup-count
1	3
2	3
3	3
4	2
5	3
6	1
7	1
8	1
9	1
10	1

Scanned by CamScanner

Step 2 : Find out whether each candidate item is present in at least 30% of transactions (As support count given is 30%).

$L_1 =$

Itemset	Sup-count
1	3
2	3
3	3
4	2
5	3

Step 3 : Generate candidate C_2 from L_1 and find the support of 2-itemsets.

$C_2 =$

Itemset	Sup-count
1,2	1
1,3	2
1,4	2
1,5	1
2,3	2
2,4	0
2,5	3
3,4	1
3,5	2

Step 4 : Compare candidate (C_2) generated in step 3 with the support count, and prune those itemsets which do not satisfy the minimum support count.

$L_2 =$

Itemset	Sup-count
1,3	2
1,4	2
2,3	2
2,5	3

Scanned by CamScanner

Step 5 : Generate candidate C_3 from L_2 and find the support.

$C_3 =$

Itemset	Sup-count
1,2,3	1
2,3,5	2
1,3,4	1

Step 6 : Compare candidate (C_3) support count with min support.

$L_3 =$

Itemset	Sup-count
2,3,5	2

Therefore the database contains the frequent itemset{2,3,5}.

Following are the association rules that can be generated from L3 are as shown below with the support and confidence.

Association Rule	Support	Confidence	Confidence %
2^3=>5	2	2/2=1	100%
3^5=>2	2	2/2=1	100%
2^5=>3	2	2/3=0.66	66%
2=>3^5	2	2/3=0.66	66%
3=>2^5	2	2/3=0.66	66%
5=>2^3	2	2/3=0.66	66%

Given minimum confidence threshold is 75%, so only the first and second rules above are output, since these are the only ones generated that are strong.

Final Rules are : Rule 1: 2^3=>5 and Rule 2 : 3^5=>2

Ex. 5.4.5 : A database has four transactions. Let min sup=60% and min conf= 80%.

TID	Date	Items-bought
T100	10/15/99	{K, A, D, B}
T200	10/15/99	{D, A, C, E, B}
T300	10/19/99	{C,A,B,E}
T400	10/22/99	{B, A, D}

Find all frequent itemsets using apriori algorithm

List strong association rules(with supports S and confidence C).

Scanned by CamScanner

Soln. :

Step 1 : Scan D for count of each candidate. The candidate list is {A,B,C,D,E,K} and find the support.

C_1=

Itemset	Sup-count
A	4
B	4
C	2
D	3
E	2
K	1

Step 2 : Compare candidate support count with minimum support count (i.e. 60%).

L_1=

Itemset	Sup-count
A	4
B	4
D	3

Step 3 : Generate candidate C_2 from L_1.

C_2=

Itemset
A,B
A,D
B,D

Step 4 : Scan D for count of each candidate in C_2 and find the support.

C_2=

Itemset	Sup-count
A,B	4
A,D	3
B,D	3

Step 5 : Compare

L_2=

Step 6 : Generate ca

C_3=

Step 7 : Scan D for co

C_3=

Step 8 : Compare can

L_3 =

Step 9 : So data conta

Therefore the asso
below with the suppor

Asso
A^B=
A^D=
B^D=
A=>.
B=>.
D=>.

Scanned by CamScanner

Step 5 : Compare candidate (C_2) support count with the minimum support count.

L_2=

Itemset	Sup-count
A,B	4
A,D	3
B,D	3

Step 6 : Generate candidate C_3 from L_2.

C_3=

Itemset
A,B,D

Step 7 : Scan D for count of each candidate in C_3.

C_3=

Itemset	Sup
A,B,D	3

Step 8 : Compare candidate (C_3) support count with the minimum support count.

L_3 =

Itemset	Sup
A,B,D	3

Step 9 : So data contain the frequent itemset(A,B,D).

Therefore the association rule that can be generated from frequent itemsets are as shown below with the support and confidence.

Association Rule	Support	Confidence	Confidence %
A^B=>D	3	3/4=0.75	75%
A^D=>B	3	3/3=1	100%
B^D=>A	3	3/3=1	100%
A=>B^D	3	3/4=0.75	75%
B=>A^D	3	3/4=0.75	75%
D=>A^B	3	3/3=1	100%

Scanned by CamScanner

If the ntinimum confidence threshold is 80% (Given), then only the SECOND, THIRD AND LAST rules above are output, since these are the only ones generated that are strong.

Ex. 5.4.6 : Apply the Apriori algorithm on the following data with Minimum support = 2.

TID	List of Item_IDs
T100	I1,I2,I4
T200	I1,I2,I5
T300	I1,I3,I5
T400	I2,I4
T500	I2,I3
T600	I1,I2,I3,I5
T700	I1,I3
T800	I1,I2,I3
T900	I2,I3
T1000	I3,I5

Soln. :

Step 1 : Scan Dfor count of each candidate. The candidate list is {I1,I2,I3,I4,I5} and find the support.

C_1=

I-Itemsets	Sup-count
I1	6
I2	7
I3	7
I4	2
I5	4

Step 2 : Compare candidate support count with minimum support count (i.e. 2).

L_1=

I-Itemsets	Sup-count
1	6
2	7
3	6
4	2
5	2

Scanned by CamScanner

Step 3 : Generate candidate C_2 from L_1 and find the support.

C_2=

2-Itemsets	Sup-count
1,2	4
1,3	4
1,4	1
1,5	3
2,3	4
2,4	2
2,5	2
3,4	0
3,5	3
4,5	0

Step 4 : Compare candidate (C_2) support count with the minimum support count.

L_2 =

2-Itemsets	Sup-count
1,2	4
1,3	4
1,5	3
2,3	4
2,4	2
2,5	2
3,5	3

Step 5 : Generate candidate C_3 from L_2.

C_3 =

Frequent 3-Itemset
1,2,3
1,2,5
1,2,4
1,3,5
2,3,5

Scanned by CamScanner

Step 6 : Scan D for count of each candidate in C_3.

C_3 =

Frequent 3-Itemset	Sup-count
1,2,3	2
1,2,5	2
1,2,4	0
1,3,5	2
2,3,5	0

Step 7 : Compare candidate (C_3) support count with the minimum support count.

L_3 =

Frequent 3-Itemset	Sup-count
1,2,3	2
1,2,5	2
1,3,5	2

Step 8 : So data contain the frequent itemsets are {I1,I2,I3} and {I1,I2,I5} and {I1,I3,I5}.

Let us assume that the data contains the frequent itemset = {I1,I2,I5} then the association rules that can be generated from frequent itemset are as shown below with the support and confidence.

Association Rule	Support	Confidence	Confidence %
I1^I2=>I5	2	2/4	50%
I1^I5=>I2	2	2/2	100%
I2^I5=>I1	2	2/2	100%
I1=>I2^I5	2	2/6	33%
I2=>I1^I5	2	2/7	29%
I5=>I1^I2	2	2/2	100%

If the minimum confidence threshold is 70% (Given), then only the SECOND, THIRD AND LAST rules above are output, since these are the only ones generated that are strong.

Similarly do for frequent itemset {I1,I2,I3} and {I1,I3,I5}.

Scanned by CamScanner

Ex. 5.4.7 : A Database has four transactions. Let Minimum support and confidence be 50%.

T_{Id}	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5
500	1,2,3
600	3,5
700	1,2,3,5
800	1,5
900	1,3

Soln. :

Step 1 : Scan D for count of each candidate. The candidate list is {1,2,3,4,5}and find the support.

$C_1 =$

Itemset	Sup-count
1	6
2	5
3	7
4	1
5	6

Step 2 : Compare candidate support count with minimum support count (i.e. 50%).

$L_1 =$

Itemset	Sup-count
1	6
2	5
3	7
5	6

Scanned by CamScanner

Step 3 : Generate candidate C_2 from L_1 and find the support.

$C_2 =$

Itemset	Sup-count
1,2	3
1,3	5
1,5	3
2,3	4
2,5	4
3,5	4

Step 4 : Compare candidate (C_2) support count with the minimum support count.

$L_2 =$

Itemset	Sup-count
1,3	5

So data contain the frequent itemset= {1,5}.

Therefore the association rule that can be generated from L_2 are as shown below with the support and confidence.

Association Rule	Support	Confidence	Confidence %
1=>3	5	5/6=0.83	83%
3=>1	5	5/7=0.71	71%

Given minimum confidence threshold is 50% , so both the rules are strong.

Final rules are :

Rule 1: 1=>3 and Rule 2 : 3=>1

Ex. 5.4.8 : Consider the five transactions given below. If minimum support is 30% and minimum confidence is 80%, determine the frequent itemsets and association rules using the a priori algorithm. **MU - Dec. 2012, 10 Marks**

Transaction	Items
T1	Bread, Jelly, Butter
T2	Bread, Butter
T3	Bread, Milk, Butter
T4	Coke, Bread
T5	Coke, Milk

Scanned by CamScanner

Soln. :

Step 1 : Scan D for Count of each candidate.

The candidate list is { Bread, Jelly, Butter, Milk, Coke }

$C_1 =$

I-Itemlist	Sup-Count
Bread	4
Jelly	1
Butter	3
Milk	2
Coke	2

Step 2 : Compare candidate support count with minimum support count (i.e. 2)

I-Itemlist	Sup-Count
Bread	4
Butter	3
Milk	2
Coke	2

Step 3 : Generate C2 from L1 and find the support

$C_2 =$

I-Itemlist	Sup Count
{Bread, Butter}	3
{Bread, Milk}	1
{Bread, Coke}	1
{Butter, Milk}	1
{Butter, Coke}	0
{Milk, Coke}	1

Step 4 : Compare candidate (C2) support count with the minimum support count

$L_2 =$

Frequent 2 - Itemset	Sup – Count
{Bread, Butter}	3

Step 5 : So data contain the frequent itemset is {Bread, Butter}

Association Rule	Support	Confidence	Confidence %
Bread ⟶ Butter	3	3/4	75%
Butter ⟶ Bread	1	3/3	100%

Scanned by CamScanner

Minimum confidence threshold is 80% (Given)

Final rule is

Butter → Bread

Ex. 5.4.9 : Consider the following transaction database.

TID	Items
01	A, B, C, D
02	A, B, C, D, E, G
03	A, C, G, H, K
04	B, C, D, E, K
05	D, E, F, H, L
06	A, B, C, D, L
07	B, I, E, K, L
08	A, B, D, E, K
09	A, E, F, H, L
10	B, C, D, F

Apply the Apriori algorithm with minimum support of 30% and minimum confidence of 70%, and find all the association rules in the data set.

MU - May 2013, May 2014, 10 Marks

Soln. :

Step 1 : Generate single item set :

Items	Support
A	6
B	7
C	6
D	7
E	6
F	3
G	2
H	3
I	1
K	4
L	4

Item set above 30 % support	
A	6
B	7
C	6
D	7
E	6
F	3
H	3
K	4
L	4

Scanned by CamScanner

Step 2 : Generate 2 item set :

Item	Support
AB	4
AC	4
AD	4
AE	3
AF	1
AH	2
AK	2
AL	2
BC	5
BD	6
BE	4
BF	1
BH	0
BK	3
BL	2
CD	5
CE	2
CF	1

Item	Support
CH	1
CK	2
CL	1
DE	4
DF	2
DH	1
DK	2
DL	2
EF	2
EH	2
EK	3
EL	3
FH	2
FK	0
FL	2
HK	1
HL	2
KL	1

Item set above 30 % support	
AB	4
AC	4
AD	4
AE	3
BC	5
BD	6
BE	4
BK	3
CD	5
DE	4
EK	3
EL	3

Step 3 : Generate 3 item set :

Item sets of 3 items

Item set	Support
ABC	3
ABD	4
ABE	2
ABK	1
ACD	3
ACE	1
ADE	2
AEK	1
AEL	1
BCD	5
BCE	2
BCK	1
BDE	3
BDK	2

Scanned by CamScanner

Item set	Support
BEK	2
BEL	1
CDE	2
DEK	2
DEL	1

Item set above 30 % support

Item set	Support
ABC	3
ABD	4
ACD	3
BCD	5
BDE	3

Step 4 : Generate 4 item set

Item set	Support
ABCD	3
ABDE	2
BCDE	2

Therefore ABCD is the large item set with minimum support 30%.

Following Rules generated

Rule	Confidence	Confidence %
$A \rightarrow BCD$	3/6 = 0.5	50%
$B \rightarrow ACD$	3/7 = 0.43	43%
$C \rightarrow ABD$	3/6 = 0.5	50%
$D \rightarrow ABC$	3/7 = 0.43	43%
$AB \rightarrow CD$	3/4 = 0.75	75%
$BC \rightarrow AD$	3/5 = 0.6	60%
$CD \rightarrow AB$	3/5 = 0.6	60%
$AC \rightarrow BD$	3/4 = 0.75	75%
$AD \rightarrow BC$	3/4 = 0.75	75%
$BCD \rightarrow A$	3/5 = 0.6	60%
$ACD \rightarrow B$	3/3 = 1	100%
$ABD \rightarrow C$	3/4 = 0.75	75%
$ABC \rightarrow D$	3/3 = 1	100%

Scanned by CamScanner

From the above Rules generated, only the rules having greater than 70% are considered as final rules. So final Rules are,

$$\boxed{\begin{array}{l} AB \rightarrow CD \\ AC \rightarrow BD \\ AD \rightarrow BC \\ ACD \rightarrow B \\ ABD \rightarrow C \\ ABC \rightarrow D \end{array}}$$

Syllabus Topic : Association Rule Generation

5.5 Association Rule Generation

Association rules are generated by finding the support and confidence of each rule as given in section **5.2.3**

Syllabus Topic : Improving the Efficiency of Apriori

5.6 Improving the Efficiency of Apriori

There are many variations of Apriori algorithm that have been proposed to improve the efficiency, few of them are given as :

- **Hash-based itemset counting :** The itemsets can be hashed into corresponding buckets. For a particular iteration a k-itemset can be generated and hashed into their respective bucket and increase the bucket count, the bucket with a count lesser than the support should not be considered as a candidate set.
- **Transaction reduction :** A transaction that does not contain k-frequent itemset will never have k+1 frequent itemset, such a transaction should be reduced from future scans.
- **Partitioning :** In this technique only two database scans are needed to mine the frequent itemsets. The algorithm has two phases, in the first phase, the transaction database is divided into non overlapping partitions. The minimum support count of a partition is min support X number of transactions in that partition. Local frequent itemsets are found out in each partition. The local frequent itemsets may or may not be frequent with respect to the entire database however a frequent itemset from database has to be frequent in atleast one of the partitions.

Scanned by CamScanner

- All the frequent itemsets with respect to each partition forms the global candidate itemsets. In the second phase of the algorithm, a second scan of database for actual support of each item is found, these are global frequent itemsets.
- **Sampling :** Rather than finding the frequent itemsets in the entire database D, a subset of transactions are picked up and searched for frequent itemsets. A lower threshold of minimum support is considered as this reduces the possibility of missing the actual frequent itemset due to a higher support count
- **Dynamic itemset counting :** In this the database is partitioned into blocks and is marked by start points. It maintains a count-so-far, if this count-so-far crosses minimum support, the itemset is added to the frequent itemset collection which can be further used to generate longer candidate itemset.

Syllabus Topic : FP Growth

5.7 FP Growth

➔ **(MU - May 2016)**

A Pattern Growth Approach for Mining Frequent Itemsets(FP-Growth).

5.7.1 Definition of FP-tree

An FP-tree is a tree structure which consists of :

- One root labeled as "null".
- A set of item prefix sub-trees with each node formed by three fields : item-name, count, node-link.
- A frequent-item header table with two fields for each entry : item-name, head of node-link.
- It contains the complete information for frequent pattern mining.
- The size of the FP-tree is bounded by the size of the database, but due to frequent items sharing, the size of the tree is usually much smaller than its original database.
- High compaction is achieved by placing more frequently items closer to the root (being thus more likely to be shared).
- The FP-Tree contains everything from the database we need to know for mining frequent.

Scanned by CamScanner

tion forms the global candidate
ond scan of database for actual
emsets.

he entire database D, a subset of
itemsets. A lower threshold of
ssibility of missing the actual

oned into blocks and is marked
-far crosses minimum support,
which can be further used to

➔ **(MU - May 2016)**

Growth).

fields : item-name, count,

y : item-name, head of

ut due to frequent items
l database.

oser to the root (being

y for mining frequent.

Patterns

- The size of the FP-tree is ≤ the candidate sets generated in the association rule mining.
- This approach is very efficient due to :
  - Compression of a large database into a smaller data structure.
  - It is a frequent pattern growth mining method or simply FP-growth.
  - It adopts a divide-and-conquer strategy.
- The database of frequent items is compressed into a FP-Tree, and the association information of items is preserved.
- Then mine each such database separately.

5.7.2 FP-Tree Algorithm

FP-tree construction algorithm given by Jiawei Han et al.

- FP-Growth: allows frequent itemset discovery without candidate itemset generation.
- Once the FP tree is generated, it is mined by calling FP_growth(FP_tree,null).

Algorithm : FP growth, Mine frequent itemsets using an FP-tree by pattern fragment growth.

Input :

- D, a transaction database.
- min_sup, the minimum support count threshold.

Output : The complete set of frequent patterns.

Method :

1. **A FP tree is constructed in the following steps**

   (a) Scan the transaction database D once, Collect F, the set of frequent items, and their support counts. Sort F by support count in descending order as L, the list of frequent items.

   (b) Create the root of an FP tree, and label it as "null". For each transaction Trans D do the following :

Scanned by CamScanner

Select and sort the frequent items in Trans according to the order of L. Let the sorted frequent item list in Trans be [p | P], where p is the first element and P is the remaining list. Call insert_tree ([p | P]·T), which is performed as follows. If T has a child N such that N_item·name = p·item·name, then increment N's count by 1; else create a new node N, and let its count be 1, its parent link be linked to T, and its node-link to the nodes with the same item·name via the node·link structure. If P is nonempty, call insert_tree (P, N) recursively.

2. **The FP-tree is mined by calling FP growth. FP tree, null/, which is implemented as follows :**

Procedure FP_growth (Tree, α) :

```
(1) if Tree contains a single path P then
(2) for each combination (denoted as β) of the nodes in the path P
(3) generate pattern β∪α with support_count = minimum support count of nodes in β;
(4) else for each a_i in the header of Tree {
(5) generate pattern β = a_i∪ αwith support_count = a_i.support_count;
(6) construct β's conditional pattern base and then β's conditional FP_tree Treeβ;
(7) if Treeβ≠ϕ then
(8) call FP_growth (Treeβ, β) ; }
```

Analysis

- Two scans of the DB are necessary. The first collects the set of frequent items and the second constructs the FP-tree.
- The cost of inserting a transaction Trans into the FP-tree is O(|Trans|), where |Trans| is the number of frequent items in Trans.

5.7.3 FP-Tree Size

- Many transactions share items due to which the size of the FP-Tree can have a smaller size compared to uncompressed data.
- **Best case scenario :** All transactions have the same set of items which results in a single path in the FP Tree.

Scanned by CamScanner

- **Worst case scenario :** Every transaction has a distinct set of items, i.e. no common items.
  - FP-tree size is as large as the original data.
  - FP-Tree storage is also higher, it needs to store the pointers between the nodes and the counter.
- FP-Tree size is dependent on the order of the items. Ordering of items by decreasing support will not always result in a smaller FP -Tree size (it's heuristic).

5.7.4 Example of FP Tree

Ex. 5.7.1 : Transactions consist of a set of items I = {a, b, c, ...} , min support = 3.

TID	Items Bought
1	f, a, c, d, g, i, m, p
2	a, b, c, f, l, m, o
3	b, f, h, j, o
4	b, c, k, s, p
5	a, f, c, e, l, p, m, n

Soln. :

Step 1 : Find the minimum support of each item.

Item	Sup.
A	3
B	3
C	4
D	1
E	1
F	4
G	1
H	1
I	1
J	1
K	1
L	2
M	3
N	1
O	2
P	3

Scanned by CamScanner

Consider items with min support = 3 (given)

Item	Sup.
A	3
B	3
C	4
F	4
M	3
P	3

Step 2 : Order all items in itemset in frequency descending order (min support=3)
(Note : Consider only items with min support = 3)

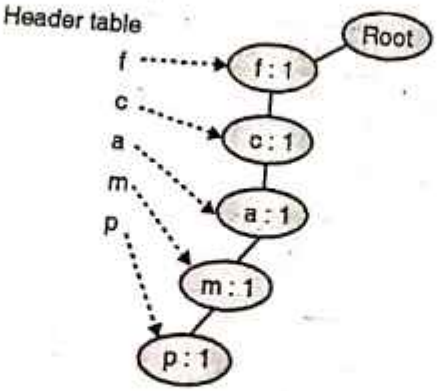
TID	Items Bought	(Ordered frequent items)
1	f, a, c, d, g, i, m, p	f, c, a, m, p
2	a, b, c, f, l, m, o	f, c, a, b, m
3	b, f, h, j, o	f, b
4	b, c, k, s, p	c, b, p
5	a, f, c, e, l, p, m, n	f, c, a, m, p

(f:4, c:4, a:3, b:3, m:3, p:3)

Step 3 : FP Tree construction

Originally Empty

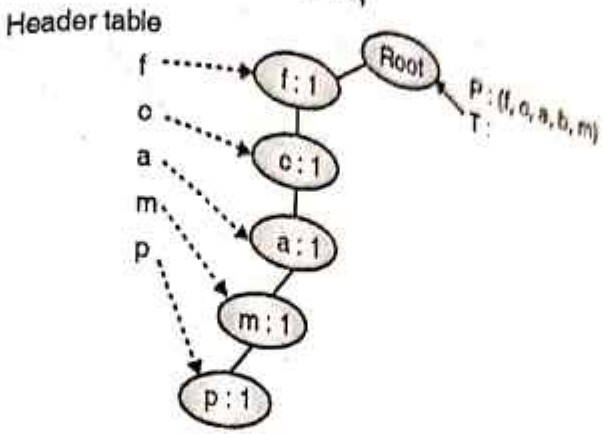
Step 4 : Insert the first Transaction (f, c, a, m, p)



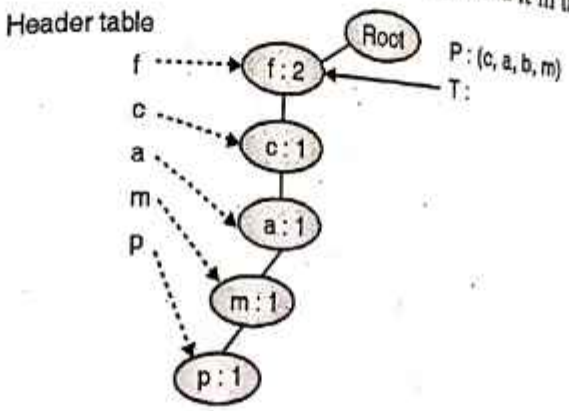
Scanned by CamScanner

Step 5 : Start the insertion of Second transaction (f,c,a,b,m)

(i) The transaction T is pointing to the root node,

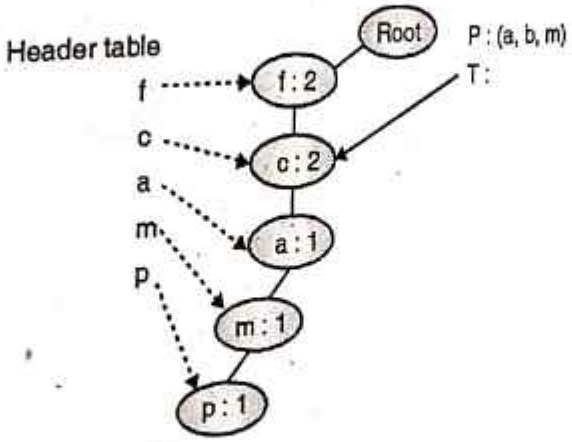


(ii) Consider the first item in the second transaction i.e. f and add it in the tree.



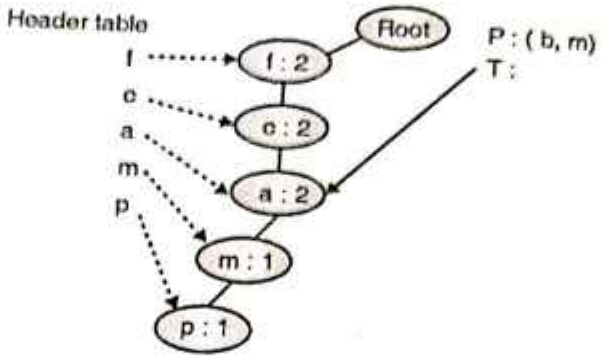
After this step we get f:2, finished adding f in the above tree.

(iii) Now consider the second item in the above transaction .i.e. c.

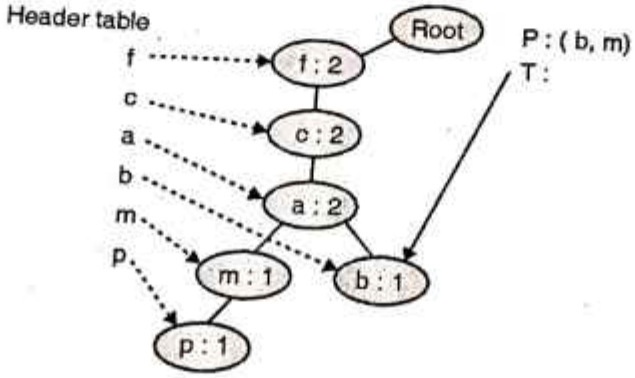


Scanned by CamScanner

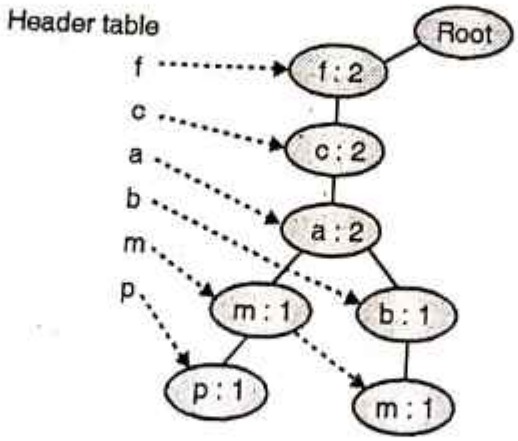
(iv) Similarly consider the next item a.



(v) Since we do not have a node b, we create one node for b below the node a (note : to maintain the path).



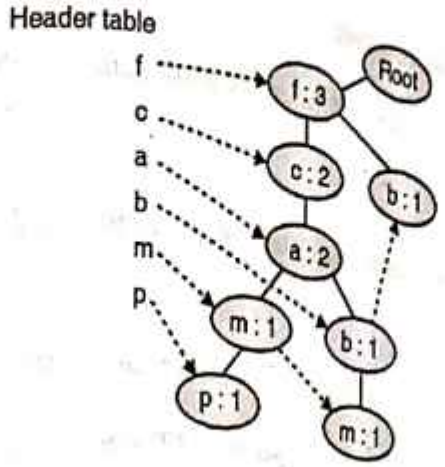
(vi) Now only m of second transaction is left. Though a node m is already exists still we can't increase its count of the existing node m as we need to represent the second transaction in FP tree, so add new node m below node b and link it with existing node m.



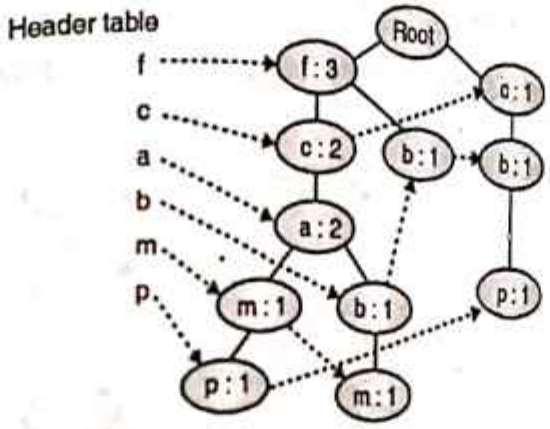
Second transaction is complete.

Scanned by CamScanner

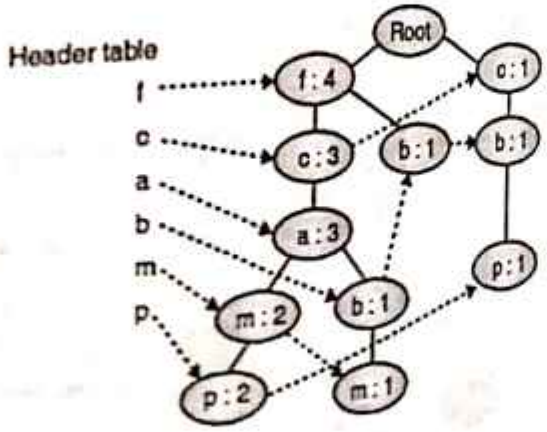
Step 6 : Similarly insert the third transaction (f, b) as explained in step 5. So After the insertion of third transaction (f, b)



Step 7 : After the insertion of fourth transaction (c, b, p)



Step 8 : After the insertion of fifth Transaction (f, c, a, m, p)



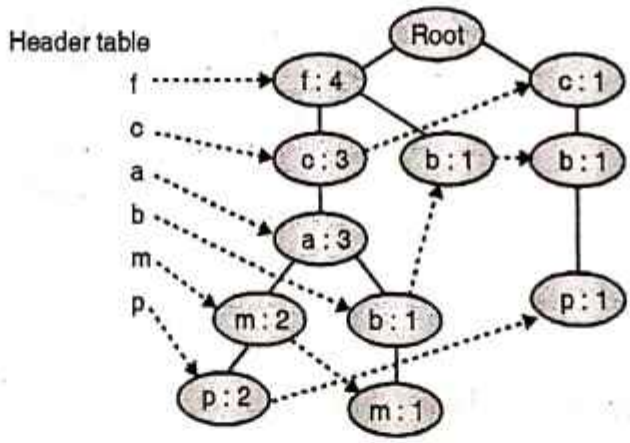
This is the final FP-Tree.

Scanned by CamScanner

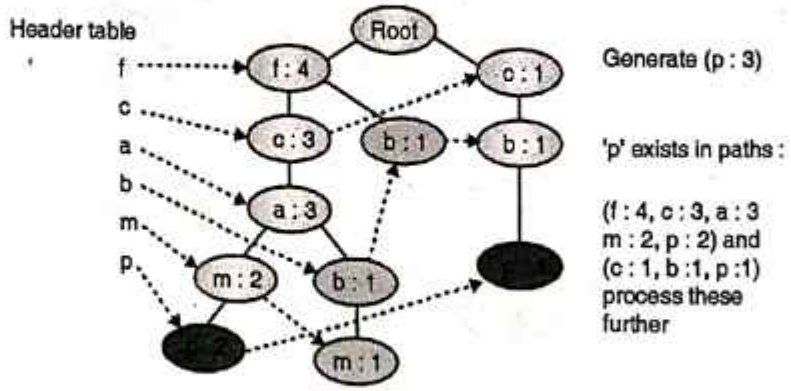
5.7.5 Mining Frequent Patterns from FP Tree

- **General idea (divide-and-conquer)**
  - Use the FP Tree and recursively grow frequent pattern path.
- **Method**
  - For each item, conditional pattern-base is constructed, and then it's conditional FP-tree.
  - On each newly created conditional FP-tree, repeat the process.
  - The process is repeated until the resulting FP-tree is empty, or it has only a single path(All the combinations of sub paths will be generated through that single path, each of which is a frequent pattern).

Example : Finding all the patterns with 'p' in the FP tree given below



- Starting from the bottom of the header table.



Scanned by CamScanner

Following are the paths with 'P'

- We got (f:4, c:3, a:3, m:2, p:2) and (c:1, b:1, p:1)
- The transactions containing 'p' have p.count
- Therefore we have (f:2, c:2, a:2, m:2, p:2) and (c:1, b:1, p:1)
- Since 'p' is part of these we can remove 'p'

Conditional Pattern Base (CPB)

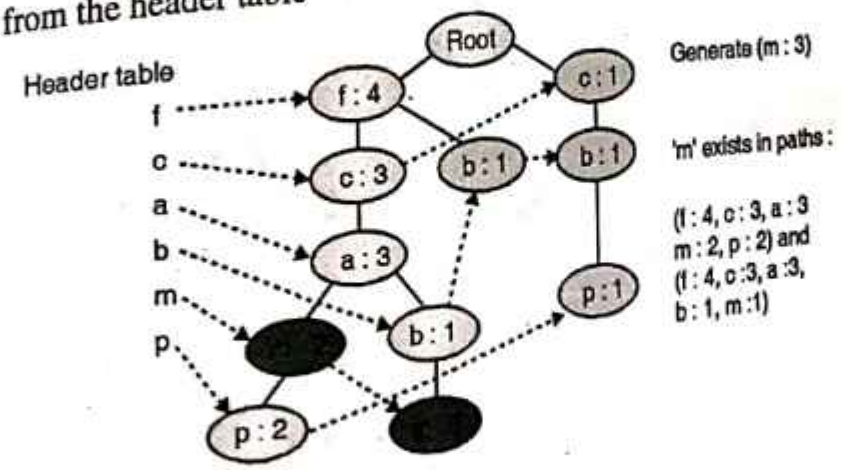
- After removing P we get : (f:2, c:2, a:2, m:2) and (c:1, b:1)
- Find all frequent patterns in the CPB and add 'p' to them, this will give us all frequent patterns containing 'p'.
- This can be done by constructing a new FP-Tree for the CPB.

Finding all patterns with 'P'.

- We again filter away all items < minimum support threshold (i.e. 3)
- (f:2, c:2, a:2, m:2), (c:1, b:1) => (c:3)
- We generate (cp:3) (Note : we are finding frequent patterns containing item p, so we append p to c as c is only item that has min support threshold.)
- Support value is taken from the sub-tree
- Frequent patterns thus far: (p:3, cp:3)

Example : Finding Patterns with 'm' but not 'p'.

Find 'm' from the header table

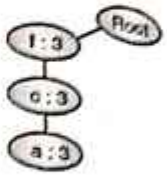


Conditional Pattern Base:

- Path 1 : (f:4, c:3, a:3, m:2, p:2) → (f:2, c:2, a:2)

Scanned by CamScanner

- In the above transaction we need to consider m:2, based on this we get f:2 and so on. Exclude p as we don't want p i.e. given in example.
- Path 2 : (f:4, c:3, a:3, b:1, m:1) → (f:1, c:1, a:1, b:1)



- Build FP tree using (f:2, c:2, a:2) and (f:1, c:1, a:1, b:1)
- Now we got (f:3, c:3, a:3 ,b:1)
- Initial Filtering removes b:1 (We again filter away all items < minimum support threshold).
- Mining Frequent Patterns by Creating Conditional Pattern-Bases.

Item	Conditional pattern-base	Conditional FP-tree
P	{(fcam:2), (cb:1)}	{(c:3)}|p
M	{(fca:2), (fcab:1)}	{(f:3, c:3, a:3)}|m
B	{(fca:1), (f:1), (c:1)}	Empty
A	{(fc:3)}	{(f:3, c:3)}|a
C	{(f:3)}	{(f:3)}|c
f	Empty	Empty

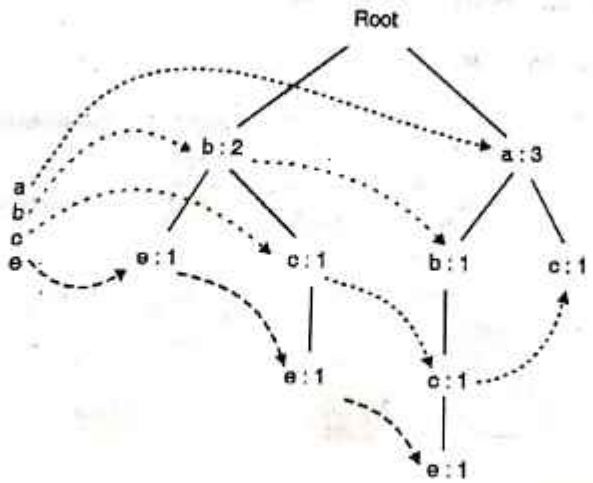
Ex. 5.7.2 : Transaction item list is given below. Draw FP tree.

T1 = b, e T2 = a, b, c, e

T3 = b, c, e T4 = a, c T5 = a

Given : minimum support = 2

Soln. :



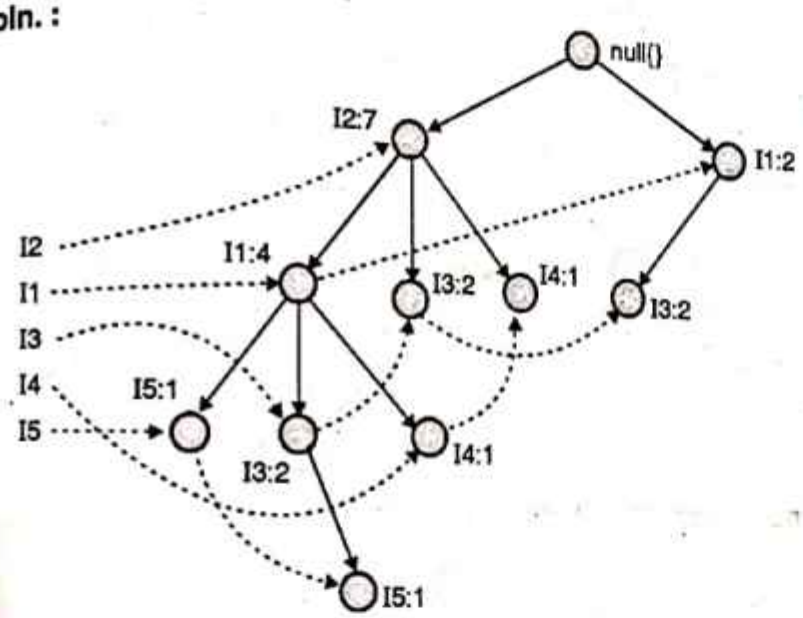
Scanned by CamScanner

Ex. 5.7.3 : Transaction database is

TID	List of item_Ids
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Min support = 2

Soln. :



Item ID	Support Count
I2	7
I1	6
I3	6
I4	2
I5	2

Mining the FP-Tree by creating conditional (sub) pattern bases.

Item	Conditional pattern base	Conditional FP-tree	Frequent patterns generated
15	{(I2 I1 : 1),(I2 I1 I3 : 1) }	{I2 : 2, I1 : 2}	I2 I5 : 2, I1 I5 : 2, I2 I1 I5 : 2
14	{(I2 I1 : 1),(I2 : 1)}	{I2 : 2}	I2 I4 : 2
13	{(I2 I1 : 2),(I2 : 2), (I1 : 2) }	(I2 : 4, I1 : 2), (I1 : 2)	I2 I3 : 4, I1, I3 : 2, I2 I1 I3 : 2
11	{(I2 : 4)}	{(I2 : 4)}	I2 I1 : 4

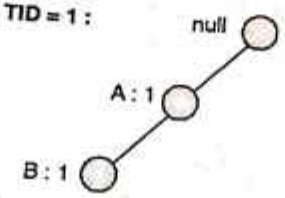
Ex. 5.7.4 : Consider the following dataset of frequent itemsets. All are sorted according to their support count. Construct the FP-Tree and find Conditional Pattern base for D.

Scanned by CamScanner

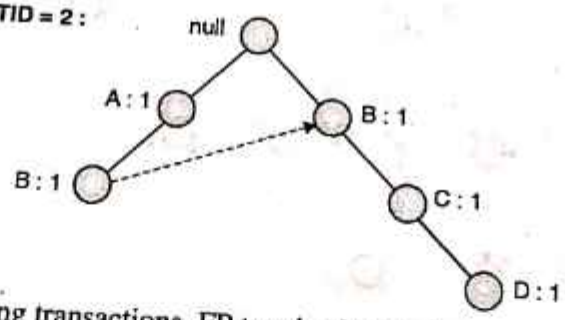
TID	Items
1	{A, B}
2	{B, C, D}
3	{A, C, D, E}
4	{A, D, E}
5	{A, B, C}
6	{A, B, C, D}
7	{B, C}
8	{A, B, C}
9	{A, B, D}
10	{B, C, E}

Soln. :

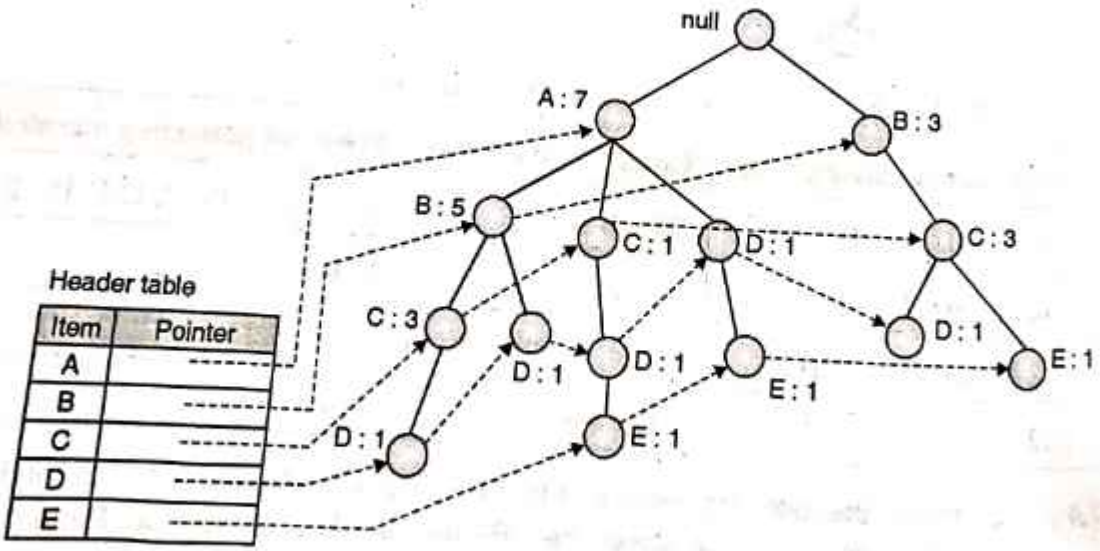
After reading TID = 1 :



After reading TID = 2 :



Similarly for all the remaining transactions, FP tree is given below.



Scanned by CamScanner

Conditional Pattern base for D

P = {(A:1,B:1,C:1),
(A:1,B:1),
(A:1,C:1),
(A:1),
(B:1,C:1)}

- We have the following paths with 'D'

  P = {(A:1,B:1,C:1),(A:1,B:1), (A:1,C:1), (A:1)} and {(B:1,C:1)}
- Support count of D =1.
- Conditional Pattern Base (CPB)
  - To find all frequent patterns containing 'D' we need to find all frequent patterns in the CPB and add 'D' to them
  - We can do this by constructing a new FP-Tree for the CPB
- Finding all patterns with 'D'
  - Again filter away all items < minimum support threshold

    (i.e.1 as Support of D =1)
  - Consider First Branch

    {(A:1,B:1,C:1),(A:1,B:1), (A:1,C:1), (A:1)} => {(A:4,B:2,C:2)}

    So append ABC with D

    We generate ABCD:1
  - Similarly for other branch of the tree {(B:1,C:1)} => {(B:1,C:1)}

    So append BC with D

    We generate BCD : 1
  - Recursively apply FP-growth
  - So Frequent Itemsets found (with sup > 1): AD, BD, CD, ACD, BCD which are generated from CPB on conditional node D.

Scanned by CamScanner

5.7.6 Benefits of the FP-Tree Structure

Completeness

- The Long pattern of any transaction is never broken.
- For frequent pattern mining complete information is preserved.
- The method can mine short as well as long frequent patterns and it is highly efficient.
- FP-Growth algorithm is much faster than Apriori Algorithm.
- The search cost is reduced.

Syllabus Topic : Mining Frequent Itemsets using Vertical Data Formats

5.8 Mining Frequent Itemsets using Vertical Data Formats

- There are usually two ways of representing transactional data, Horizontal Data format and Vertical data format.
- In Horizontal data format the transactional data is represented as TID-itemset where TID is the transaction id and itemset is the set of items bought in that particular transaction.

E.g. of Horizontal data Format.

Transaction ID	Items Bought
T100	1,2,3
T200	1,3
T300	1,4
T400	2,3

- In Vertical data format, transactional data is represented as item-TID-set. Item is the item name and TID-set is the set of transactions containing that item.
  - E.g. of Vertical Data Format

Items Bought	Transaction ID set
1	T100, T200, T300
2	T100, T400
3	T100,T200, T400
4	T300

Scanned by CamScanner

- 2-itemset in vertical data format

Items Bought	Transaction ID set
12	T100
13	T100, T200
14	T300
23	T100, T400

- 3-itemset in vertical data format

Items Bought	Transaction ID set
123	T100

Therefore there is only one frequent 3 -itemset {1,2,3}.

- The above process is repeated by the intersection of TID_sets of the frequent k-itemsets to compute the TID_sets of the corresponding (k+1) itemsets. The process is stopped when until no frequent itemsets or candidate itemsets can be found.

Syllabus Topic : Introduction to Mining Multilevel Association Rules

5.9 Introduction to Mining Multilevel Association Rules

(MU - May 2012, May 2013, Dec. 2013, Dec. 2014, May 2016)

- Items are always in the form of hierarchy.
- Items which are at leaf nodes are having lower support.
- An item can be either generalized or specialized as per the described hierarchy of that item and its levels can be powerfully preset in transactions.
- Rules which combine associations with hierarchy of concepts are called Multilevel Association Rules.

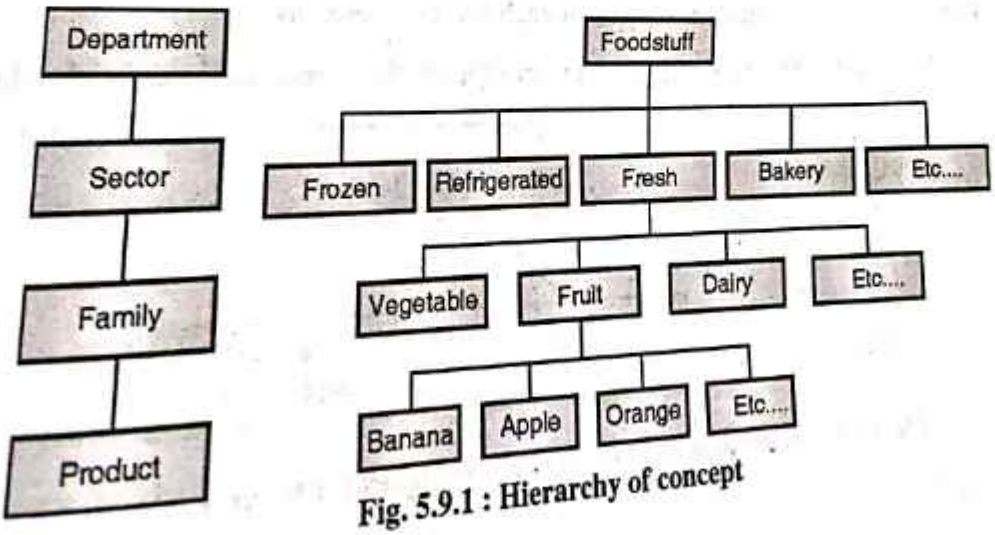


Fig. 5.9.1 : Hierarchy of concept

Scanned by CamScanner

Support and confidence of multilevel association rules

- The support and confidence of an item is affected due to its generalization or specialization value of attributes.
- The support of generalized item is more than the support of specialized item.
- Similarly the support of rules increases from specialized to generalized itemsets.
- If the support is below the threshold value then that rule becomes invalid.
- Confidence is not affected for general or specialized.

Two approaches of multilevel association rule

1. **Using uniform minimum support for all levels**
   - Consider the same minimum support for all levels of hierarchy.
   - As only one minimum support is set, so there is no necessity to examine the items of itemset whose ancestors do not have minimum support.
   - If very high support is considered then many low level association may get missed.
   - If very low support is considered then many high level association rules are generated.

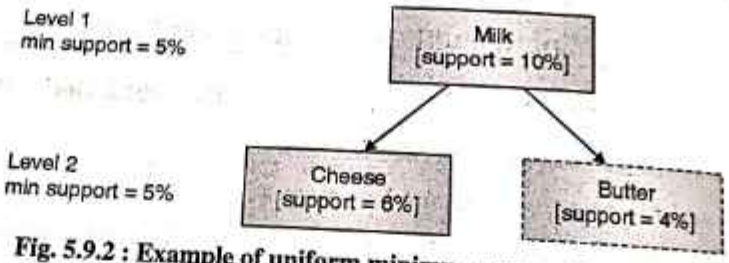


Fig. 5.9.2 : Example of uniform minimum support for all levels

2. **Using reduced minimum support at lower level**
   - Consider separate minimum support at each level of hierarchy.
   - As every level is having its own minimum support, the support at lower level reduces.

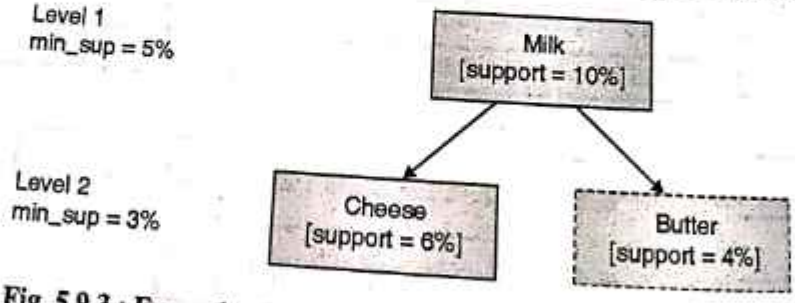


Fig. 5.9.3 : Example of reduced minimum support for lower levels

Scanned by CamScanner

There are 4 search strategies :

(i) **Level-by-level independent**

- Its a full-breadth search method.
- The parent node is checked whether it's frequent or not frequent and based on that node is examined.

(ii) **Level-cross filtering by single item**

The children of only frequent nodes are checked.

(iii) **Level-cross filtering by k-itemset**

- Find the frequent k itemset at the parent level
- Only the k itemset at next level is checked.

(iv) **Controlled level-cross filtering by single item**

- This is the modified version of Level-cross filtering by single item.
- Some minimum support threshold is set for lower level.
- So the items which do not satisfy minimum support are checked for minimum support threshold this is also called "Level Passage Threshold".

Syllabus Topic : Multidimensional Association Rules

5.10 Mining Multidimensional (MD) Association Rules

→ **(MU - May 2013, Dec. 2014, May 2016)**

- **Single-dimensional rules :** The rule contains only one distinct predicate. In the following example the rule has only one predicate "buys".

  buys(X, "Butter") ⇒ buys(X, "Milk")

- **Multi-dimensional rules :** The rule contains two or more dimensions or predicates.
  - **Inter-dimension association rules :** The rule doesn't have any repeated predicate

    gender(X,"Male") ∧ salary(X, "High") ⇒ buys(X, "Computer")

  - **Hybrid-dimension association rules :** The rule have many occurrences of same predicate i.e. buys.

    gender(X,"Male") ∧ buys(X, "TV") ⇒ buys(X, "DVD")

Scanned by CamScanner

- **Categorical attributes :** This have finite number of possible values and there is no ordering among values. Example : brand, color.
- **Quantitative attributes :** These are numeric values and there is implicit ordering among values. Example : age, income.

Techniques for Mining MD Associations

1. Using static discretization of quantitative attributes

- Using concept hierarchy, discretize the quantitative attributes.
- Convert the numeric values by ranges or categorical values.
- To get the all frequent k-predicate , k or k+1 table scans are required and generate a cuboid which is more suited for mining
- Mining is always more faster on data cube.
- Predicate sets are determined by the cells of n- dimensional cuboid.
- In the example below, the base cuboid contains the three predicates gender, salary and buys. Each cuboid represents a different group-by.

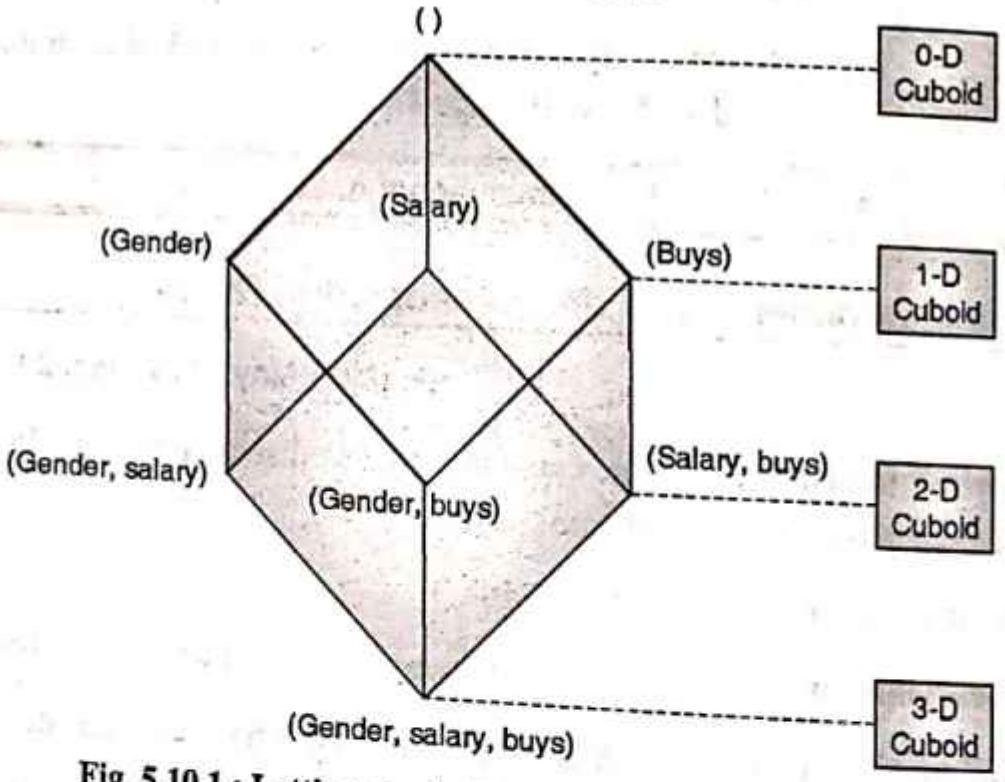


Fig. 5.10.1 : Lattice of cuboids making up a 3-D data cube

2. Quantitative association rules

- Numeric attributes are dynamically discretized such that the confidence or compactness of the rules mined is maximized.

Scanned by CamScanner

- In quantitative association rule, the left hand side of rule contains 2-D quantitative attributes and right hand side of rule contains one categorical attribute.

$$A_{quantitative1} \wedge A_{quantitative2} \Rightarrow C_{ategorical}$$

- Example : If we are interested in association rule where two quantitative measures are age and salary and the type of the phone that customer buy.

  Age(CUST,"30-34") ∧ Salary (CUST,"24K – 48K") ⇒ buys(CUST, "Sony Xperia Z")

- The approach used to find such rules is the Association Rule Clustering System (ARCS).
- Steps involved in ARCS are :

(a) **Binning :** Partition the ranges of quantitative attributes into intervals. These intervals are considered as bins. ARCS are equiwidth binning method where each bin has same interval size.

(b) **Finding frequent predicate set :** It finds the frequent predicate sets which satisfy the minimum support and also minimum confidence. Rule Generation algorithm is used to generate strong association rules.

(c) **Clustering the association rule :** Strong association rules obtained from above step is mapped to a 2-D grid as shown in Fig. 5.10.2 :

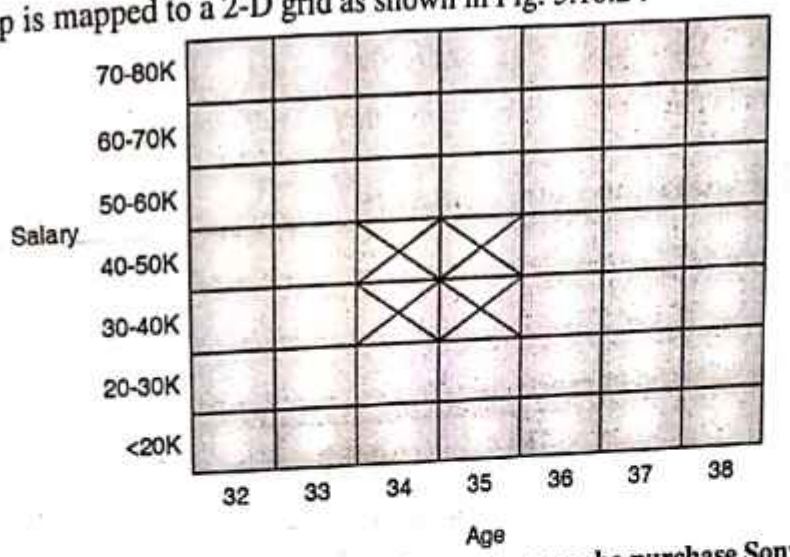


Fig. 5.10.2 : A 2-D grid for tuples representing customers who purchase Sony Xperia Z

- Following four CUSTs correspond to the rules :
  - age(CUST,34) ∧Salary(CUST,"30 – 40K") → buys(CUST," Sony Xperia Z")
  - age(CUST,35) ∧Salary (CUST,"30 – 40K") → buys(CUST," Sony Xperia Z")
  - age(CUST,34) ∧Salary (CUST,"40 – 50K") → buys(CUST," Sony Xperia Z")
  - age(CUST,35) ∧Salary (CUST,"40 – 50K") → buys(CUST," Sony Xperia Z")

Scanned by CamScanner

- As above rules are close to each other, they can be clustered together to form the following rule :
  - age(CUST, "34 – 35") ∧Salary (CUST,"30 – 50K") → buys(CUST," Sony Xperia Z")

Limitations of ARCS

- It works only for quantitative attributes on LHS of rules.
- The limitation is 2D i.e. two attributes on LHS only so doesn't work for more dimensions.

3. Distance-based association rules

- This is a dynamic discretization process that considers the distance between data points.
- This mining process has only two steps :
  - Perform clustering to find the interval of attributes involved.
  - Obtain association rules by searching for groups of clusters that occur together.
- The resultant rules of this method must satisfy :
  - Clusters in the rule antecedent are strongly associated with clusters of rules in the consequent.
  - Clusters in the antecedent occur together.
  - Clusters in the consequent occur together.

5.11 University Questions and Answers

May 2010

Q. 1 Consider the following transactions :

TID	Items
01	1, 3, 4, 6
02	2, 3, 5, 7
03	1, 2, 3,5, 8
04	2, 5, 9, 10
05	1,4

Apply the Apriori Algorithm with minimum support of 30% and minimum confidence of 75% and find the large item set L. *(Ans. : Refer Ex. 5.4.4)* **(10 Marks)**

Scanned by CamScanner

clustered together to form the

50K") → buys(CUST," Sony

t work for more dimensions.

the distance between data

lved.

ers that occur together.

h clusters of rules in the

m confidence of

(10 Marks)

Dec. 2010

Q.2 Consider the transaction database given below. Use Apriori Algorithm with minimum support count 2. Generate the association rules along with its confidence : *(Ans. : Refer Ex. 5.4.3)* **(10 Marks)**

TID	List of Items
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

May 2011

Q.3 What is Association Rule Mining ? Give the Apriori algorithm. Apply AR Mining to find all frequent itemsets from the following table : **(8 Marks)**

Transaction - ID	Items
100	1, 2, 5
200	2, 4
300	2, 3
400	1, 2, 4
500	1, 3
600	2, 3
700	1, 3, 2, 5
800	1, 3
900	1, 2, 3

Minimum Support Count = 2

Minimum – Confidence = 70%. *(Ans. : Refer Ex. 5.4.3)*

Scanned by CamScanner

Dec. 2011

Q. 4 Explain how Apriori algorithm is useful in identifying frequent item set ?

(Ans. : Refer sections 5.2.3 and 5.4.1) **(10 Marks)**

May 2012

Q. 5 Explain what is meant by association rule mining. For the table given below perform apriori algorithm. Also :

(i) Determine the k-item sets (frequent) obtained.

(ii) Justify the strong association rule that has been determined i.e. specify which is the strongest rule obtained.

The table is as follows :

TID	Items
01	1, 3, 4, 6
02	2, 3, 5, 7
03	1, 2, 3, 5, 8
04	2, 5, 9, 10
05	1, 4

Assume Minimum support of 30% and Minimum confidence of 75%.

(Ans. : Refer Ex. 5.4.4) **(10 Marks)**

Q. 6 What is meant by market-basket analysis? Explain with an example. State and explain with formula the meaning of the terms :

(i) Support (ii) Confidence

Hence explain how to mine multi level association rules from transaction databases, with example for each.

(Ans. : Refer sections 5.1, 5.2.3 and 5.9) **(10 Marks)**

Dec. 2012

Q. 7 Consider the 5 transactions given below. If minimum support is 30% and minimum confidence is 80%, determine the frequent item sets and association rules using the a priori algorithm. *(Ans. : Refer Ex. 5.4.8)* **(10 Marks)**

Scanned by CamScanner

Transaction	Items
T1	Bread, Jelly, Butter
T2	Bread, Butter
T3	Bread, Milk, Butter
T4	Coke, Bread
T5	Coke, Milk

May 2013

Consider the following transaction database:

TID	Items
01	A, B, C, D,
02	A, B, C, D, E, G
03	A, C, G, H, K
04	B, C, D, E, K
05	D, E, F, H, L
06	A, B, C, D, L
07	B, I, E, K, L
08	A, B, D, E, K
09	A, E, F, H, L
10	B, C, D, F

Apply the Apriori algorithm with minimum support of 30% and minimum confidence of 70%, and find all the association rules in the data set.
(Ans. : Refer Ex. 5.4.9) **(10 Marks)**

Define multidimensional and multilevel association mining.
(Ans. : Refer sections 5.9 and 5.10) **(10 Marks)**

Dec. 2013

What is meant by market,- basket analysis? Explain with an example. State and explain with formula the meaning of following terms

(i) Support (ii) Confidence

Hence explain how to mine multilevel Association rules from transaction databases, with examples. *(Ans. : Refer sections 5.1, 5.2.3 and 5.9)* **(10 Marks)**

Scanned by CamScanner

Q. 11 Consider the following transactions :

TID	Items
01	1, 3, 4, 6
02	2, 3, 5, 7
03	1, 2, 3, 5, 8
04	2, 5, 9, 10
05	1, 4

Apply the Apriori Algorithm with minimum support of 30 % and minimum confidence of 75 and find the large item set L. *(Ans. : Refer Ex. 5.4.4)* **(10 Marks)**

May 2014

Q. 12 Consider the following transaction database :

TID	Items
01	A, B, C, D
02	A, B, C, D, E, G
03	A, C, G, H, K
04	B, C, D, E, K
05	D, E, F, H, L
06	A, B, C, D, L
07	B, I, E, K, L
08	A, B, D, E, K
09	A, E, F, H, L
10	B, C, D, F

Apply the **Apriori** algorithm with minimum support of 30% and minimum confidence of 70% and find all the association rules in the data set.
(Ans. : Refer Ex. 5.4.9) **(10 Marks)**

Dec. 2014

Q. 13 Define multidimensional and multilevel association mining.
(Ans. : Refer sections 5.9 and 5.10) **(10 Marks)**

May 2016

Q. 14 Discuss association rule mining and apriori algorithm. Apply AR mining to find all frequent item sets and association rules for the following dataset :

Minimum support count = 2

Minimum confidence = 70 % *(Ans. : Refer Ex. 5.4.3)* **(10 Marks)**

Scanned by CamScanner

Transaction_ID	Items
100	1,2,5
200	2,4
300	2,3
400	1,2,4
500	1,3
600	1,3
700	1,3,2,5
800	1, 3
900	1, 2, 3

Q. 15 Write short note on : FP tree *(Ans. : Refer section 5.7)* **(5 Marks)**

Q. 16 Write short note on : Multilevel and multidimensional association rule *(Ans. : Refer sections 5.9 and 5.10)* **(5 Marks)**

□□□

Chapter Ends

Scanned by CamScanner

CHAPTER 6

Module 6

Spatial and Web Mining

Syllabus :

Spatial Data, Spatial Vs. Classical Data Mining, Spatial Data Structures, Mining Spatial Association and Co-location Patterns, Spatial Clustering Techniques : CLARANS Extension, Web Mining : Web Content Mining, Web Structure Mining, Web Usage mining, Applications of Web Mining.

Syllabus Topic : Spatial Data

6.1 Spatial Data

- Spatial data refers to all types of data objects or elements that are present in a geographical space or horizon. It supports the global finding and locating of individuals or devices anywhere in the world.
- Spatial data is also known as geospatial data, spatial information or geographic information.

Spatial Data Mining

Spatial data mining is the process of discovering interesting, useful, non-trivial patterns from large spatial datasets.

Non-trivial Search

- Large (e.g. exponential) search space of reasonable hypothesis.
- Example : Asiatic cholera : causes : water, food, air, insects, ...; water delivery mechanisms - numerous pumps, rivers, ponds, wells, pipes, ...

Scanned by CamScanner

Module 6

Web Mining

Data Structures, Mining Spatial
ring Techniques : CLARANS
Structure Mining, Web Usage

nts that are present in a
d locating of individuals or

ormation or geographic

ul, non-trivial patterns

..; water delivery

Interesting

- Useful in certain application domain.
- Example : Shutting off identified Water pump => saved human life.

Unexpected

- Pattern is not common knowledge.
- May provide a new understanding of world.
- Example : Water pump - Cholera connection lead to the "germ" theory.

6.1.1 Spatial Pattern

Pattern showing the interaction of two or more spatial objects or space-depending attributes according to a particular spacing or set of arrangements.

6.1.2 What is NOT Spatial Data Mining ?

- Simple Querying of Spatial Data
  - Find neighbors of Canada given names and boundaries of all countries.
  - Find shortest path from Mumbai to Delhi in a freeway map.
  - Search space is not large (not exponential).
- Testing a hypothesis via a primary data analysis
  - Ex. Female chimpanzee territories are smaller than male territories.
  - Search space is not large.
  - SDM : Secondary data analysis to generate multiple probable hypotheses.
- Uninteresting or obvious patterns in spatial data
  - Heavy rainfall in Minneapolis is correlated with heavy rainfall in St. Paul, given that the two cities are 10 miles apart.
  - Common knowledge : Nearby places have similar rainfall.

6.1.3 Why Learn about Spatial Data Mining ?

- New understanding of geographic processes for Critical questions
  - Example : How is the health of planet Earth?
  - Example : Characterize effects of human activity on environment and ecology.

Scanned by CamScanner

- Example : Predict effect of weather, and economy.

- Traditional approach : Manually generate and test hypothesis but, spatial data is growing too fast to analyze manually.
  - Satellite imagery, GPS tracks, sensors on highways.
- Number of possible geographic hypothesis too large to explore manually
  - Large number of geographic features and locations.
  - Number of interacting subsets of features grow exponentially.
  - Ex. Find teleconnections between weather events across ocean and land areas.
- SDM may reduce the set of possible hypothesis
  - Identify hypothesis supported by the data.
  - For further exploration using traditional statistical methods.

6.1.4 Spatial Data Mining : Actors

Domain Expert

- Identifies SDM goals, spatial dataset.
- Describe domain knowledge, e.g. well-known patterns, e.g. correlates.
- Validation of new patterns.

Data Mining Analyst

- Helps identify pattern families, SDM techniques to be used.
- Explain the SDM outputs to Domain Expert.

Joint effort

- Feature selection.
- Selection of patterns for further exploration.

6.1.5 Characteristics of Spatial Data Mining

- Auto correlation.
- Patterns usually have to be defined in the spatial attribute subspace and not in the complete attribute space.

Scanned by CamScanner

t, spatial data is growing

ually

d land areas.

the

- Longitude and latitude (or other coordinate systems) are the glue that link different data collections together.
- People are used to maps in GIS; therefore, data mining results have to be summarized on the top of maps.
- Patterns not only refer to points, but can also refer to lines, or polygons or other higher order geometrical objects.
- Large, continuous space defined by spatial attributes.
- Regional knowledge is of particular importance due to lack of global knowledge in geography (→ spatial heterogeneity).

Syllabus Topic : Spatial Vs. Classical Data Mining

6.2 Spatial Vs. Classical Data Mining

Parameters	Classical data mining	Spatial data mining
Data definition	Simple	Complex
Relationships	Explicit	Implicit
Data organization	Indexed	Vertical : stratification Horizontally : Spatial Indexing
Operation	Local	Local, focal and zonal
Statistical	Independence of context	Spatial autocorrelation
Output	Set based	Spatial based
Algorithm	Generic : divide and conquer, ordering, hierarchical structure	MBR, Spatial indexing, Plane sweeping, Computational geometry

Syllabus Topic : Spatial Data Structures

6.3 Spatial Data Structures

Spatial data consists of spatial objects made up of points, lines, regions, rectangles, surfaces, volumes, and even data of higher dimension which includes time. Examples of spatial data include cities, rivers, roads, counties, states, crop coverages, mountain ranges, parts in a CAD system, etc.

Scanned by CamScanner

6.3.1 R-tree

- To organize a collection of spatial objects, R-tree and its variants are designed.
- R-tree is represented by d-dimensional rectangles. Each node encloses its child node and corresponds to the smallest d-dimensional rectangle.
- Leaf nodes of tree having the pointers to original objects in the database rather than child node.
- During spatial query, few nodes are visited as often node correspond to disk page.
- Though the object is associated with one node, but spatially it may contained in several nodes. Even bonding rectangles may overlap though they belongs to different nodes.
- Therefore spatial query may require to visit many nodes to find a particular object.
- Rules for R-tree are similar to B-tree.
- Example given by Hanan Samet is given below :
  - All leaf nodes appear at the same level. Each entry in a leaf node is a 2-tuple of the form (R, O) such that R is the smallest rectangle that spatially contains object O. Each entry in a non-leaf node is a 2-tuple of the form (R, P) such that R is the smallest rectangle that spatially contains the rectangles in the child node pointed at by P. An r-tree of order (m, M) means that each node in the tree, with the exception of the root, contains between m <= [M/2] and M entries. The root node has at least two entries unless it is a leaf node.

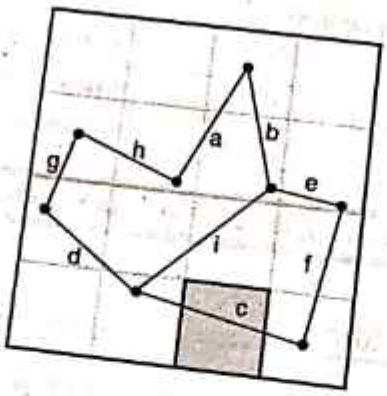
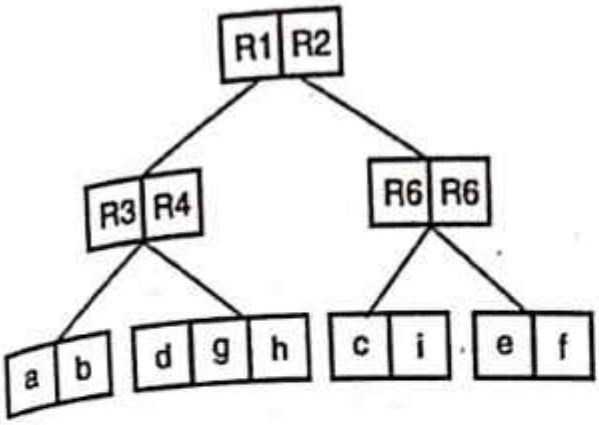
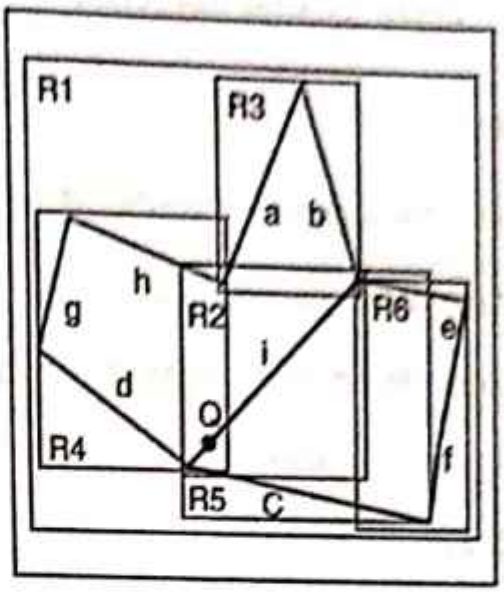


Fig. 6.3.1 : Example collection of line segments embedded in a 4 × 4 grid

- Example : Consider above example which is collection of line in 4 × 4 grid. Let us consider M =3 and m = 2. So one of the possible r-tree is given below.

Scanned by CamScanner





(a) R-tree for the collection of line segments (b) the spatial extents of the bounding rectangles

Fig. 6.3.2

- R-tree is not unique, rectangles depend on how objects are inserted and deleted from the tree.
- Problem is that to find some object you might have to go through several rectangles or whole database.

6.3.2 R+-tree

- R+-tree and Cell Trees used approach of discomposing space into cells and deals with collection of objects bounded by rectangles.
- Cell tree deals with collection of objects bounded by convex polyhedra.
- R+-trees is extension of k-d-B-tree.
- Try not to overlap the rectangles.
- If object is in multiple rectangles, it will appear multiple times.

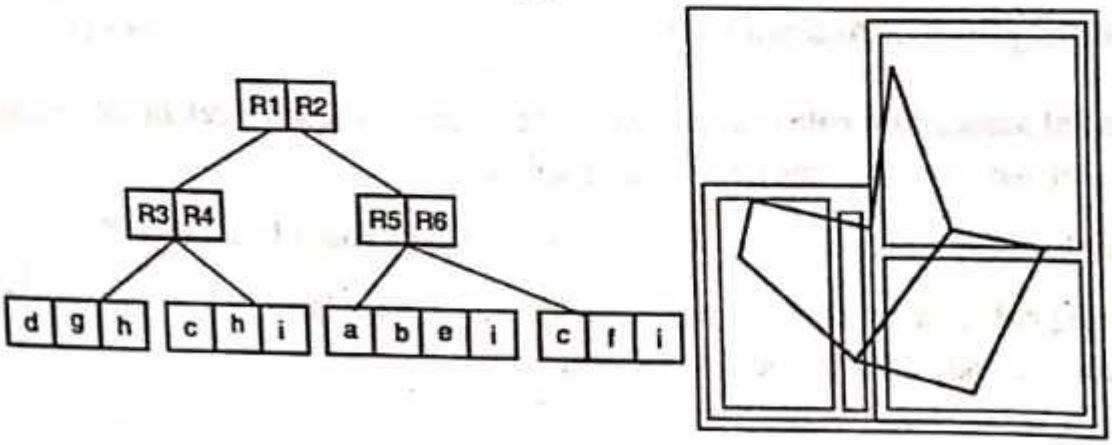


Fig. 6.3.3 : Decomposition of space into disjoint cells

Scanned by CamScanner

6.3.3 More Spatial Indexing

Uniform Grid

- Ideal for uniformly distributed data.
- More data-independence then R+-trees.
- Space decomposed on blocks on uniform size.
- Higher overhead.

Quadtree

- Space is decomposed based on data points.
- Sensitive to positioning of the object.
- Width of the blocks is restricted to power of two.
- Good for Set-theory type operations, like composition of data.

Syllabus Topic : Mining Spatial Association and Co-location Patterns

6.4 Mining Spatial Association and Co-location Patterns

- Looking for interesting, useful, unexpected spatial patterns of information embedded in large databases.
- Spatial Data Mining is searching for spatial patterns :
  - **Non-trivial search** - as "automated" as possible reduce human effort.
  - Interesting, useful and unexpected spatial pattern.

6.4.1 Mining Spatial Association

- **A spatial association rule** is a rule indicating certain association relationship among a set of spatial and possibly some non-spatial predicates.
- **Example :** "Most big cities in Canada are close to the Canada U.S. border"
- **A strong rule** indicates that the patterns in the rule have relatively frequent occurrences in the database and strong implication relationships.

Scanned by CamScanner

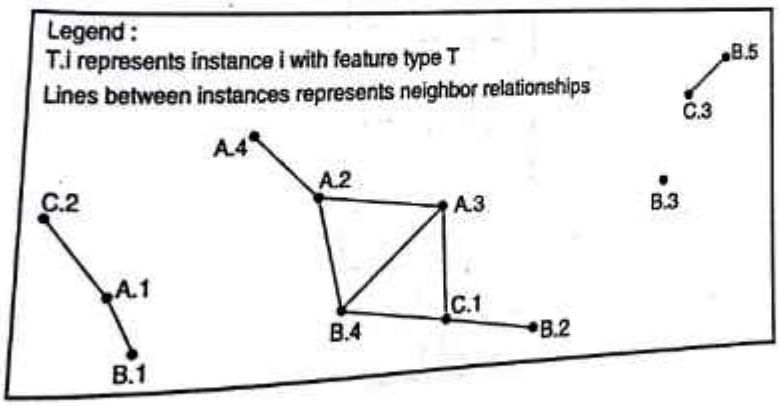
- A **spatial association rule** is a rule in the form of :

  $P1 \wedge \ldots \wedge Pm \rightarrow Q1 \wedge \ldots \wedge Qn \quad (c\%)$,

  where at least one of the predicates P 1 ,...., Pm , Q 1 ,...., Qn is a spatial predicate, and c% is the ***confidence*** of the rule [which indicates that c% of objects satisfying the antecedent of the rule will also satisfy the consequent of the rule].
- A set of predicates P is **large** in set S at level k if the support of P is no less than its minimum support threshold σ`k for level k, and all ancestors of P from the concept hierarchy are large at their corresponding levels.
- The confidence of a rule "P → Q/S" is **high** at level k if its confidence is no less than its corresponding minimum confidence threshold φ`k
- A rule "P → Q/S" is **strong** if predicate "P ∧ Q" is **large** in set S and the confidence of "P → Q/S" is high.

6.4.2 Mining Co-location Patterns

- Colocation is a set of *spatial features* that frequently occur in *together*.
- For Example :
  - Ecology : Symbiotic relationship in animals or plants
  - Public health : Environmental factors and cancers
  - Public safety : Crime generators and crime events
- Colocation pattern is a subset of spatial event types or instances of these event types *frequently occur together.*
- Consider following example :



Scanned by CamScanner

- From the above example,
  - Spatial event types are A, B, C
  - Spatial event instances are A.1, A.2, A.3, B.1, B.2 ... ...
  - Candidate Colocation are the combinations of event types (A, B), (B, C), (A,C)...
  - Neighbor relationship represented by solid line between event instances (A.1, B.1), (A.1, C.2), (A.3, C.1).....
  - Table instance of (A, B) i.e. direct relationship between events A and B from above graph :

    (A.1, B.1)

    (A.2, B.4)

    (A.3, B.4)

- Interestingness Measure are defined by participation ratio (pr) and participation index (pi)

1. Participation ratio is defined for the colocation pattern

$$C = (f_1, f_2, ..., f_k)$$

$$pr(c, f_i) = \frac{|\pi f_i \text{ Table.Instance } (c)|}{|\text{Table.Instance } (f_i)|}$$

2. Participation index for given c can be calculated as

$$pi(c) = \min_i \{pr(c, f_i)\}$$

- **Example :** Find pi

Table instance for each pattern :

T1	T2	T3
A	B	C
A.1	B.1	C.1
A.2	B.2	C.2
A.3	B.3	C.3
A.4	B.4	
	B.5	

Scanned by CamScanner

- Table instance of (A, B, C) is as given below because only A.3, B.4, C.1 have direct relationship

Table instance for ABC		
A.3	B.4	C.1

- Calculate the participation ratio using formula given above :

$$pr((A, B, C), A) = \frac{1}{4}$$

(i.e. number of table instances of ABC / number of instances of A)

$$pr((A, B, C), B) = \frac{1}{5}$$

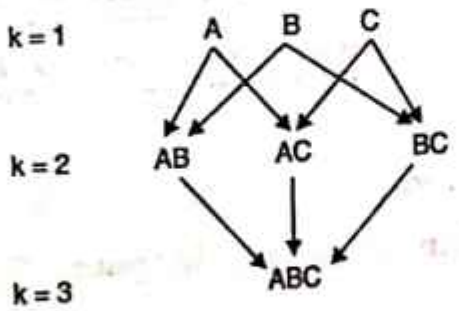
$$pr((A, B, C), C) = \frac{1}{3}$$

$$pi(c) = \min(1/4, 1/5, 1/3) = 1/5$$

Colocation Mining Algorithm

- Starting with k = 1
- Iterative until no prevalent pattern
- Generate size k colocation patterns $\{c_k\}$
- Generate table instance of each c_k
- Compute each $pi(c_k)$, add to result if prevalent

$$k = k + 1$$



Scanned by CamScanner

T1	T2	T3
A	B	C
A.1	B.1	C.1
A.2	B.2	C.2
A.3	B.3	C.3
A.4	B.4	
	B.5	

T4	T5	T6
A B	A C	B C
A.1, B.1	A.1, C.2	B.2, C.1
A.2, B.4	A.3, C.1	B.4, C.1
A.3, B.4		B.5, C.3

T7
A B C
A.3, B.4, C.1

- Calculate the pi for T4, T5, T6 and T7

  pi(T4) = 0.4,

  pi(T5) = 0.5

  pi(T6) = 0.6

  pi(T7) = 0.2

- If a participation index threshold delta = 0.5 then find all colocation patterns c :
- From the above pi, final colocation patterns are (A, C) and (B, C) as pi(T5) and pi(T6) is greater and equal to threashold.
- **Algorithm given by Huang Yan is :**

Colocation Mining Algorithm : Filter-Based approach :

Symbol	Description
c_k	candidate colocation of size k
C_k	all candidate colocation of size k
P_k	all prevalent colocation of size k

- Starting with k=1
- Iterative until no prevalent pattern
- Generate size k candidate patterns C_k from prevalent patterns P_{k-1}
- For each candidate $c_k \in C_k$
- Check all subset patterns
- If any subset pattern not prevalent, ***prune out*** c_k

Scanned by CamScanner

- Generate coarse table instance of each remaining $c_k \in C_k$
- Compute each $pi(c_k)$
- If $pi(c_k)$ based on coarse resolution below threshold,

  prune out c_k
- Generate table instance of each remaining $c_k \in C_k$
- Compute each $pi(c_k)$, if above threshold, add c_k to set P_k

$k = k+1$

Syllabus Topic : Spatial Clustering Techniques - CLARANS Extension

6.5 Spatial Clustering Techniques : CLARANS

CLARANS (A Clustering Algorithm based on Randomized Search)

- Clustering is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes (Ester, M., Frommelt, A., Kriegel, H.-P., and Sander, J, 1998). In spatial data sets, clustering permits a generalization of the spatial component like explicit location and extension of spatial objects which define implicit relations of spatial neighbourhood.
  - CLARANS improves on CLARA by using multiple different samples.
  - For every step of search CLARANS draws a sample neighbour.
  - So this is not confining a search to localized area.
  - It uses two additional parameters numlocal and maxneighbor.
  - Numlocal indicates the number of samples to be taken.
  - Maxneighbor is the number of neighbors of a node to which any specific node can be compared.

Algorithm CLARANS given by Jiawei Han

1. Input parameters numlocal and maxneighbor. Initialize i to 1, and mincost to a large number.
2. Set current to an arbitrary node in Gn;k.
3. Set j to 1.

Scanned by CamScanner

4. Consider a random neighbor S of current, and based on 5, calculate the cost differential of the two nodes.
5. If S has a lower cost, set current to S, and go to Step 3.
6. Otherwise, increment j by 1. If j maxneighbor, go to Step 4.
7. Otherwise, when j > maxneighbor, compare the cost of current with mincost. If the former is less than mincost, set mincost to the cost of current and set bestnode to current.
8. Increment i by 1. If i > numlocal, output bestnode and halt. Otherwise, go to Step 2.

Steps 3 to 6 above search for nodes with progressively lower costs. But, if the current node has already been compared with the maximum number of the neighbors of the node (specified by maxneighbor) and is still of the lowest cost, the current node is declared to be a "local" minimum. Then, in Step 7, the cost of this local minimum is compared with the lowest cost obtained so far. The lower of the two costs above is stored in mincost. Algorithm CLARANS then repeats to search for other local minima, until numlocal of them have been found.

Syllabus Topic : Web Mining

6.6 Web Mining

Introduction to Web Mining

– Web Mining refers to application of data mining techniques to web data.

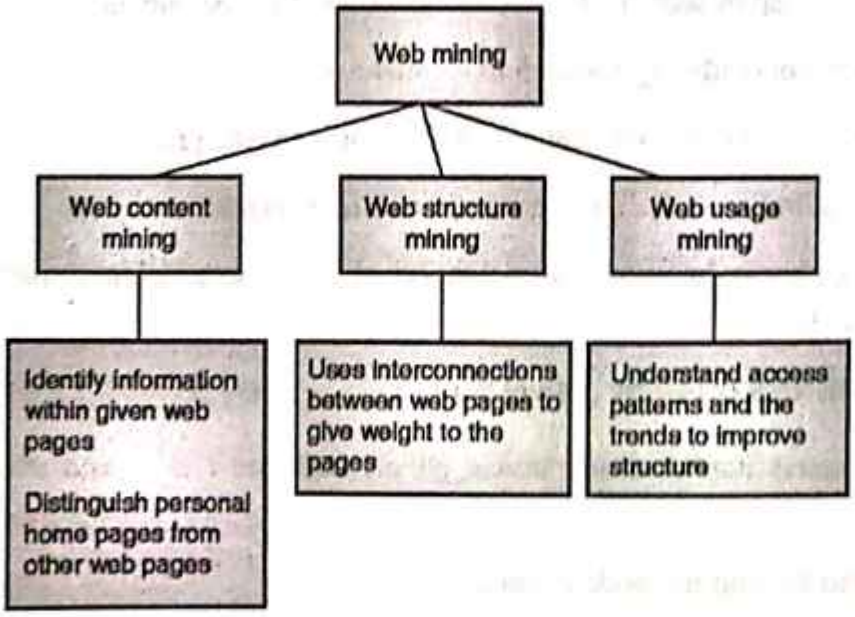


Fig. 6.6.1 : Web Mining Taxonomy

Scanned by CamScanner

on 5, calculate the cost differential of

p 4.

urrent with mincost. If the former
set bestnode to current.

Otherwise, go to Step 2.

wer costs. But, if the current
of the neighbors of the node
e current node is declared to
minimum is compared with
above is stored in mincost.
minima, until numlocal of

ta.

- It helps in solving the problem of how users are using the web sites.
- The process involves mining logs or analysis of the logs to get meaningful data from them.
- It is the process of discovering the useful and previously unknown information from the web data.
- Web data is :
  - Web content – text, image, records, etc.
  - Web structure – hyperlinks, tags, etc.
  - Web usage – http logs, app server logs, etc.

6.6.1 How Web Mining is Different from Classical DM ?

- Web Mining is similar to data mining.
- It differs in data collection.
- **Data Mining :** The collection of data is already done and stored in a data warehouse.
- **Web Mining :** Data collection is done by crawling through a number of target web pages.

6.6.2 Benefits of Web Data Mining :

- Match your available resources to visitor interests.
- Increase the value of each visitor.
- Improve the visitor's experience at the website.
- Perform targeted resource management.
- Collect information in new ways.
- Test the relevance of content and web site architecture.

Syllabus Topic : Web Content Mining

6.7 Web Content Mining

6.7.1 Introduction to Web Content Mining

- The process to discover useful information from the content of a web page.

Scanned by CamScanner

- The type of the web content may consist of :
  - Text
  - Image
  - Audio
  - Video
  - Web Content Mining is also known as Web Text Mining
  - Web Content Mining uses the following techniques
  - Natural Language Processing(NLP)
  - Information Retrieval (IR)

6.7.2 Text Mining

- The process of deriving high quality information from text.
- Text mining is an interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics, and computational linguistics. As most information (over 80%) is currently stored as text, text mining is believed to have a high commercial potential value.
- Using Statistical Pattern learning, high quality of information is derived.
- Text mining involves the process of :
  - Structuring the input text by Parsing.
  - Addition of some derived linguistic features and the removal of others.
  - Subsequent insertion into a database.
  - Deriving patterns within the structured data.
  - Evaluation and interpretation of the output.
- 'High quality' in text mining refers to some combination of :
  - Relevance
  - Novelty
  - Interestingness

Scanned by CamScanner

nation retrieval, data
As most information
e a high commercial

l.

s.

- Typical text mining tasks include :
  - Text categorization
  - Text clustering
  - Concept/entity extraction
  - Document summarization
  - Entity relation modeling (i.e. learning relations between named entities)

Syllabus Topic : Web Usage mining

6.8 Web Usage Mining

6.8.1 What is Web Usage Mining ?

- Web Usage mining is the type of web mining activity which predicts about which pages are likely to be visited in near future based on the active users' behavior. Such pages can be pre-fetched to reduce access times.
- The usage data records the user's behaviour when the user browses or makes transactions on the web site in order to better understand and serve the needs of users or Web-based applications.
- Automatic discovery of patterns from one or more Web servers.
- Organizations often generate and collect large volumes of data; most of this information is usually generated automatically by Web servers and collected in server log. Analyzing such data can help these organizations to determine :
  - The value of particular customers.
  - Cross marketing strategies across products.
  - The effectiveness of promotional campaigns, etc.
- The first web analysis tools simply provided mechanisms to report user activity as recorded in the servers. Using such tools, it was possible to determine such information as :
  - The number of accesses to the server.
  - The times or time intervals of visits.
  - The domain names and the URLs of users of the Web server.

Scanned by CamScanner

- A very little or no analysis is provided by these tools of the data relationships among the accessed files and directories within web space.
- The tools for discovery and analysis of patterns may be classified into following two categories :
  - Pattern Discovery Tools
  - Pattern Analysis Tools
- Web servers, Web proxies, and client applications can quite easily capture **Web Usage data.**
- **Web Server Log :** It is a file that is created by the server to record all the activities it performs.
- For example when a user enters URL into the browsers address bar or requests by clicking on a link.
- The page request sent to the web server maintains the following information in its log :
  - Information about the URL.
  - Whether the request was successful.
  - The users IP address.
  - Time and date about the page request completed.
  - The referrer header.
- Web Mining systems use the web usage data and based on the analysis the system can discover useful knowledge about a system's usage characteristics and user's interest which finds its application in :
  - Personalization and Collaboration in Web-based systems,
  - Marketing,
  - Web site design and evaluation,
  - Decision support.

6.8.2 Purpose of Web Usage Mining

- Web usage mining has been used for various purposes :
  - A knowledge discovery process for mining marketing intelligence information from Web data.

Scanned by CamScanner

- In order to improve the performance of the website, web usage logs can be used to extract useful web traffic patterns.

- Search engine transaction logs also provide valuable knowledge about user behaviour on Web searching.
- Such information is very useful for a better understanding of users' Web searching and information seeking behaviour and can improve the design of Web search systems.
- Web usage mining is useful in finding interesting trends and patterns which can provide important knowledge about the users of a system.
- Several Machine learning and data mining techniques for e.g. Association rule mining, classification and clustering can be applied.
- Web usage data provides a extremely useful way to learn user's interest.
- Web logs may be used in web usage mining to help identify users who have accessed similar web pages. These patterns may be useful in web searching and filtering.
- For e.g. Amazon.com uses collaborative filtering for recommending books to their customers based on the data collected from other customers similar interests or purchasing history.

6.8.3 Web Usage Mining Activities

- Pre-processing Web log :
  - Cleanse.
  - Remove extraneous information.
  - Sessionize.
- **Session :** All the requests that a single client makes to a web server.
- Pattern Discovery :
  - Uncovering the traversal patterns.
  - Traversal pattern is a set of pages visited by user in a session.
  - Count patterns that occur in sessions.
  - Pattern is sequence of pages references in session.
  - Pattern Discovery uses techniques such as statistical analysis, association rules, clustering, classification, sequential pattern, dependency Modeling.

Scanned by CamScanner

- Pattern Analysis :
  - Once the patterns have been identified, they must be analyzed to determine how that information can be used.
  - A process to gain Knowledge about how visitors use Website in order to Prevent.
    - (i) Disorientation and help designers to place important information/functions exactly where the visitors look for and in the way users need it.
    - (ii) Build up adaptive Website server.

6.8.4 Web Server Log

6.8.4(A) Structure of Web Log

The Web log contains the following information :

(i) The user's IP address,

(ii) The user's authentication name,

(iii) The date-time stamp of the access,

(iv) The HTTP request,

(v) The response status,

(vi) The size of the requested resource, and optionally also,

(vii) The referrer URL (the page the user "came from"),

(viii) The user's browser identification.

Note : Different servers have different log formats.

6.8.4(B) Web Server Log - An Example

Web log fields

- IP
  - 152.152.98.11
  - IP address - can be converted to host name, such as xyz.example.com
- Name
  - The name of the remote user (usually omitted and replaced by a dash "-")

Scanned by CamScanner

- Login
  - Login of the remote user (also usually omitted and replaced by a dash "-")
- Date/Time/TZ
  - 16/Nov/2005:16:32:50 -0500
- Request, Status code, Object size, Referrer, User agent.

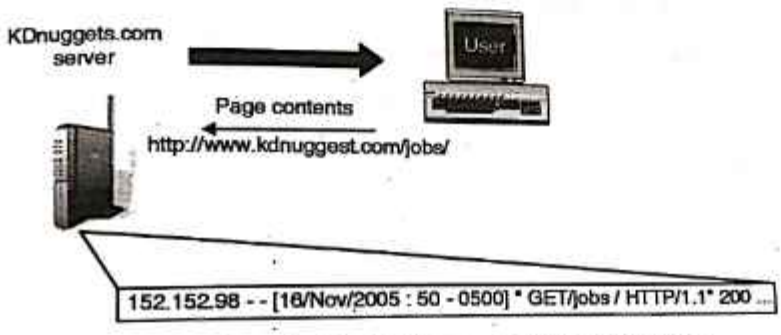


Fig. 6.8.1 : Example of a Web Server and Client Interaction

Ex. 6.8.1 : The following is a fragment from the server logs for loganalyzer.net. All the relative URL's are for the base URL http://www.loganalyzer.net/.

Soln. : First let's look at a fragment of log file....

```
66.249.65.107 - - [08/Oct/2007:04:54:20 -0400] "GET /support.html HTTP/1.1" 200 11179 "-"
"Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

111.111.111.111 - - [08/Oct/2007:11:17:55 -0400] "GET / HTTP/1.1" 200 10801
"http://www.google.com/search?q=log+analyzer&ie=utf-8&oe=utf-8 &aq=t&rls=org.mozilla:en-
US:official&client=firefox-a" "Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7)
Gecko/20070914 Firefox/2.0.0.7"
111.111.111.111 - - [08/Oct/2007:11:17:55 -0400] "GET /style.css HTTP/1.1" 200 3225
"http://www.loganalyzer.net/" "Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7)
Gecko/20070914 Firefox/2.0.0.7"
```

The above fragment shows two visitors :

1. A google spider, which is 66.249.65.107. The pages are retrieved and indexed them for their search engine.
2. Another visitor is from IP address 111.111.111.111 who gave a search for Nihuo Web Log analyzer homepage.

Scanned by CamScanner

A few things to note :

1. A single hit is represented by each line in the file, for a file on web server.
2. A web page hit is a page view. For e.g. if a web page contains 4 images that a hit on that page will generate 5 hits on the web server, one hit for the web page and 4 for the images.
3. IP address or cookie is used to determine a unique visitor. A visit session is terminated if the user falls inactive for more than 30 minutes. So if a unique visitor visits the same website twice then it gets reported as two visits.

Description of fields

Let's look at one line from the above fragment.

```
111.111.111.111
-
-
[08/Oct/2007:11:17:55 -0400]
"GET / HTTP/1.1"
200
10801
"http://www.google.com/search?q=log+analyzer&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a"
"Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7) Gecko/20070914 Firefox/2.0.0.7"
```

1. **IP address : "111.111.111.111"**

   This is the IP address who visited the site Nihuo web log Analyzer.

2. **Remote log name: "-"**

   This field will return a dash until Identity Check is set on the web server.

3. **Authenticated user name : "-"**

   This information is available only when accessing content which is password protected by the authenticate system of web server.

4. **Timestamp : [08/Oct/2007:11:17:55 -0400]**

   This field represents the timestamp as seen by the web server.

Scanned by CamScanner

5. **Access request : "GET / HTTP/1.1"**

This represents the HTTP method called as GET and the version of the HTTP protocol which is HTTP/1.1

6. **Result status code : "200"**

This field represents the status of the request, code 200 represents success. Other code like (e.g. HTTP 404 "File Not Found" or HTTP 500 "Internal Server Error").

7. **Bytes transferred : "10801"**

This field tells the amount of data transferred to the user. In this example its 10801 bytes or 10k is the size of the homepage.

8. **Referrer URL**

"http://www.google.com/search?q=log+analyzer&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a"

The referring URL. This tells the page the visitor was on when they clicked to come to this page.

9. **User Agent**

"Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7) Gecko/20070914 Firefox/2.0.0.7"

The "User Agent" identifier. The user agent is whatever application the visitor used to access the site, it could include a browser, a web robot, a link checker an FTP client.

In this case "Mozilla/5.0" probably means visitor's browser is Mozilla compatible," Windows NT 5.2" indicates Windows 2003, "en-US" probably implies it's an English version, "Firefox/2.0.0.7" means Firefox 2.0. In the first line, "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)" means this hit is caused by googlebot(spider).

Syllabus Topic : Web Structure Mining

6.9 Web Structure Mining

6.9.1 Introduction to Web Structure Mining

- The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages.

Scanned by CamScanner

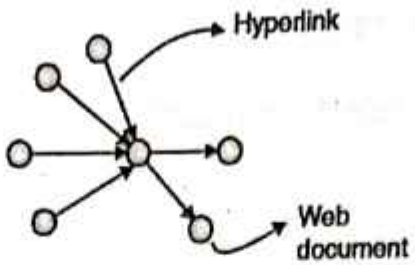


Fig. 6.9.1 : Web Graph Structure

- The process of using the graph theory to analyze the node and connection structure of a web site. Web structure mining can be divided into two kinds :
  - Extract patterns from hyperlinks in the web. A hyperlink is a structural component that connects the web page to a different location.
  - Mining the document structure. It is using the tree-like structure to analyze and describe the HTML or XML tags within the web page.
- Web structure mining has been largely influenced by research in :
  - Social network analysis.
  - Citation analysis (bibliometrics).
  - **In-links :** The hyperlinks pointing to a page.
  - **Out-links :** The hyperlinks found in a page.
  - Usually, the larger the number of in-links, the better a page is.
  - By analyzing the pages containing a URL, we can also obtain.
  - **Anchor text :** How other Web page authors annotate a page and can be useful in predicting the content of the target page.
- **Web Structure Mining :** Discovers the structure information from the web
  - This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level.
  - The research at the hyperlink level is also called **Hyperlink Analysis.**
- Motivation to study Hyperlink Structure :
  - Hyperlinks serve two main purposes.
  - Pure Navigation.
  - Point to pages with authority on the same topic of the page containing the link.
  - This can be used to retrieve useful information from the web.

Scanned by CamScanner

connection structure of a

rlink is a structural

cture to analyze and

be useful in

level or at

Web Structure Terminology

- **Web-graph :** Directed graph representing the web.
- **Node :** Each Web page represents a node of the Web-graph.
- **Link :** Each hyperlink on the Web is a directed edge of the Web-graph.
- **In-degree :** The number of distinct links that point to a node.
- **Out-degree :** The number of distinct links originating at a node that point to other nodes.
- **Directed Path :** It is a sequence of links, starting from a node say r that can be followed to reach another node say t.
- **Shortest Path :** The path with the shortest length out of all the paths between nodes p and q. (number of links on it)
- **Diameter :** It is the maximum of all the shortest paths between a pair of nodes p and q, for all pairs of nodes p and q in the Web-graph.

6.9.2 Techniques of Web Structure Mining

1. Page Rank
2. CLEVER

6.9.2(A) Page Rank Technique Used by Google

- Prioritize pages returned from search by looking at Web structure.
- Importance of page is calculated based on number of pages which point to it - Backlinks.
- Weighting is used to provide more importance to backlinks coming from important pages.
- Page Rank of Page P is defined as PR(p)

$$PR(p) = c\,(PR(1)/N_1 + \ldots + PR(n)/N_n)$$

PR(i) : Page Rank for a page i which points to target page p.

N_i : number of links coming out of page i

c : the constant having value between 0 and 1 and is used for normalization

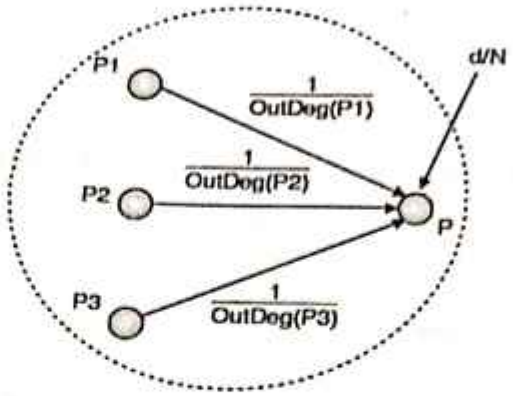
Example

Google's Page Rank

Rank of a web page depends on the rank of the web pages pointing to it.

he New Revised Sy
niversity
ic year 2018-2019
ce Based Credit

Scanned by CamScanner



$$PR(P) = d/N + (1-d)\left(\frac{PR(P1)}{OutDeg(P1)} + \frac{PR(P2)}{OutDeg(P2)} + \frac{PR(P3)}{OutDeg(P3)}\right)$$

The Page Rank Algorithm

Set PR $\leftarrow [r_1, r_2, \ldots r_N]$, where r_i is some initial rank of page I , and N the number of Web pages in the graph;

$d \leftarrow 0.15$; $D \leftarrow [1/N \ldots\ldots 1/N]^T$;

A is the adjacency matrix as described above;

do

$$PR_{i+1} \leftarrow A^T * PR_i ;$$

$$PR_{i+1} \leftarrow (1-d) * PR_{i+1} + d*D;$$

$$\delta \leftarrow \| PR_{i+1} - PR_i \|_1$$

While $\delta < \varepsilon$, where ε is a small number indicating the convergence threshold.

return PR.

6.9.2(B) CLEVER Technique

- Identify authoritative and hub pages by creating weights.

  (i) Authoritative Pages :

  - Highly important pages.
  - Best source for requested information.

Scanned by CamScanner

(ii) Hub Pages :

- Contain links to highly important pages.

Hubs and **authorities** are 'fans' and 'centers' in a bipartite core of a web graph.

- A good hub page is one that points to many good authority pages.
- A good authority page is one that is pointed to by many good hub pages.

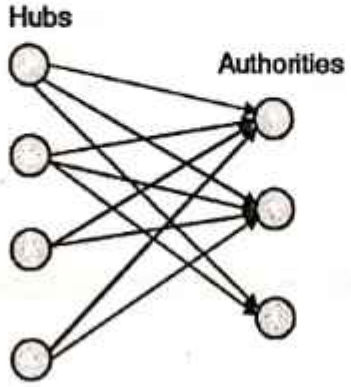


Fig. 6.9.2 : Bipartite Core

HITS Algorithm (Hyperlink-induced topic Search)

- HITS (Hypertext Induced Topic Search) algorithm is a kind of rank algorithm that analyzes Web resource based on local link.
- The difference between Page Rank and HITS is that HITS is related to query, and Page Rank is a kind of query unrelated algorithm. Page Rank algorithm gives each page a rank value which is unique and unrelated to query keyword, but HITS algorithm gives each page two values, which are Authority value and Hub value.
- Authority page and Hub page are two important concepts in HITS algorithm, which are the concepts all related to query keyword.
- Authority page refers to some page that is most related to query keyword and combination.
- Hub page is the page that includes multiple Authority pages. Hub page itself may not have direct relation to query content, but through it, the Authority page with direct relation can be linked.

Analysis of HITS Algorithm

The central idea of HITS algorithm is that :

- Firstly, use text-based retrieval algorithm to obtain a Web subset, and the pages in this subset all have relativity to user query.

Scanned by CamScanner

- Then, HITS performs link analysis on this subset, and find out the Authority pages and Hub pages related to query in the subset.
- The selection of subnet in HITS algorithm is acquired by means of keyword matching. This subnet is defined as root set R, then use link analysis to acquire set S from root set R. S is the page that includes Authority page and ultimately meet the query requirement. The process from R to S is called "Neighborhood Expand".

The algorithm procedure for computing S is shown below:

Step 1 : Use text keyword matching to acquire root set R, which includes thousands of URL or more;

Step 2 : Define S to R, that is S and R are equal;

Step 3 : To each page p in R, put the hyperlinks included by p into set S; put the pages referring to p into set S;

Step 4 : S is the acquired expanded neighbourhood set.

HITS algorithm needs three parameters, which are query keyword, maximum capability of root set R, and maximum capability of expanded neighbourhood S. After using the algorithm above, the pages in S will have more Authority pages and Hub pages which meet the query keyword.

Analysis of HITS Link

- The process of HITS link analysis takes advantage of the attribute that Authority and Hub are interacting to identity them from expanded set S.
- Assume the pages in expanded set S are respectively 1, 2, ..., n.
- B (i) represents the page set referring to page i,
- F (i) represents the page set referred by page i,
- HITS generates an authority value a_i and Hub value h_i for each page in S.
- The initial value of computing initial a_i and h_i can be an arbitrary value, similar to PageRank. HITS can use iterative method to acquire convergence value.
- There are two steps in its iterative procedure, which are step I and step O.
  - In step I, the authority value of each page is the sum of the Hub values of pages referring to it.

Scanned by CamScanner

- In step O, the Hub value of each page is the sum of the authority values of pages referring to it. That is :

$$I: \quad a_i = \sum_{j \in B(i)} h_j$$

$$O: \quad h_i = \sum_{j \in B(i)} a_j$$

The two steps, I and O, are based on the fact that one Authority page is always referred by many Hub pages, and one Hub page includes many Authority pages.

HITS algorithm iteratively computes the two steps, I and O, till they converge. At last, a_i and h_i are the Authority and Hub value of page i.

The procedure is shown below :

Step 1 : Initialize a_i, h_i;

Step 2 : Iterate procedure I, O; Perform iteration I; Perform iteration O; Normalize the value of a and h, let $\sum_i a_i^2 = 1$; $\sum_i h_i^2 = 1$

Step 3 : Complete iteration

- Fig. 6.9.3 shows the application of HITS algorithm in a subgraph including 6 nodes.

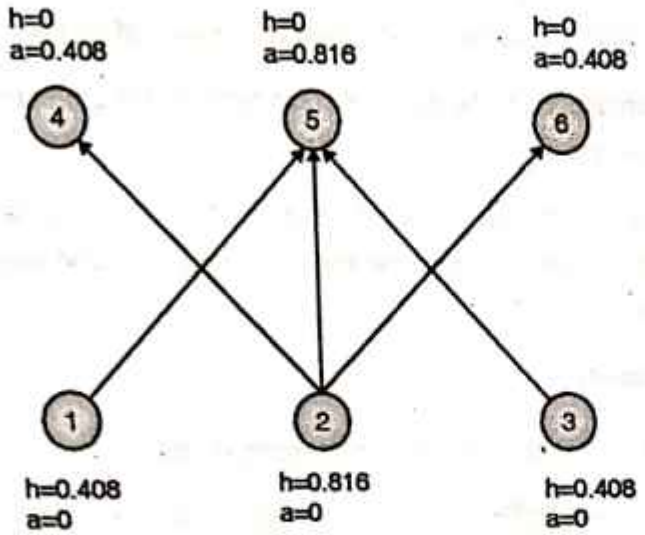


Fig. 6.9.3 : HITS algorithm on six-nodes graph

Scanned by CamScanner

- As shown in Fig. 6.9.3, the authority value of node 5 is equal to the Hub values of nodes 1, 3 which refers to it, after normalization, the value is 0.816.
- Assume $A_{m \times m}$ is the matrix of subgraph, then the value of position i, j in matrix A is equal to 1 (if page i refers to j), or 0. Set a to be the authority value vector $[a_1, a_2, ..., a_n]$, h to be the hub value vector $[h_1, h_2, ..., h_n]$, then the iteration I, O can be expressed as $a = Ah$, $h = A^T a$. After completing the iteration, the values of authority and hub respectively satisfy $a = c_1 AA^T a$, $h = c_2 A^T Ah$, which c_1 and c_1 are constant in order to satisfy the normalization condition. Thus, vector a and vector b respectively become the eigenvector of matrix AA^T and matrix $A^T A$. This feature is similar to Page Rank algorithm, their convergence speeds are decided by eigenvector.

6.10 Web Crawlers

- A web crawler is an automated program that scans or crawls through the internet pages to create an index of the data.
- A web crawler is also known as Web spider, web robot, bot, crawler and automatic indexer.
- Search engines makes use of web crawler to collect information about the data on public web pages, their primary purpose is to collect data so that when a user enters a search term on their site, they can be quickly be provided with relevant web sites.
- The search engine's web crawler visits a web page, it collects information like visible text, hyperlinks and the various tags like keyword rich meta tag. This information may be used by the search engine to determine what the site is about and index the information. The website is then included in the search engines database and its page ranking process.
- Web crawling is considered to be an important method for collecting data and keeping up with the expanding internet.
- A vast number of pages are added on a continuous basis and information keeps changing continuously, a web crawler is a method for search engines and others to ensure that their database is upto date.

Different Types of Crawler

- **Traditional crawler :** Visits entire Web and replaces index.
- **Periodic crawler :** Visits portions of the Web and updates subset of index.
- **Incremental crawler :** Selectively searches the Web and incrementally modifies index.
- **Focused crawler :** Visits pages related to a particular subject.

Scanned by CamScanner

- A focused web crawler takes a set of well-selected web pages exemplifying the user interest.
- The focused crawler starts from the given pages and recursively explores the linked web pages.
- Traditional crawlers uses breadth-first and searches the whole web, while a focused crawler searches only a portion of the web using best first directed by the users interest.
- A focused crawler searches for multimedia content in the web instead of plain HTML documents.

– Components of focused crawler are :
  - **Classifier :** Assigns relevance score to each page based on crawl topic.
  - **Distiller :** Identifies hub pages. **Hub Pages** contain links to many relevant pages. Must be visited even if not high relevance score.
  - **Crawler :** Visits pages to based on crawler and distiller scores.

Context focused crawler

– **Context Graph :**
  - Context graph created for each seed document.
  - Root is the seed document.
  - Nodes at each level show documents with links to documents at next higher level.
  - Updated during crawl itself.

Approach

– Construct context graph and classifiers using seed documents as training data.

– Perform crawling using classifiers and context graph created.

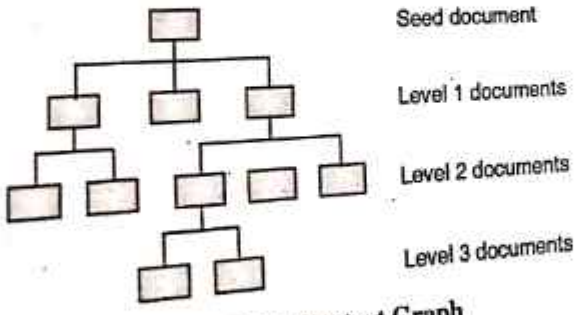


Fig. 6.10.1 : Context Graph

Scanned by CamScanner

Syllabus Topic : Applications of Web Mining

6.11 Applications of Web Mining

1. E-Learning.
2. Digital libraries.
3. E –Government.
4. Electronic commerce.
5. E-Politics and E-Democracy.
6. Security and Crime Investigation.
7. Electronic Business.

□□□

Chapter Ends

Scanned by CamScanner

Appendix - A

Solved University Question Paper of May 2019

May 2019

Q. 1(a) What are Spatial Data Structures? Outline their importance in GIS. **(Section 6.3)**

Ans. :

(5 Marks)

- Spatial data are data that have spatial or location components. Spatial data can be viewed as data about objects that themselves are located in a physical space. This may be implemented with a specific location attribute(s) such as address or latitude/ longitude or by a partitioning of the database based on location.
- Geographic Information Systems (GIS) are used to store information related to geographic locations on the surface of the earth.

Q. 1(b) What is metadata? Why do we need metadata when search engines like google seem so effective? **(Section 1.7)**

Ans. :

(5 Marks)

- Simple search engines retrieve relevant documents usually using a keyword-based retrieval technique similar to those found in traditional IR systems.
- The data warehouse also stores all the Meta data (data about data) definitions used by all processes in the warehouse.
- It is used for variety of purpose including :
  - **The extraction and loading process :** Meta data is used to map data sources to a common view of information within the warehouse.
  - **The warehouse management process :** Meta data is used to automate the production of summary tables.
  - **As part of Query Management process :** Meta data is used to direct a query to the most appropriate data source.
- The structure of Meta data will differ in each process, because the purpose is different.

Scanned by CamScanner

Q. 1(c) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. **(Section 3.12.3)** **(5 Marks)**

Q. 1(d) With respect to web mining, is it possible to detect visual objects using meta-objects? **(5 Marks)**

Ans. :

It is possible to detect visual objects using meta-objects as visual objects are represented by small rectangle that completely contains that object which is minimum bounding rectangle (MBR).

Q. 2(a) Suppose that a data warehouse for DB- 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. **(10 Marks)**

(i) Draw a snowflake schema diagram for the data warehouse.

(ii) 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.

Ans. :

(i) A snowflake schema is as shown

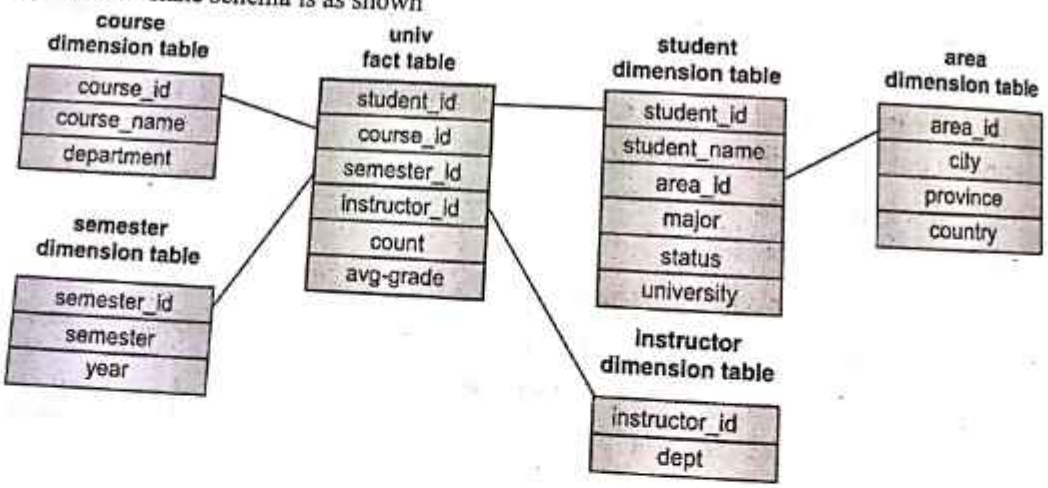


Fig. 1-Q. 2(a)

(ii) 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 DB-University student. The specific OLAP operations to be performed are :

- Roll-up on course from course id to department.

Scanned by CamScanner

- Roll-up on student from student id to university.
- Dice on course, student with department = "CS" and university = "DB-University".
- Drill-down on student from university to student name.

Q. 2(b) What is the relationship between data warehousing and data replication? Which form of replication (synchronous or asynchronous) is better suited for data warehousing? Why? Explain with appropriate example. **(10 Marks)**

Ans. :

- Data warehouses are carefully designed databases that hold integrated data for secondary usage. If you only need data from one system, but can't impact the performance of that system, then it is suggested to take a copy i.e. a replicated data store unless the complexities and width of data is very large and the various ways to require access to it are very high, a data warehouse would be overkill.
- A replicated data store is a database that holds schemas from other systems, but doesn't truly integrate the data. This means it is typically in a format similar to what the source systems had. The value in a replicated data store is that it provides a single source for resources to go to inorder to access data from any system without negatively impacting the performance of thatsystem.
- Data replication is simply a method for creating copies of data in a distributed environment.
- Replication technology can be used to capture changes to source data.
  - **Synchronous replication :** Synchronous replication is used for creating replicas in real time. In synchronous replication data is written to primary storage and the replica is done simultaneously. Primary copy and the replica should always remain synchronized.
  - **Asynchronous replication :** It is used for creating time delayed replicas. In asynchronous replication data is written to the primary storage first and then copy data to the replica.
- Synchronous replication is best suited for data warehouse as it creates replicas in real time.

Q. 3(a) The following table consists of training data from an employee database. The data have been generalized. For example, "31::: 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row. **(10 Marks)**

Scanned by CamScanner

Department	Status	Age	Salary	Count
Sales	Senior	31 ... 35	46 K ... 50 K	30
Sales	Junior	26 ... 30	26 K ... 30 K	40
Sales	Junior	31 ... 35	31 K ... 35 K	40
Systems	Junior	21 ... 25	46K ... 50 K	20
Systems	Senior	31 ... 35	66K ... 70K	5
Systems	Junior	26 ... 30	46K ... 50K	3
Systems	Senior	41 ... 45	66K ... 70K	3
Marketing	Senior	36 ... 40	46 K ... 50 K	10
Marketing	Junior	31 ... 35	41K ... 45K	4
Secretary	Senior	46 ... 50	36K ... 40K	4
Secretary	Junior	26 ... 30	26K ... 30 K	6

Let status be the class label attribute.

(i) How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple (i.e., of each row entry)?

(ii) Use your algorithm to construct a decision tree from the given data.

Ans. :

(i) The basic decision tree algorithm should be modified as follows to take into consideration the count of each generalized data tuple. The count of each tuple must be integrated into the calculation of the attribute selection measure (such as information gain). Take the count into consideration to determine the most common class among the tuples.

(ii) Age and Salary have been discretized into intervals. You can consider them like ordinal attributes. When trying multi-splitting, you can merge values by their closeness. For example, if you have a three-way split of age, you can have [26-30] at one branch, [31 35] at one, and [36 40] [41 45] [46 50] at one. It is ok as long as you try a number of reasonable splits

Step 1 : Choose the root node

- It is an alteration of the information gain that reduces its favouritism on high-branch attributes.

Scanned by CamScanner

- Gain ratio should be big when data is evenly spread and small when all data belong to one branch. So, it considers number of branches and size of branches when it selects attribute to split.
- If a data set T contains examples from n classes,

$$\text{Gain (A)} = 1 - \sum_{j=1}^{n} p_j^2$$

Where, p_j is the relative frequency of class j in A.

- Intrinsic information (SplitInfo) is the entropy of distribution of instances into branches.

$$\text{SplitInfo (S, A)} = -\sum \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

- Gain ratio normalizes info gain by using following formula:

$$\text{Gain Ratio (S,A)} = \frac{\text{Gain(S,A)}}{\text{SplitInfo(S,A)}}$$

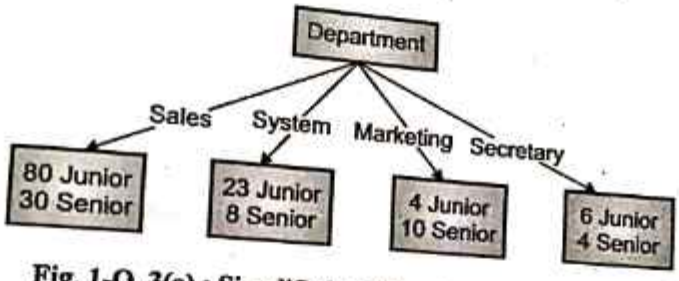


Fig. 1-Q. 3(a) : Simplified decision tree for Department

$$\text{GainRATIO} = \frac{\text{GAIN}}{\text{SplitINFO}}$$

$$\text{GAIN} = \left(1 - \left(\frac{113}{165}\right)^2 - \left(\frac{52}{165}\right)^2\right)$$

$$- \frac{110}{165}\left(1 - \left(\frac{80}{110}\right)^2 - \left(\frac{30}{110}\right)^2\right) - \frac{31}{165}\left(1 - \left(\frac{23}{31}\right)^2 - \left(\frac{8}{31}\right)^2\right)$$

$$- \frac{14}{165}\left(1 - \left(\frac{4}{14}\right)^2 - \left(\frac{10}{14}\right)^2\right) - \frac{10}{165}\left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2\right) = 0.4316 - 0.4001 = 0.0315$$

$$\text{SplitINFO} = -\left(\frac{110}{165}\log\frac{110}{165} + \frac{31}{165}\log\frac{31}{165} + \frac{14}{165}\log\frac{14}{165} + \frac{10}{165}\log\frac{10}{165}\right) = 0.9636$$

$$\text{GainRATIO} = \frac{0.0315}{0.9636} = 0.0327$$

Scanned by CamScanner

For Age

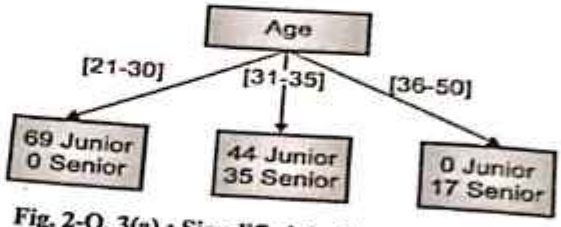


Fig. 2-Q. 3(a) : Simplified decision tree for Age.

$$\text{Gain RATIO} = \frac{\text{GAIN}}{\text{Split INFO}}$$

$$\text{GAIN} = \left(1-\left(\frac{113}{165}\right)^2-\left(\frac{52}{165}\right)^2\right) - \frac{69}{165}\left(1-\left(\frac{69}{69}\right)^2-\left(\frac{0}{69}\right)^2\right) - \frac{79}{165}\left(1-\left(\frac{44}{79}\right)^2-\left(\frac{35}{79}\right)^2\right) - \frac{17}{165}\left(1-\left(\frac{0}{17}\right)^2-\left(\frac{17}{17}\right)^2\right) = 0.4316-0.2363$$

$$= 0.1953$$

$$\text{SplitINFO} = -\left(\frac{69}{165}\log\frac{69}{165}+\frac{79}{165}\log\frac{79}{165}+\frac{17}{165}\log\frac{17}{165}\right) = 0.9513$$

$$\text{GainRATIO} = \frac{0.1953}{0.9513} = 0.2053$$

For Salary (merging values by their closeness, as mentioned in the Hint, and having so three branches)

For Salary

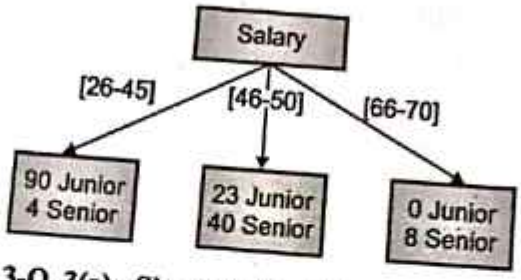


Fig. 3-Q. 3(a) : Simplified decision tree for salary.

$$\text{GAIN} = \left(1-\left(\frac{113}{165}\right)^2-\left(\frac{52}{165}\right)^2\right) - \frac{94}{165}\left(1-\left(\frac{90}{94}\right)^2-\left(\frac{4}{94}\right)^2\right) - \frac{63}{165}\left(1-\left(\frac{23}{63}\right)^2-\left(\frac{40}{63}\right)^2\right) - \frac{8}{165}\left(1-\left(\frac{0}{8}\right)^2-\left(\frac{8}{8}\right)^2\right)$$

$$= 0.4316-0.2234$$

$$= 0.2082$$

Scanned by CamScanner

$$\text{SplitINFO} = -\left(\frac{94}{165}\log\frac{94}{165} + \frac{63}{165}\log\frac{63}{165} + \frac{8}{165}\log\frac{8}{165}\right) = 0.8349$$

$$\text{GainRATIO} = \frac{0.2082}{0.8349} = 0.2494$$

The biggest GainRATIO is obtained with Salary, therefore Salary is the root node

For Department

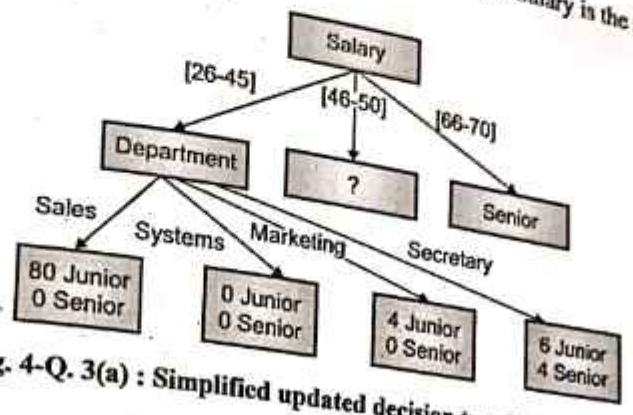


Fig. 4-Q. 3(a) : Simplified updated decision tree for Department

$$\text{GAIN} = \left(1-\left(\frac{90}{94}\right)^2-\left(\frac{4}{94}\right)^2\right)-\frac{80}{94}\left(1-\left(\frac{80}{80}\right)^2-\left(\frac{0}{80}\right)^2\right)$$
$$-\frac{4}{94}\left(1-\left(\frac{4}{4}\right)^2-\left(\frac{0}{4}\right)^2\right)-\frac{10}{94}\left(1-\left(\frac{6}{10}\right)^2-\left(\frac{4}{10}\right)^2\right)$$
$$= 0.0815 - 0.051$$
$$= 0.0305$$

$$\text{SplitINFO} = -\left(\frac{80}{94}\log\frac{80}{94} - \frac{4}{94}\log\frac{10}{94} + \frac{10}{94}\right) = 0.5099$$

$$\text{GainRATIO} = \frac{0.0305}{0.5099} = 0.0598$$

For Age

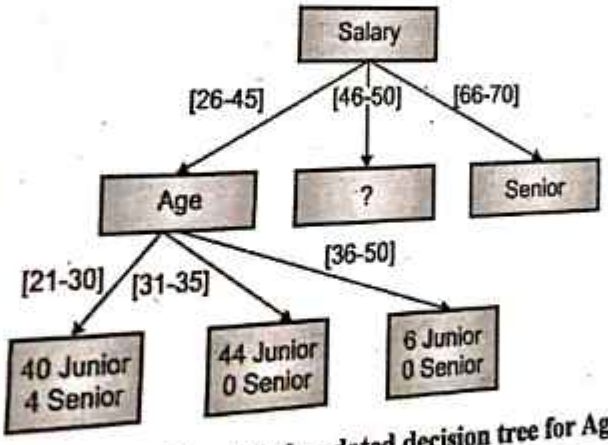


Fig. 5-Q. 3(a) : Simplified updated decision tree for Age

Scanned by CamScanner

$$GAIN = \left(1-\left(\frac{90}{94}\right)^2-\left(\frac{4}{94}\right)^2\right)-\frac{44}{94}\left(1-\left(\frac{40}{44}\right)^2-\left(\frac{44}{94}\right)^2\right)$$
$$-\frac{44}{94}\left(1-\left(\frac{44}{44}\right)^2-\left(\frac{0}{44}\right)^2\right)-\frac{6}{94}\left(1-\left(\frac{6}{6}\right)^2-\left(\frac{0}{6}\right)^2\right)$$
$$= 0.0815-0.0773$$
$$= 0.0042$$

$$SplitINFO = -\left(\frac{80}{94}\log\frac{80}{94}-\frac{4}{94}\log\frac{10}{94}+\frac{10}{94}\right)=0.8863$$

$$GainRATIO = \frac{0.0042}{0.8863}=0.0047$$

The biggest GainRATIO is obtained with Department, therefore Department is the connection with Salary = [26 – 45].

Notes : For this salary level we don't have samples for Department = Systems. For Department = Secretary the connections to two levels of Age are automatically, as shown in Fig. 6-Q. 3(a).

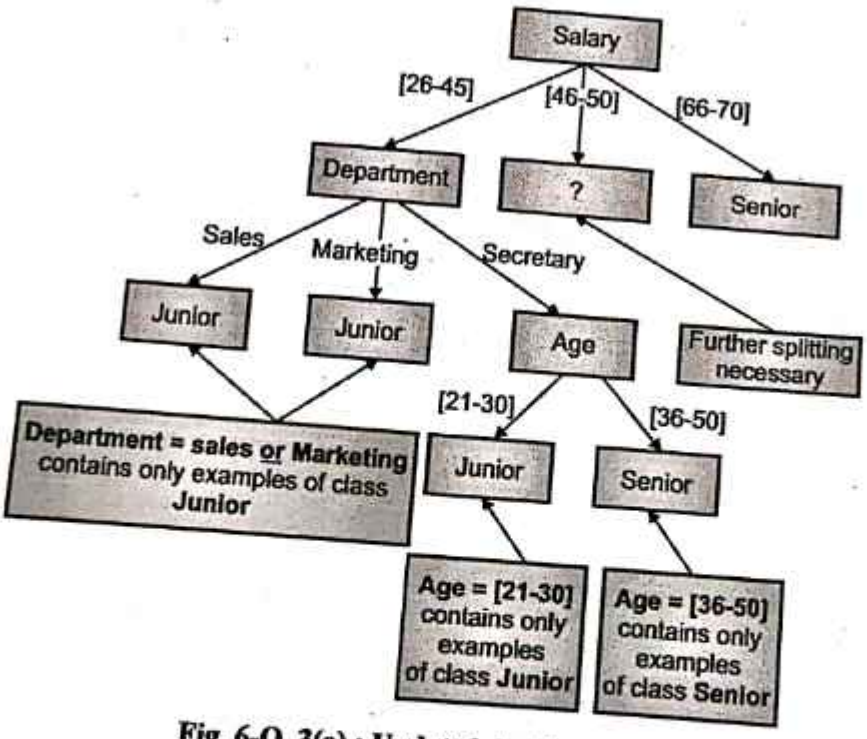


Fig. 6-Q. 3(a) : Updated decision tree

When we connect the node Age in Salary = [40 – 50] we already arrive in leaf nodes for all the age classes. Therefore, this is our final tree (Fig. 7 - Q. 3(a))

Scanned by CamScanner

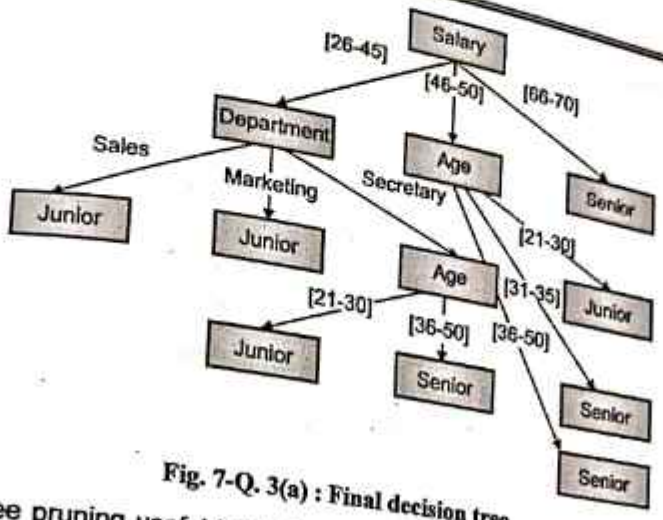


Fig. 7-Q. 3(a) : Final decision tree.

Q. 3(b) Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning? Given a decision tree, you have the option of (i) converting the decision tree to rules and then pruning the resulting rules, or (ii) pruning the decision tree and then converting the pruned tree to rules. What advantage does (i) have over (ii)? **(10 Marks)**

Ans. :

Tree Pruning

- Because of noise or outliers, the generated tree may over fit due to many branches.
- To avoid over fitting, prune the tree so that it is not too specific.

Prepruning

- Start pruning in the beginning while building the tree itself.
- Stop the tree construction in early stage.
- Avoid splitting a node by checking the threshold with the goodness measure falling below a threshold.
- Selection of correct threshold is difficult in prepruning.

Postpruning

- Build the full tree then start pruning, remove the branches.
- Use different set of data than training data set to get the best pruned tree.
- The decision tree built may overfit the training data. There could be too many branches, some of which may reflect anomalies in the training data due to noise or outliers. Tree pruning addresses this issue of overfitting the data by removing the least reliable branches

Scanned by CamScanner

(using statistical measures). This generally results in a more compact and reliable decision tree that is faster and more accurate in its classification of data.

- The drawback of using a separate set of tuples to evaluate pruning is that it may not be representative of the training tuples used to create the original decision tree. If the separate set of tuples are skewed, then using them to evaluate the pruned tree would not be a good indicator of the pruned tree's classification accuracy. Furthermore, using a separate set of tuples to evaluate pruning means there are less tuples to use for creation and testing of the tree. While this is considered a drawback in machine learning, it may not be so in data mining due to the availability of larger data sets.
- If pruning a subtree, we would remove the subtree completely with method (b). However, with method (a), if pruning a rule, we may remove any precondition of it. The latter is less restrictive.

Q. 4(a) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters. A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9). The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the k-means algorithm to show only

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

(ii) The final three clusters **(Example 4.8.6)** **(10 Marks)**

Q. 4(b) Briefly outline with example, how to compute the dissimilarity between objects described by the following :

(i) Nominal Attributes **(10 Marks)**

(ii) Asymmetric binary attributes

Ans. :

(i) Nominal Attributes

- Nominal attributes are also called as Categorical attributes and allow for only qualitative classification.
- Every individual item has a certain distinct categories, but quantification or ranking the order of the categories is not possible.
- The nominal attribute categories can be numbered arbitrarily.
- Arithmetic and logical operations on the nominal data cannot be performed.
- Typical examples of such attributes are:

Scanned by CamScanner

Car Owner :	1. Yes 2. No
Employment Status :	1. Unemployed 2. Employed

- The dissimilarity computed between objects described by nominal attributes is that the dissimilarity between two objects i and j can be computed based on the ratio of mismatches: $d(i, j) = (p - m)/p$, where m is the number of matches (i.e., the number of attributes for which i and j are in the same state), and p is the total number of attributes describing the objects.

(ii) Asymmetric binary attributes

- If the outcomes of the states are not equally important.An example of such a variable is the presence or absence of a relatively rare attribute. For example : Person is "handicapped or not handicapped". The most important outcome is usually coded as 1 (present) and the other is coded as 0 (absent).
- Object dissimilarity can be computed for objects described by nominal attributes and by asymmetric binary attributes are as follows:
- Measures of dissimilarity are used in data mining applications such as clustering, outlier analysis, and nearest-neighbour classification. Examples include the Jaccard coefficient for asymmetric binary attributes and Euclidean, Manhattan, Minkowski, and supremum distances for numeric attributes.

Q. 5(a) Frequent pattern mining algorithms considers only distinct items in a transaction. However, multiple occurrences of an item in the same shopping basket, such as four cakes and three jugs of milk, can be important in transactional data analysis. How can one mine frequent item sets efficiently considering multiple occurrences of items? Generate Frequent Pattern Tree for the following transaction with 30% minimum support :

(10 Marks)

Transaction ID	Items
T1	E, A, D, B
T2	D, A, C, E, B

Scanned by CamScanner

Transaction ID	Items
T3	C, A, B, E
T4	B, A, D
T5	D
T6	D, B
T7	A, D, E
T8	B, C

Ans. :

Step 1 : Find support for each item

Item	Support
A	5
B	6
C	3
D	6
E	4

Step 2 : Consider items with min. support = 30% = 3.

As all items have support > = 3, consider all items and arrange the transaction items in descending order i.e. B, D, A, E, C.

Transaction ID	Items Boughts	Ordered Frequent Items
T1	E, A, D, B	B, D, A, E
T2	D, A, C, E, B	B, D, A, E, C
T3	C, A, B, E	B, A, E, C
T4	B, A, D	B, D, A
T5	D	D
T6	D, B	B, D

Scanned by CamScanner

Transaction ID	Items Boughts	Ordered Frequent Items
T7	A, D, E	D, A, E
T8	B, C	B, C

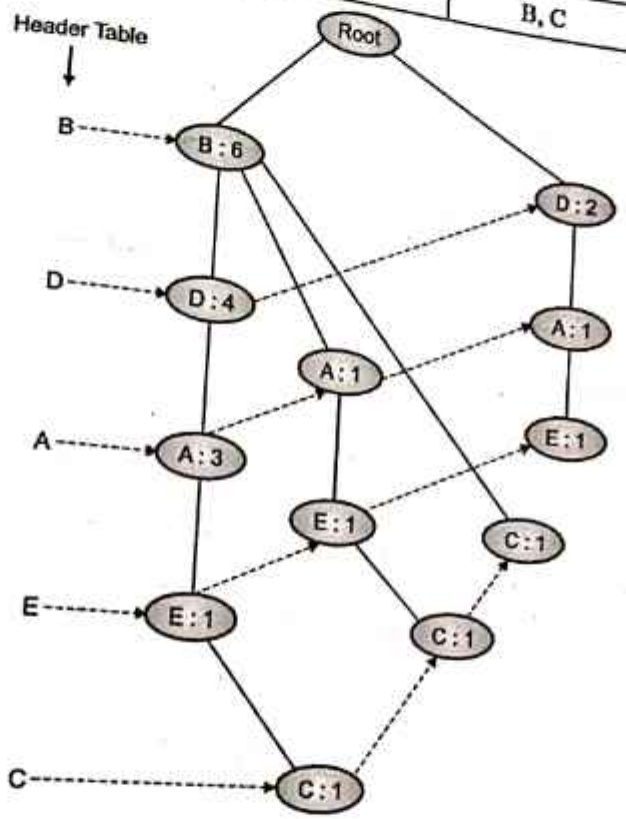


Fig. 1-Q. 5(a) : FP – Tree

Q. 5(b) Differentiate between simple linkage, average linkage and complete linkage algorithms. Use complete linkage algorithm to find the clusters from the following dataset. **(10 Marks)**

X	4	8	15	24	24
Y	4	4	8	4	12

Scanned by CamScanner

Ans. :

Single linkage : Minimum distance is considered

Complete linkage : Maximum distance is considered

Average linkage : Average distance is considered

Calculate Distance Matrix using Euclidean distance formula

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	4	0			
P_3	11.704	8.06	0		
P_4	20	16	9.848	0	
P_5	21.54	17.888	9.848	8	0

Distance matrix after merging P_1 and P_2

	(P1,P2)	P3	P4	P5
(P1,P2)	0			
P3	8.06	0		
P4	16	9.848	0	
P5	17.888	9.848	8	0

Distance matrix after merging P_4 and P_5

	(P_1,P_2)	P_3	(P_4,P_5)
(P_1,P_2)	0		
P_3	8.06	0	
(P_4,P_5)	16	9.848	

Distance matrix after merging (P_1, P_2) and P_3

	(P_1, P_2, P_3)	(P_4, P_5)
(P_1, P_2, P_3)	0	
(P_4, P_5)	9.848	0

Scanned by CamScanner

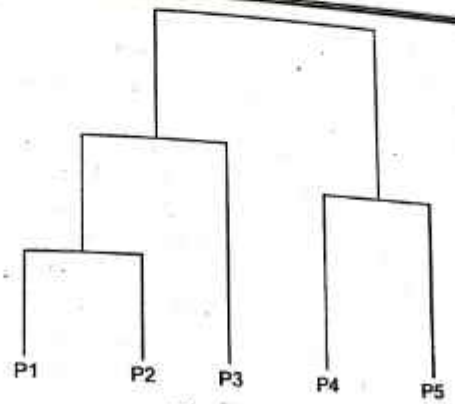


Fig. 1-Q. 5(b)

Q. 6(a) Data quality can be assessed in terms of several issues, including accuracy, completeness, and consistency. For each of the above three issues, discuss how data quality assessment can depend on the intended use of data, giving examples. Propose two other dimensions of data quality. **(10 Marks)**

Ans. :

Other dimensions that can be used to assess the quality of data include timeliness, believability, value added, interpretability and accessibility, described as follows:

- **Timeliness :** Data must be available within a time frame that allows it to be useful for decision making.
- **Believability :** Data values must be within the range of possible results in order to be useful for decision making.
- **Value added :** Data must provide additional value in terms of information that offsets the cost of collecting and accessing it.
- **Interpretability :** Data must not be so complex that the effort to understand the information it provides exceeds the benefit of its analysis.
- **Accessibility :** Data must be accessible so that the effort to collect it does not exceed the benefit from its use.

Q. 6(b) Present an example where data mining is crucial to the success of a business. What data mining functions does this business need? Can they be performed alternatively by data query processing or simple statistical analysis? **(10 Marks)**

Ans. :

- A suitable example could be found from practically any business that sells items or services. Such business would require both cross-market analysis (Finding associations between product sales) and customer profiling (what types of customers buy what products). Based on the acquired profiles predictions can be made on what kind of marketing strategies would be most effective.

Scanned by CamScanner

- In theory this knowledge can be acquired with data query processing or simple statistical analysis, but it would require a considerable amount of manual work by expert market analysts, both in order to decide which queries to use or how to interpret the statistics and due to the huge amount of data.
- A department store, for example, can use data mining to assist with its target marketing mail campaign. Using data mining functions such as association, the store can use the mined strong association rules to determine which products bought by one group of customers are likely to lead to the buying of certain other products. With this information, the store can then mail marketing materials only to those kinds of customers who exhibit a high likelihood of purchasing additional products.
- Data query processing is used for data or information retrieval and does not have the means for finding association rules. Similarly, simple statistical analysis cannot handle large amounts of data such as those of customer records in a department store.

Scanned by CamScanner