

**A COMPREHENSIVE APPROACH TOWARDS
BUILDING OF AN EFFICIENT INFORMATION
DELIVERY SYSTEM**

THESIS

Submitted in fulfilment of the requirement of the degree of

**DOCTOR OF PHILOSOPHY
COMPUTER ENGINEERING**

to

YMCA UNIVERSITY OF SCIENCE & TECHNOLOGY

by

NAVEEN DAHIYA

Registration No. YMCAUST/Ph. 56/2K11

Under the Supervision of

**DR. MANJEET SINGH
(SUPERVISOR)
ASSOCIATE PROFESSOR
DEPT. OF CE
YMCAUST, FARIDABAD**

**DR. VISHAL BHATNAGAR
(CO-SUPERVISOR)
ASSOCIATE PROFESSOR
DEPT. OF CSE
AIACT&R, GEETA COLONY, DELHI**



**Department of Computer Engineering
Faculty of Engineering and Technology
YMCA University of Science & Technology
Sector-6, Mathura Road, Faridabad, Haryana, India**

AUGUST, 2015

DEDICATED

To

My Family

DECLARATION

I hereby declare that this thesis entitled **A COMPREHENSIVE APPROACH TOWARDS BUILDING OF AN EFFICIENT INFORMATION DELIVERY SYSTEM** by **NAVEEN DAHIYA**, being submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy in **COMPUTER ENGINEERING** under Faculty of Engineering and Technology of **YMCA University of Science & Technology Faridabad**, during the academic year 2015, is a bonafide record of my original work carried out under guidance and supervision of **DR. MANJEET SINGH, ASSOCIATE PROFESSOR, DEPARTMENT OF COMPUTER ENGINEERING (SUPERVISOR), YMCAUST, FARIDABAD** and **DR. VISHAL BHATNAGAR, ASSOCIATE PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE ENGINEERING (CO SUPERVISOR), AIACT&R, DELHI** and has not been presented elsewhere.

I further declare that the thesis does not contain any part of any work which has been submitted for the award of any degree either in this university or in any other university.

(Naveen Dahiya)

Registration No-YMCAUST/Ph.56/2011

CERTIFICATE

This is to certify that this thesis entitled **A COMPREHENSIVE APPROACH TOWARDS BUILDING OF AN EFFICIENT INFORMATION DELIVERY SYSTEM** by **NAVEEN DAHIYA**, submitted in fulfilment of the requirement for the Degree of Doctor of Philosophy in **COMPUTER ENGINEERING** under Faculty of Engineering & Technology of YMCA University of Science & Technology Faridabad, during the academic year 2015, is a bonafide record of work carried out under our guidance and supervision.

We further declare that to the best of our knowledge, the thesis does not contain any part of any work which has been submitted for the award of any degree either in this university or in any other university.

(Signature of Supervisor)

Dr. Manjeet Singh

ASSOCIATE PROFESSOR

Department of Computer Engineering

Faculty of Engineering

YMCA University of Science & Technology Faridabad,

(Signature of Co-Supervisor)

Dr. Vishal Bhatnagar

ASSOCIATE PROFESSOR

Department of Computer Science & Engineering

AIACT&R, Geeta Colony, Delhi.

Dated: 08-08-2015

ACKNOWLEDGEMENT

Foremost, I have no words to express my gratitude to almighty GOD, for showering his blessings to provide the will power, courage, sincerity and peace of mind to complete the research work.

I would like to express my sincere gratitude to my supervisors **Dr. Vishal Bhatnagar** and **Dr. Manjeet Singh** for giving me the opportunity to work in this area. It would never have been possible for me to pursue the research work to this level without their constant motivation, inexpressible support and innovative ideas.

My special thanks to **Dr. Naresh Chauhan**, Chairman, YMCAUST, Faridabad for his kind wishes, support and encouragement. I am also very much thankful to **Dr. Komal Kumar Bhatia**, Associate Professor, for his persistent encouragement and timely helps.

I wish to offer my thanks to all the faculty members, staff members, friends and fellow colleagues of Maharaja Surajmal Institute of Technology, New Delhi as well as of YMCAUST, Faridabad especially **Mr. Sushil Kumar**, Reader and **Mr. Pradeep Sangwan**, HoD, ECE Department, Maharaja Surajmal Institute of Technology, New Delhi whose kind wishes and moral support were always with me and will always be there with me in future as well.

My special thanks to my dear kids Srishti & Bhavay, my wife Shefali for their constant motivation, patience, compromises and support. Last but not the least, I am thankful to my parents, in-laws, my sister for their love and encouragement for carrying out Ph.D.

(Naveen Dahiya)

Registration No-YMCAUST/ Ph. 56/2011

Abstract

The research work carried out presents comprehensive approach towards building of an efficient information delivery system. An efficient information delivery system is a system that delivers right/correct information to the users in a real/near real time for strategic decision making in a business scenario which is generally available on the device of their choice such as PC (personal computer), laptops and mobile sets. In the development of an information delivery system, there are three major components: i) data acquisition, ii) design of a data warehouse and iii) data extraction system. These components in turn do have many sub-components. The present work focuses on quality evaluation of conceptual models which is a sub-component of data warehouse design and information extraction aspect of information delivery systems. The quality of conceptual models can be evaluated using quality metrics that are based on size and structural complexities of conceptual models. The quality metrics can evaluate the quality of conceptual models along certain parameters like understandability, efficiency, effectiveness. The current research work aims towards quality evaluation of design at conceptual level which follows a stepwise approach. The first step towards quality evaluation of conceptual models is proposal of new quality metric NRFD (number of relations between fact and dimensions). The metric is theoretically validated to prove its practical utility and relevance towards quality evaluation of data warehouse conceptual models.

The theoretical validation is followed by empirical validation carried out using a controlled experiment in which 22 conceptual models are used and 80 subjects participate for a total of 13 quality metrics. All the volunteers had adequate knowledge of data warehousing and UML concepts because they were studying the subject 'Data Warehousing and Data Mining' as a part of their B. Tech. course curriculum in third year. The participation of the subjects was taken up voluntarily. The experiment was conducted in two separate rooms with strength of 40 students in each room. A supervisor was appointed for monitoring the students in each room. Before the start of experiment the students were given a description of the tasks to be performed and a tutorial to brush up their related concepts. A sample model was taken, calculation of values of metrics from model was shown, how the questions

were to be answered for the sample, where and in what format the answers were to be placed, where and in what format the starting time/ending time of the tasks were to be recorded. The students were given the printed hard copy of samples along with questionnaires in a variable order. The seating arrangement in each room was such that every two consecutive student had different set of models to be answered. Each of the alternate students was given a set 10 models first and the other was given other 12 models. A wall clock was installed in each room for recording the start and end time of each task. The time taken by each participant for answering the tasks of each model was recorded and gathered. From the collected data, average time for each model was calculated, which acts as understanding time for each conceptual model.

We proposed several hypotheses for quality evaluation of conceptual models. The proposed hypotheses are as follows:

- Null Hypothesis H_{01} : Quality metrics have no impact/contribution towards prediction of understandability of conceptual data warehouse models.
- Null Hypothesis H_{02} : All the principal components of the model summary are significant to predict the understandability of models.
- Null Hypothesis H_{03} : The models having similar values of quality metrics do not have any relation in respect of their understanding times.
- Alternate Hypothesis H_{01} : Quality metrics have significant effect/contribution towards prediction of understandability of conceptual data warehouse models.
- Alternate Hypothesis H_{02} : Not all the principal components of the model summary are significant to predict the understandability of models.
- Alternate Hypothesis H_{03} : The models having similar values of quality metrics have significant relation in respect of their understanding times.

Empirical validation techniques namely correlation, regression, principal component analysis, nearest neighbour analysis are applied. The results of empirical validation prove that several metrics including the one newly proposed have significant effect towards quality evaluation of conceptual data warehouse models i.e. understandability of conceptual models.

In order to measure the individual effect of a quality metric towards evaluation of quality of conceptual models along parameters namely understandability, efficiency

and effectiveness the metrics are ranked using a fuzzy multi-criteria ranking methodology based on expert opinion. The opinions of experts are recorded in a pre-defined format in the form of fuzzy linguistic variables for further processing. The criteria are defined qualitatively and the significance of quality metrics along the criteria varies according to user requirements, situations and expert opinion. This is the need of a fuzzy based system that can deal with imprecise and qualitative (non-numeric) data based on actual human (expert) decision making. Ranking of metrics along variable criteria (understandability, efficiency and effectiveness) lead to multiple-criteria decision making problem which is specified as follows:

A team of n experts ($E_1, E_2, E_3, \dots, E_n$), has to analyse and grant weights to k criteria ($C_1, C_2, C_3, \dots, C_k$) and the ratings to m quality metrics ($Q_1, Q_2, Q_3, \dots, Q_m$) for each of the k criteria. Let W_{ij} ($i=1,2,3,\dots,k; j=1,2,3,\dots,n$) be the weight assigned to criteria C_i by expert E_j . Let R_{ijt} ($i=1,2,3,\dots,m; j=1,2,3,\dots,n; t=1,2,3,\dots,k$) be the rating given to metric Q_i by expert E_j under criteria C_t .

The results of the proposed methodology are compared with the results of aggregated expert opinion, to prove their correctness. In the study, five experts from data warehouse domain having up to date knowledge of technological advances and rich practical hands on experience, with more than 10-20 years of experience are selected. Out of five experts three are from academics and two are from software industry. The academic experts are chosen as they have good experimental knowledge and are well acquainted with up to date technical advancements. The industrial experts have good insights into issues related to cost and benefits. To predict the understandability of data warehouse conceptual models, efforts are involved towards design of conceptual models, identification of subjects, preparation of questionnaires based on structural properties of models, collection of data in the form of time and then further aggregation of collected data. To minimize the efforts involved in prediction of understanding time, the need for a system that could predict the understanding time of conceptual models arise. The aim of minimizing the efforts involved in prediction of understanding time is achieved by creating a fuzzy rule base (based on expert opinion and ranking of metrics) to predict the understandability of conceptual models. The values of quality metrics are given as input to the system and understanding time is taken as output. The predicted results about understandability are compared with

actual results obtained by controlled experiment conducted, as mentioned above. The predicted results are highly accurate and hence show the significance and relevance of the automatic prediction system developed.

Efficient information extraction has also been identified as one of the main components towards development of an efficient information delivery system. All the information delivery systems are supported by huge data warehouses at back end. Complex queries run on data warehouses and results related to extraction of strategic decision making are obtained. Query response time is one of the major factors affecting the quality of data warehouses. Thus query optimization needs to be achieved for efficient information extraction from data storehouses towards building of efficient information delivery systems. This issue is also taken care of in the current research work and it is found that one of the major factors affecting query optimization is optimal selection of materialized views. Few terms must be known while discussing materialized view selections which are as follows:

- View: A derived relation/result in response to a query. It is defined in terms of base relation and/or combination of attributes. Each cell in multidimensional cubes forms a view.
- Materialized View: A view is materialized if its result in response to query is stored in memory. It is the set of materialized views whose optimal selection improves query optimization.
- View Selection: It aims at selecting a set of materialized views given some database to optimize query response time. The optimal view selection improves query response time and is one of the main factors affecting query optimization of decision support systems.

Regarding information extraction aspect of data warehouse development, we present a refined greedy view selection approach. The greedy approach for view selection is taken up as the base approach for reference. We enhance the basic greedy approach to a refined greedy selection approach using forward references to give a better selection of views. The same is proved using experimental results. The view selection is further enhanced by including space constraints to the results of greedy and refined greedy approach using knapsack implementation.

TABLE OF CONTENTS

| | |
|--|------|
| Candidate's Declaration | i |
| Certificate of the Supervisor | ii |
| Acknowledgement | iii |
| Abstract | iv |
| Table of Contents | viii |
| List of Tables | xii |
| List of Figures | xiv |
| List of Abbreviations | xvi |
| CHAPTER I INTRODUCTION | |
| 1.1 EFFICIENT INFORMATION DELIVERY SYSTEMS (EIDS) | 1 |
| 1.2 NEED FOR BUILDING EFFICIENT DATA WAREHOUSE | 2 |
| 1.3 DESIGN OF EFFICIENT DATA WAREHOUSE | 2 |
| 1.3.1 Conceptual Design Phase | 3 |
| 1.3.2 Logical Design Phase | 4 |
| 1.3.3 Physical Design Phase | 4 |
| 1.4 SIGNIFICANCE AND RELEVANCE OF CONCEPTUAL DESIGN PHASE | 5 |
| 1.5 CONCEPTUAL MODEL AND QUALITY METRICS | 7 |
| 1.6 QUALITY METRICS: ORDERING APPROACH | 9 |
| 1.7 A FUZZY RULE BASE SYSTEM FOR PREDICTING THE QUALITY OF CONCEPTUAL MODEL | 10 |
| 1.8 EFFICIENT INFORMATION EXTRACTION | 11 |
| 1.9 PROBLEM DEFINITION | 12 |
| 1.10 ORGANIZATION OF THESIS | 13 |
| CHAPTER II LITERATURE REVIEW | |
| 2.1 INTRODUCTION | 15 |
| 2.2 DATA WAREHOUSE DEVELOPMENT | 15 |
| 2.3 QUALITY EVALUATION OF CONCEPTUAL DATA WAREHOUSE MODELS | 21 |
| 2.4 CLASSIFICATION AND ORDERING OF QUALITY METRICS | 26 |

| | |
|---|----|
| 2.5 EFFICIENT INFORMATION EXTRACTION | 28 |
| 2.6 PROBLEM DEFINITION: REVISED | 29 |
| 2.7 DATA COLLECTION AND ANALYSIS | 32 |
| 2.8 TOOLS | 33 |
| 2.9 UNITS USED FOR ANALYSIS | 34 |
| 2.10 THESIS OUTLINE | 34 |
| CHAPTER III THEORETICAL AND EMPIRICAL STUDY TOWARDS BUILDING OF EIDS | |
| 3.1 INTRODUCTION | 35 |
| 3.2 CONCEPTUAL FRAMEWORK FOR EFFICIENT DATA WAREHOUSE SYSTEM | 35 |
| 3.3 NRFD (NUMBER OF RELATIONS BETWEEN FACTS AND DIMENSIONS), NEW PROPOSED METRIC AND ITS THEORETICAL VALIDATION | 40 |
| 3.3.1 How the Idea Generated | 41 |
| 3.3.2 Importance of Proposed Metric | 42 |
| 3.3.3 Metric Creation | 43 |
| 3.3.4 Theoretical Validation | 46 |
| 3.4 EMPIRICAL VALIDATION | 49 |
| 3.4.1 Preliminaries | 50 |
| 3.5 EXPERIMENTAL SETUP | 51 |
| 3.5.1 Goal | 52 |
| 3.5.2 Model | 52 |
| 3.5.3 Subjects | 53 |
| 3.5.4 Hypothesis | 53 |
| 3.6 EMPIRICAL DATA COLLECTION | 54 |
| 3.6.1 Independent Variables | 54 |
| 3.6.2 Dependent Variables | 54 |
| 3.6.3 Data Validation | 54 |
| 3.7 RESULT ANALYSIS | 55 |
| 3.7.1 Correlation Analysis | 56 |
| 3.7.2 Regression Analysis | 57 |
| 3.7.3 Principal Component Analysis (PCA) | 62 |
| 3.7.4 Nearest Neighbour Analysis | 64 |

| | |
|--|-----|
| 3.7.5 ROC Classification | 66 |
| 3.8 THREATS TO VALIDITY AND LIMITATIONS | 68 |
| 3.9 SUMMARY | 70 |
| CHAPTER IV QUALITY EVALUATION BASED ON RANKING, INFERENCE APPROACH TOWARDS BUILDING OF EIDS | |
| 4.1 INTRODUCTION | 71 |
| 4.2 PRELIMINARIES | 71 |
| 4.2.1 Introduction to Fuzzy Sets | 71 |
| 4.2.2 Triangular Fuzzy Membership Functions | 72 |
| 4.2.3 Fuzzy Linguistic Terms and Variables | 73 |
| 4.2.4 Quality Metrics Ranking Problem and Fuzzy Solution | 75 |
| 4.2.5 Criteria Matrix | 75 |
| 4.2.6 Permanent of Matrix | 76 |
| 4.2.7 Expert Opinion Ranking Methodology | 77 |
| 4.3 RESEARCH METHODOLOGY | 77 |
| 4.3.1 Identification of Quality Metrics for Conceptual Data Warehouse Models | 78 |
| 4.3.2 Identification and Selection of Experts | 78 |
| 4.3.3 Selection of Ranking Criteria | 79 |
| 4.3.4 Fuzzy Evaluation and Formation of Criteria Matrix | 80 |
| 4.3.5 Calculating Permanent of Criteria Matrix | 80 |
| 4.3.6 Ranking of Metrics | 80 |
| 4.4 PRACTICAL APPLICATION | 80 |
| 4.5 RESULT ANALYSIS AND COMPARISON | 84 |
| 4.6 FUZZY RULE BASE FOR PREDICTING UNDERSTANDABILITY OF CONCEPTUAL MODELS | 87 |
| 4.6.1 Component Classification | 87 |
| 4.6.2 Stepwise Approach | 88 |
| 4.7 RESULT ANALYSIS | 93 |
| 4.8 SUMMARY | 95 |
| CHAPTER V EFFICIENT INFORMATION EXTRACTION TOWARDS BUILDING OF EIDS | |
| 5.1 INTRODUCTION | 97 |
| 5.2 ADVANTAGES OF LATTICE FRAMEWORK | 100 |

| | |
|--|-----|
| 5.3 OPTIMAL GREEDY SELECTION OF MATERIALIZED VIEWS | 100 |
| 5.3.1 Greedy Algorithm for View Selection | 101 |
| 5.3.2 Need for Refined Greedy Approach | 102 |
| 5.4 REFINED GREEDY ALGORITHM WITH FORWARD REFERENCING | 103 |
| 5.5 EXPERIMENTAL RESULTS | 105 |
| 5.6 COMPARISON AND ANALYSIS | 112 |
| 5.7 KNAPSACK IMPLEMENTATION FOR THE RESULTS OF GREEDY AND REFINED GREEDY ALGORITHM | 114 |
| 5.8 SUMMARY | 118 |
| CHAPTER VICONCLUSION AND FUTURE RESEARCH SCOPE | |
| 6.1 CONCLUSION | 119 |
| 6.2 FUTURE SCOPE | 121 |
| 6.2.1 Iterative Approach for Identification of New Metrics | 122 |
| 6.2.2 Wide Sampling for Generalization | 122 |
| 6.2.3 Application to Real Projects | 122 |
| 6.2.4 Extending Domain of Experts, Criteria and Ranking Methodologies | 122 |
| 6.2.5 Factor Expansion for Query Optimization | 123 |
| REFERENCES | 125 |
| BRIEF PROFILE OF RESEARCH SCHOLAR | 137 |
| PUBLICATIONS | 138 |

LIST OF TABLES

| Table | Title | Page No. |
|--------------|---|-----------------|
| Table 2.1 | Distribution of articles according to proposed Classification Model | 18 |
| Table 2.2 | Techniques for data warehouse development | 20 |
| Table 3.1 | Metric Values | 51 |
| Table 3.2 | Table of metrics for models | 52 |
| Table 3.3 | Descriptive Statistics for Understanding Time | 56 |
| Table 3.4 | Pearson correlation analysis | 57 |
| Table 3.5 | Univariate Linear Regression | 58 |
| Table 3.6 | Model Summary of Univariate Linear Regression | 59 |
| Table 3.7 | Multiple Regression | 60 |
| Table 3.8 | Model Summary of Understandability | 61 |
| Table 3.9 | Results of PCA | 62 |
| Table 3.10 | Rotated Component Matrix | 63 |
| Table 3.11 | Nearest Neighbour Analysis | 66 |
| Table 3.12 | Summary results of ROC | 68 |
| Table 4.1 | Fuzzy membership values for weights assigned to criteria | 74 |
| Table 4.2 | Fuzzy membership values for rating assigned to quality metrics | 74 |
| Table 4.3 | Fuzzy membership values and linguistic representation for ranking | 80 |
| Table 4.4 | Fuzzy membership values and linguistic representation for quality metrics | 81 |
| Table 4.5 | Aggregated weights for criteria ranking | 82 |
| Table 4.6 | Aggregated rating for quality metrics | 82 |
| Table 4.7 | Values of crisp scores for rating quality metrics | 83 |
| Table 4.8 | Ranking values and rank of quality metrics | 83 |
| Table 4.9 | Comparison and analysis with other technique | 84 |
| Table 4.10 | Input to rank based on expert opinion | 85 |
| Table 4.11 | Comparison based on various parameters | 86 |
| Table 4.12 | Fuzzy linguistic variables for metrics | 91 |
| Table 4.13 | Fuzzy linguistic variables for understanding time | 92 |
| Table 4.14 | Predicted vs Calculated Results | 93 |

| Table | Title | Page No. |
|--------------|-----------------------------------|-----------------|
| Table 5.1 | Greedy selection | 102 |
| Table 5.2 | Comparison of Benefits | 112 |
| Table 5.3 | Comparison of Cumulative Benefits | 112 |
| Table 5.4 | Comparison of Cumulative Space | 113 |

LIST OF FIGURES

| Figure | Title | Page No. |
|---------------|---|-----------------|
| Figure 1.1 | Data Warehouse Design Process | 5 |
| Figure 1.2 | Information and Data Warehouse Quality | 7 |
| Figure 1.3 | Measures of Quality Metrics | 8 |
| Figure 2.1 | Classification framework for efficient data warehouse design and development | 16 |
| Figure 2.2 | Bar plot | 19 |
| Figure 2.3 | Line graph | 21 |
| Figure 2.4 | Research Framework | 32 |
| Figure 3.1 | Layered approach for development of an efficient data warehouse system | 35 |
| Figure 3.2 | Proposed Detailed Research Framework for Development of Efficient Data Warehouse System | 38 |
| Figure 3.3 | Quality metrics along with proposed metric | 42 |
| Figure 3.4 | A conceptual model showing sales of items of a store | 44 |
| Figure 3.5 | A conceptual model showing sales of items of a store | 45 |
| Figure 3.6 | UML Class Diagram for manufacturing parts | 50 |
| Figure 3.7 | Scree plot of PCA | 63 |
| Figure 3.8 | Predictor space for selected model 2 | 64 |
| Figure 3.9 | Peer chart for selected model 2 | 65 |
| Figure 3.10 | ROC Curve plot | 67 |
| Figure 4.1 | Triangular fuzzy membership function graph | 73 |
| Figure 4.2 | Fuzzy membership graph for weighting criteria | 74 |
| Figure 4.3 | Fuzzy membership graph for rating quality metrics | 74 |
| Figure 4.4 | Research methodology | 77 |
| Figure 4.5 | Performa 1 | 79 |
| Figure 4.6 | Performa 2 | 79 |
| Figure 4.7 | Fuzzy inference system | 87 |
| Figure 4.8 | Fuzzification of NC | 89 |
| Figure 4.9 | Fuzzification of NA | 90 |
| Figure 4.10 | Fuzification of NH | 90 |
| Figure 4.11 | Fuzzification of DHP | 90 |

| Figure | Title | Page No. |
|---------------|--|-----------------|
| Figure 4.12 | Fuzzification of NRFD | 91 |
| Figure 4.13 | Fuzification of Understanding Time | 91 |
| Figure 4.14 | Fuzzy rule base | 92 |
| Figure 4.15 | Output of rule viewer for model no. 1 | 94 |
| Figure 5.1 | Lattice Framework 1 | 98 |
| Figure 5.2 | Lattice Framework 2 | 102 |
| Figure 5.3 | Example Lattice Framework | 103 |
| Figure 5.4 | Cube Materialization Form | 106 |
| Figure 5.5 | Node Entry Form | 106 |
| Figure 5.6 | Lattice Entry Form | 107 |
| Figure 5.7 | Lattice Storage Form | 107 |
| Figure 5.8 | Pass Entry Form | 108 |
| Figure 5.9 | Greedy Algorithm Pass by Pass Output | 108 |
| Figure 5.10 | Refined Greedy Algorithm Pass by Pass Output | 109 |
| Figure 5.11 | Benefits vs. Pass Graph | 113 |
| Figure 5.12 | Space vs. Pass Graph | 114 |
| Figure 5.13 | Space Constraint Form | 115 |
| Figure 5.14 | Knapsack Benefits for Greedy and Refined Greedy Approach | 116 |
| Figure 5.15 | Benefits vs. Size Graph | 116 |
| Figure 5.16 | Number of nodes selected vs. Size Graph | 117 |

LIST OF ABBREVIATIONS

| | |
|--------|--|
| ACM | Association for Computing Machinery |
| ANOVA | Analysis of Variance |
| DBMS | Data Base Management System |
| DF | Dimension Fact |
| EIDS | Efficient Information Delivery System |
| ER | Entity Relationship |
| SPSS | Statistical Package for Social Sciences |
| IEEE | Institution for Electrical and Electronics Engineers |
| MATLAB | Matrix Laboratory |
| NRFD | Number of Relations between Facts and Dimensions |
| OODM | Object Oriented Dimensional Model |
| OOMD | Object Oriented Multi-Dimensional Model |
| PCA | Principal Component Analysis |
| ROC | Receiver Operating Characteristic |
| UML | Unified Modelling Language |

CHAPTER I

INTRODUCTION

1.11 EFFICIENT INFORMATION DELIVERY SYSTEMS (EIDS)

Data is a valuable resource of any organization for strategic decision making. In today's world, every organization has a database storing data used for its daily operations as well as for strategic decision making. The organizations feel the necessity to transform data into valuable information. To achieve this aim of transforming data into valuable information and to extract the valuable information efficiently, the organizations feels the acute need of efficient information delivery systems (EIDS) for critical business decisions.

An efficient information delivery system helps in making critical decisions in an economic cost-effective manner, thereby giving organizations competitive advantage, increasing employee productivity and enhancing overall reputation. An efficient information delivery system plays a vital role in business intelligence/analytic strategy.

An efficient information delivery system extracts useful/right information from large repositories of data (internal data, external data and archived data) which are also part of information delivery system. Data repositories extract variable data from multiple sources and transform it in a form suitable for making strategic decisions. The data repositories are data warehouses. The basic components of an efficient delivery system are data, efficient data warehouses and efficient information extraction systems.

In this chapter, we first focus on various issues related to efficient data warehouses such as need of data warehouse, design of data warehouse, quality of data warehouse and efficient information extraction from the data warehouse towards building of an efficient information delivery system. Secondly, we present problem formulation followed by organization of thesis.

1.12 NEED FOR BUILDING EFFICIENT DATA WAREHOUSE

To gain competitive edge in the ever expanding, complex scenarios of the business world, the managers and executives became desperate for information that can be used for strategic decision making. The information for strategic decision making includes knowledge of variable trends in customer needs and preferences, emerging technologies, sales and marketing techniques, quality of products and services, market trends over a period of time. This arise the need of efficient data warehouses that are capable enough to provide strategic information and to handle fierce competition in the business world.

‘A data warehouse is a subject oriented, integrated, time variant and non-volatile collection of data elements in support of the management’s decision making processes’[1, 2, 3]

The data warehouses can effectively handle (integrate and transform) large amounts of variable scattered data residing on multiple platforms, having variable data structures, having separate origins and make it suitable for analysis towards strategic decision making process. Salient features of efficient data warehouse [4] are following:

- Designed for analytical tasks
- Data from multiple applications
- Direct interaction with users
- Contains current and historical data
- Availability of users to run queries and get online results
- Ease of use

The need for an efficient data warehouse led to the study of various design issues involved in its creation. The issues related to design of efficient data warehouse towards building of an efficient information delivery system are introduced in next section.

1.13 DESIGN OF EFFICIENT DATA WAREHOUSE

In the above section the need for building an efficient data warehouse was discussed. As the need of efficient data warehouse system was felt by organizations, efforts for

designing efficient data warehouse towards building of an efficient delivery system were simultaneously made. We know facts and dimensions are basic components of a data warehouse. Facts are generally the subjects, having numerical values that are analysed along various parameters called dimensions. Dimensions are non- numerical attributes along which facts are analysed.

The design of a data warehouse starts with the knowledge of expectations of users (strategic decision makers) from data warehouse, as to what task they want the data warehouse to perform. From these expectations the actual functioning of data warehouse can be summarized and also facts and dimensions for a particular data warehouse can be identified prior to the design of a data warehouse.

After identification of facts and dimensions, the process of data warehouse design is initiated. The design of data warehouse system is a three phase process starting with conceptual model design phase [5, 6, 7, 8] to logical model design phase and finally physical model design phase, as shown in Figure 1.1. It can be seen from Figure 1.1 that the conceptual design phase of data warehouse [9] lays the foundation for design of data warehouse system. The design techniques for each of the design phases of data warehouse development are given in following sub-sections:

1.3.1 Conceptual Design Phase

In a conceptual design phase entities and the relations between them are specified. No attributes or primary keys are specified in conceptual design phase. StarER, Multidimensional ER (ME/R), Object Oriented Multidimensional Model (OOMD), Dimension Fact Model (DF) are some of the models used for conceptual design phase [10].

- StarER: It combines the star structure (fact in middle surrounded by dimensions) with the constructs of ER diagrams to depict hierarchies in dimensions more effectively.
- Multidimensional ER: This technique of designing conceptual models includes ER constructs along with dimension hierarchies showing specialization hierarchies and operators such as roll up/down.

- Dimension Fact technique: Here the conceptual model is collection of tree structured fact models whose elements are facts, dimensions and hierarchies.
- Object oriented dimensional modeling: This conceptual modeling technique makes use of object oriented features like class, objects, specialization, generalization and derived attributes. UML is widely accepted language for object oriented design models.

All of the conceptual design models specify entity names and the relationships existing between various entities.

1.3.2 Logical Design Phase

A logical design phase specifies more level of details than conceptual design phase. All the entities, attributes, relationships between entities, primary keys for each entity, foreign keys specifying various relationships and normalization is performed at this level. Fact Constellation model, Snowflake model and Star model are models used for logical data warehouse design [10].

- Star model: This is simplest of logical model design model in which a fact table is surrounded by various smaller dimension tables.
- Fact constellation model: It is a set of star models with hierarchical links between fact tables. The links between fact tables gives the ability to drill across various factual levels.
- Snowflake model: It is a variant of star model in which all the hierarchies of dimensions are explicitly specified and dimensions are not denormalized.

All of the logical design models specify entity names, relationships, primary keys, foreign keys, surrogate keys along with the data structures of all the design elements.

1.3.3 Physical Design Phase

A physical design phase shows how the model is actually implemented. It shows table structures including column names, columns data type, column constraints, primary keys, foreign keys and relationships. This phase makes the actual implementation of the data warehouse using various tools and software. Queries are thrown to the data warehouse system and results obtained accordingly. The queries are executed using operations like aggregation, drill down, drill across, slicing, dicing. Various tools exist

for information extraction from data warehouse namely SPSS for statistical analysis, Weka/Clementine for data mining.

The focus of study was on conceptual design phase due to its utmost importance in data warehouse development which is introduced in next section.

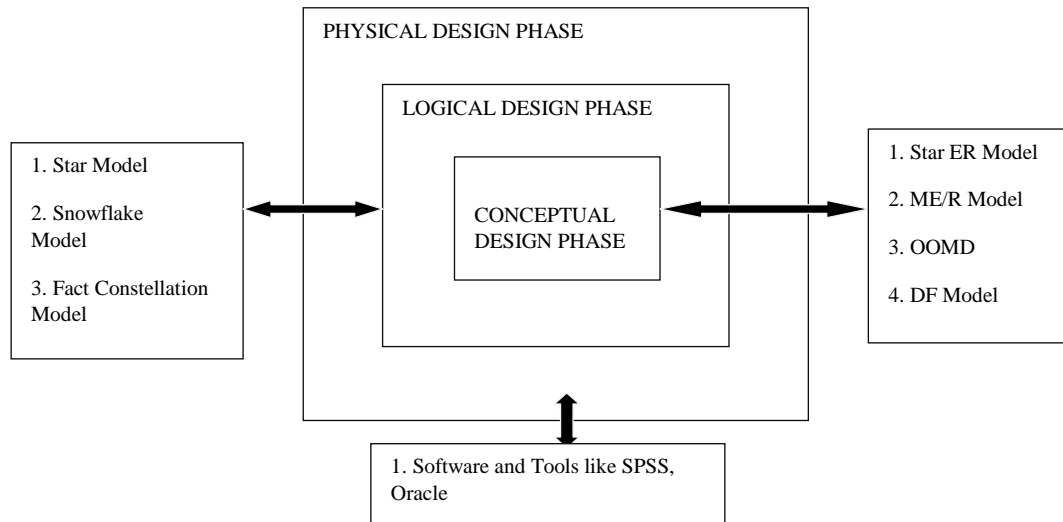


Figure 1.1 Data Warehouse Design Process

1.4 SIGNIFICANCE AND RELEVANCE OF CONCEPTUAL DESIGN PHASE

Conceptual phase is the initial phase in data warehouse design process as depicted in Figure 1.1 and stated in above section. It is always said that a good foundation of a building paves the way for a strong and lifelong sustainable building erected on it. Thus, a conceptual phase is like the foundation of a data warehouse design process. It is 100 times more economical to detect and rectify errors at conceptual level than later design phases [1, 2, 3]. The output of conceptual design phase is a conceptual model. A model is a simplified mathematical or non-mathematical representation of system containing information about customer needs and preferences, quality of products, sales trend and analysis as collected from users, business analysts and executives.

During the conceptual design phase, the end user requirements are transformed into abstract representations that are understandable to the end users and are independent of the implementations details [11]. But the representation is formal and complete that can be transformed into next phase (i.e. logical design phase) of data warehouse

design [12]. A good quality conceptual model leads to the development of a good quality logical and physical model.

A good conceptual data warehouse model should possess following characteristics [4]:

- A conceptual model should be user friendly in the sense that the user can understand at a glance what the model intends to present.
- It should be formal and complete. It should present the users with all the identified facts and dimensions along with the relationships between them.
- It should be software/hardware platform independent. It should be free of any concern regarding any hardware/software to read and understand it.
- The information collected from users should be stored in such an efficient manner that it can be transformed into next logical and physical models easily.

There exist various conceptual modelling design techniques like Star, StarER, Dimension Fact modelling, OOMD (introduced in section 1.2 above). Of all these design techniques OOMD is most preferred, as discussed by Mishra et al [10]. This approach uses UML (Unified Modelling Language) for designing conceptual models which takes advantages of object oriented properties to specify class diagrams of conceptual models such as inheritance, specialization/generalization, which are not taken into consideration by other approaches. It maps conceptual models to real world entities, which makes the models more realistic. Further it is more adaptable to constantly changing user requirements and models software reusable packages like class, objects, use cases. OODM approach for modelling conceptual models is used in this research study, based on its significance towards building of an efficient information delivery system.

As can be seen from Figure 1.1, the design process of data warehouse starts with conceptual design phase, the output of which is conceptual model. The conceptual model is input for logical design phase, whose output is logical model. The logical model is given as input to physical design phase, whose output is physical model. It can be visualized that better the quality of conceptual model, better will be the quality of logical and physical design model. [1, 2, 3]

Keeping in consideration significance of conceptual design phase/models, more in depth study was performed on improving the quality of conceptual design models towards building of an efficient information delivery system. It was observed during study that various quality metrics were significant enough to evaluate and improve the quality of data warehouse conceptual models. The quality metrics are introduced in next section.

1.5 CONCEPTUAL MODEL AND QUALITY METRICS

The previous section describes the significance of conceptual design phase in data warehouse design process. A good quality conceptual model leads to the development of efficient data warehouse systems [13]. Due to the significance of data warehouse in making strategic decisions, there is a need to assure the quality of multidimensional models at conceptual levels. Figure 1.2 shows hierarchical representation of various factors influencing data warehouse quality and leading to the building up of a good quality data warehouse. Information quality of an information delivery system depends on data warehouse quality and presentation quality [11]. Data warehouse quality depends on data quality, DBMS quality, data model quality. Data model quality depends on conceptual model quality, logical model quality and physical model quality. Figure 1.2 shows that quality of conceptual models can be measured along various quality metrics [14, 15].

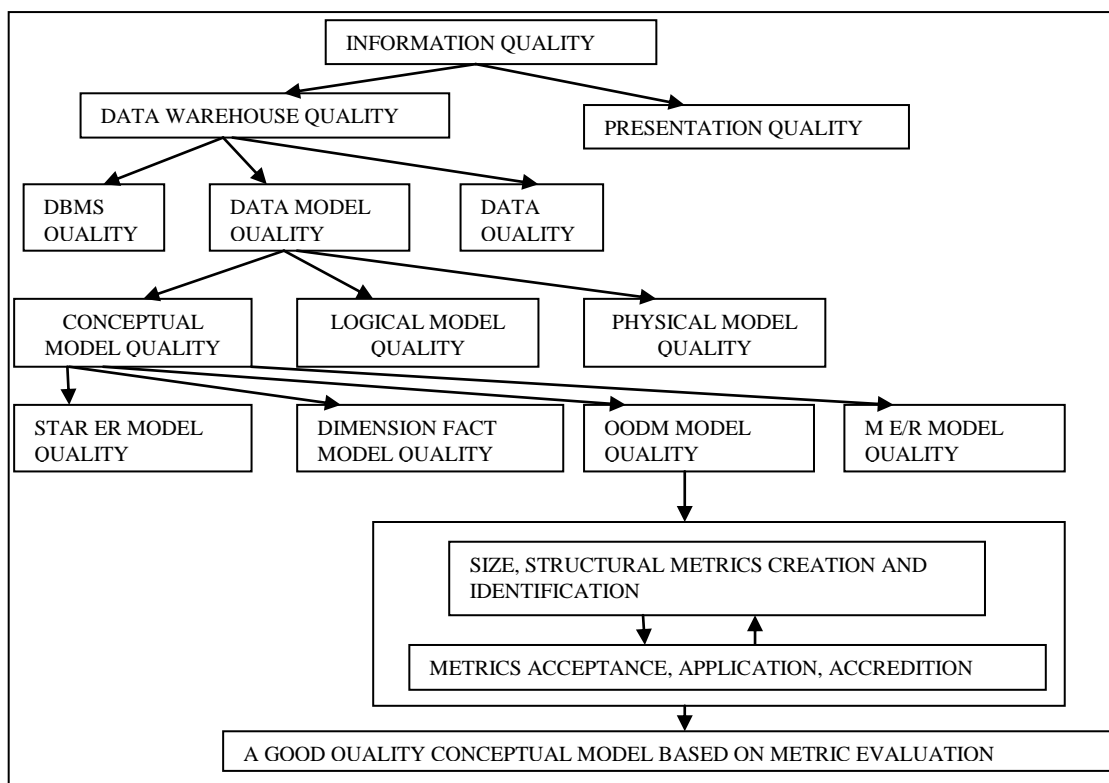


Figure 1.2 Information and Data Warehouse Quality

The metrics provides evaluation criteria[16] to judge the understandability (time taken to understand a conceptual model), efficiency (ability to accomplish task with minimum expenditure of time), relevance (measure of how closely an object matches user search for information) and effectiveness (capability of producing a desired result) of data warehouse conceptual models as can be seen from Figure 1.3.

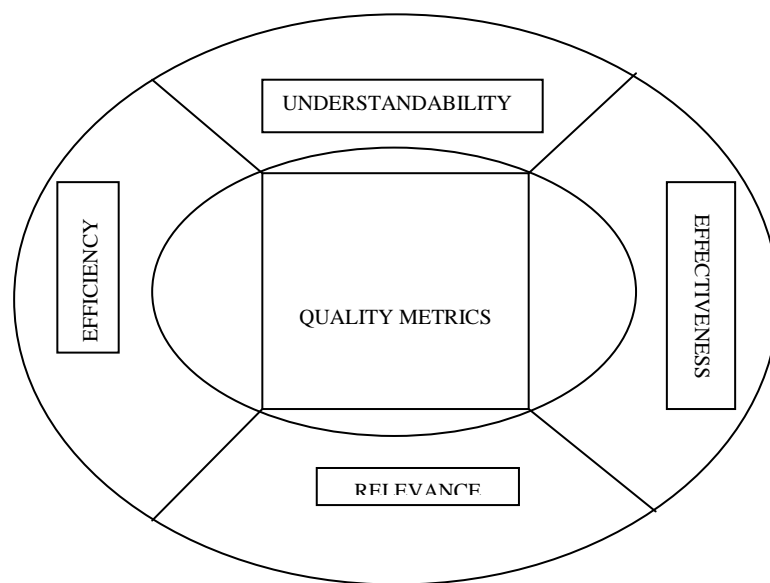


Figure 1.3 Measures of Quality Metrics

The metrics once created need to be theoretically and empirically validated to prove their utility in prediction of understandability of conceptual models. Validity can be said to be the degree to which a test supports what it claims. The creation process [16] of a metric undergoes two phases:

- **Theoretical Validation:** This phase checks whether the metric can numerically measure the defined attribute based on formal properties of data model. It gives a mathematical proof of relevance of metric towards measurement of some defined attribute. Of the various existing techniques for theoretical validation, like DISTANCE framework [17], Briand et al. framework [18] and Zuse framework [19], a measurement-based approach (DISTANCE framework) is selected for theoretical validation in this study due to its simple,

understandable nature based on the concepts of distance/dissimilarity between several entities of data warehouse conceptual models.

- **Empirical Validation:** This phase succeeds theoretical validation phase, where the metrics proved to be theoretically correct are applied to empirical validation techniques. An empirical validation is used for hypothesis testing. A hypothesis [11] is assumption that there exists a causal relationship among constructs of theoretical interest. The constructs are independent and dependent variables. The variables whose values remain constant during the course of experiment are independent variables and the variables whose values change during the course of experiment are dependent variables. The degree to which independent variables affect dependent variables is one of the major concerns of empirical validation. A questionnaire is prepared, data collected in response to the questionnaire, followed by analysis of collected data. This involves experiments, qualitative studies, surveys. Experiment is a form of empirical study carried out under controlled conditions to test the hypothesis. The output of empirical validation decides the acceptance, redefinition or rejection of proposed metric.

The metrics that pass the tests of validation (theoretical and empirical validation) successfully are valid metrics. Valid metrics are significant enough to predict the quality of data warehouse conceptual models [11, 16]. During the study of quality metrics, it was observed and found that each of the valid quality metric has its own contribution towards quality evaluation of conceptual data warehouse models towards building of an efficient information delivery system. Thus, another critical issue that emerged was evaluation of contribution of each valid metric of conceptual data warehouse model. This evaluation of conceptual model metrics can be used in design of an efficient information delivery system. The mechanism for calculating the importance of each valid quality metric and thereby ordering valid quality metrics was further focused on in the research study. The ordering of quality metrics for quality evaluation of conceptual models is introduced in next section.

1.6QUALITY METRICS: ORDERING APPROACH

The previous section discussed the role played by quality metrics towards quality evaluation of conceptual data warehouse models and the issue of ordering the quality metrics to discover the relative importance of each quality metric. The ordering [20]

of a quality metrics towards quality evaluation of conceptual models can improve the design quality of conceptual data warehouse models as the metrics at higher position in the order can be given higher consideration than the metrics at lower position.

The ordering of quality metrics can be measured along several criteria like understandability, efficiency and effectiveness. This leads to multi-criteria ordering of quality metrics. The criteria are defined qualitatively with non-numeric values. The significance of quality metrics along the criteria varies according to user requirements, situations and expert opinion, which is also non-numeric and not crisp. Further study in respect of techniques capable of dealing with non-crisp, ambiguous, non-numeric data was made. The best possible technique capable to handle imprecise and qualitative (non-numeric) data based on actual human (expert) decision making was identified to be fuzzy logic [21]. The fuzzy based approach could be one of the many existing techniques that can find way in ordering of quality metrics against multiple criteria towards building of an efficient information delivery system. The fuzzy based approach considers uncertainties, ambiguities, biases involved in human thought process and take into account all possible interdependencies of attributes involved.

Making use of ordering of quality metrics towards quality evaluation of conceptual models for building of an efficient information delivery system can be another aspect which needs further consideration. Proceeding in this direction, with the aim to make use of ordering of metrics towards building of an efficient information delivery system, it was discovered that a fuzzy rule base, based on the ordering of metrics, can be constructed for quality evaluation of conceptual data warehouse models. A broad introduction to fuzzy rule base system is presented in next section.

1.7 A FUZZY RULE BASE SYSTEM FOR PREDICTING THE QUALITY OF CONCEPTUAL MODEL

The previous section discussed the need of fuzzy based approach for ordering quality metrics for conceptual models. Once a hierarchical ordering of metrics is achieved, it could be used to generate a fuzzy rule based system [22] for predicting the understanding time of conceptual models. The input to the fuzzy rule base system can be the values of metrics of the conceptual models and the output can be the crisp

understanding time of models. The system could find the best conceptual model with least understanding time from its multiple variants. Also the system could reduce the effort involved in calculating the understanding time of conceptual models which involved preparation of questionnaire, surveying and analyzing survey to get the results in terms of understanding times. The issue of main concern is the validity of understanding times generated by the fuzzy rule base system. Research work has been carried out in this research study towards building of a rule base that could predict understanding times near to real world understanding times.

Another major aspect, identified during the research study was efficient information extraction from data repositories. Queries are thrown on data storehouses to get correct response in real time. The response time of queries to give correct results can also be considered a major aspect towards building of an efficient information delivery system. A good quality information delivery system should give quick and correct responses to queries thrown on it. Introduction towards improving query response time for information extraction from an information delivery system is presented in next section.

1.8 EFFICIENT INFORMATION EXTRACTION

Till now research study has focused on conceptual design phase of the data warehouse development process. An efficient data warehouse gives fruitful information and help managers to make intelligent decisions in response to the complex queries thrown on it. The response time to queries is a very crucial factor in governing the quality of data warehouse systems. Reducing the response time of queries by selection of only few and not all materialized views to give better trade off in terms of space/benefits is another issue of concern towards development of efficient information delivery system. Researchers have proposed several query optimizing techniques to improve query response time. One of the base query optimization approaches was proposed by Harinarayan et al. [23]. The approach works on lattice framework of data that is capable enough to show inter dependencies of data. The approach uses greedy approach for selection (views with maximum current benefits in terms of space) of materialized views. Further study revealed that greedy approach proposed by Harinarayan et al., is unable to deal with the situation as to which view to materialize when two or more views have same current benefits. Research work can also be

carried on in the direction to improve greedy selection approach for a better trade off in terms of space/benefits in response to queries, when two or more views have same benefits. The view selection can be further enhanced by including space constraints to the results of greedy and refined greedy approach.

Based on issues discussed in all of the above sections towards development of an efficient information delivery system, a detailed review of literature was conducted in each of the subdomains with the aim of throwing light on the research already conducted and to come up with some new research paradigms. The successive chapter presents a detailed literature review.

1.9 PROBLEM DEFINITION

The various issues related to the development of an efficient information delivery system have been introduced in the sections above. Basically there are three major components/phases of an efficient information delivery system namely data, efficient data warehouses and efficient information extraction systems. There is a wide scope of improvement and research in each of the phases of efficient information delivery system as seen from the introduction of various issues related to development of an efficient information delivery system. The overall purpose of research can be summarized as building of an efficient information delivery system. An efficient information delivery system is the need of time as it helps to gain competitive advantage to business organizations in ever growing competitive environment. An efficient information delivery system helps the managers to have a deep and correct knowledge of variable trends about customer needs and preferences, emerging technologies, sales and marketing techniques, quality of products and services over a period of time. This knowledge helps the business analysts and managers to take correct strategic decisions by analysis of past trends towards prediction of future market strategies. Based on the introduction to related issues towards development of an efficient information delivery system and on the basis of literature survey conducted in Chapter II of this thesis, we have set objectives for a comprehensive approach towards building of an efficient information delivery system. The set objectives are following:

- The primary objective towards development of an efficient information delivery system is to improve the quality of data base from which information

is to be extracted. This data base is data warehouse. Data warehouse quality is dependent on the quality of conceptual design model, logical design model and physical design model. The primary objective is to improve the quality of conceptual design model owing to its significance that has been discussed in various sections of literature review in Chapter II of thesis.

- The second objective towards development of an efficient information delivery system is efficient information extraction from data warehouse. Information extraction is said to be efficient if correct information can be provided by the system in minimum possible time.

To achieve the set objectives, we have set certain aims towards building of an efficient information delivery system, defined as follows:

- To study the development of data warehouse and come out with a detailed classification framework.
- To propose new quality metric towards quality evaluation of conceptual data warehouse models.
- To analyze the validity and significance of proposed metric along with existing metrics in quality evaluation of conceptual models.
- To evolve some methodology for ranking of quality metrics.
- To develop a rule base for predicting the understandability of conceptual models.
- To develop new technique/modify existing technique for efficient information extraction from a data warehouse.

1.10 ORGANIZATION OF THESIS

The organization of thesis is presented as follows:

- The thesis begins with an introduction to efficient information delivery systems and the various research aspects related to building of efficient information delivery systems as presented in Chapter I. This chapter also introduces a broad view of problem definition as identified by research scholar based on the study of various related issues concerned with the building of an efficient information delivery system.
- Chapter II presents a detailed review of literature conducted by research scholar in the domain of efficient information delivery systems. Based on this

literature review, the problem definition is again discussed in the light of existing research work in the domain of efficient information delivery system.

- Chapter III focuses on the research work carried by research scholar towards quality evaluation of efficient information delivery systems. In this chapter the theoretical and empirical validation of new proposed metric along with already existing metrics is carried out and the results obtained are analyzed.
- In Chapter IV Chapter the ranking of metrics using fuzzy based approach is presented and a automatic fuzzy inference system for predicting the understanding time of conceptual models is proposed. The results obtained are discussed and analyzed in detail.
- Chapter V concentrates on the research work carried to extract information efficiently from data storehouse and thus leading to the building of an efficient information delivery system. Chapter VI concludes the thesis with the directions of further research that can be conducted by researchers towards building of efficient information delivery systems.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

The literature survey related to the development of an efficient information delivery system has been conducted and is presented at depth in the chapter. An efficient information delivery system makes use of efficient data warehouse for information delivery as has been discussed in Chapter I. An efficient data warehouse system is capable to provide strategic information to the managers which include knowledge of variable trends about customer needs and preferences, emerging technologies, sales and marketing techniques, quality of products and services over a period of time. Efficient data warehouses are need of the time as they help to gain competitive advantage to business organizations in ever growing competitive environment.

The process of data warehouse development follows an incremental approach. Towards development of an efficient data warehouse system four incremental phases are identified namely requirement gathering, design, quality evaluation, data extraction and then further sub classification of each of the identified phases can be performed [1,2,3]. Detailed study of existing literature for each of the identified phases was conducted which gave directions for further research work towards accomplishment of the aim and objectives for building of an efficient information delivery system. In the successive sections a detailed study of literature for each of the identified phases towards building of an efficient data warehouse system is discussed and presented.

2.2 DATA WAREHOUSE DEVELOPMENT

The first major step towards building of an efficient data warehouse system is its design and development. This involves identification of various phases of data warehouse development, techniques and methods used in each phase and discovering possible sources used for design of each phase. To provide a comprehensive bibliography of academic literature on data warehouse development techniques the following online journals and conference databases were searched:

- Springer Publication
- Sage Publication
- Science Direct Publication
- ACM Publication
- IEEE Publication
- Wiley Publication
- IGI Global Publication
- Inderscience Publication
- Emerald Publication

Based on the study of literature, we framed a classification framework for efficient data warehouse design and development shown in Figure 2.1.

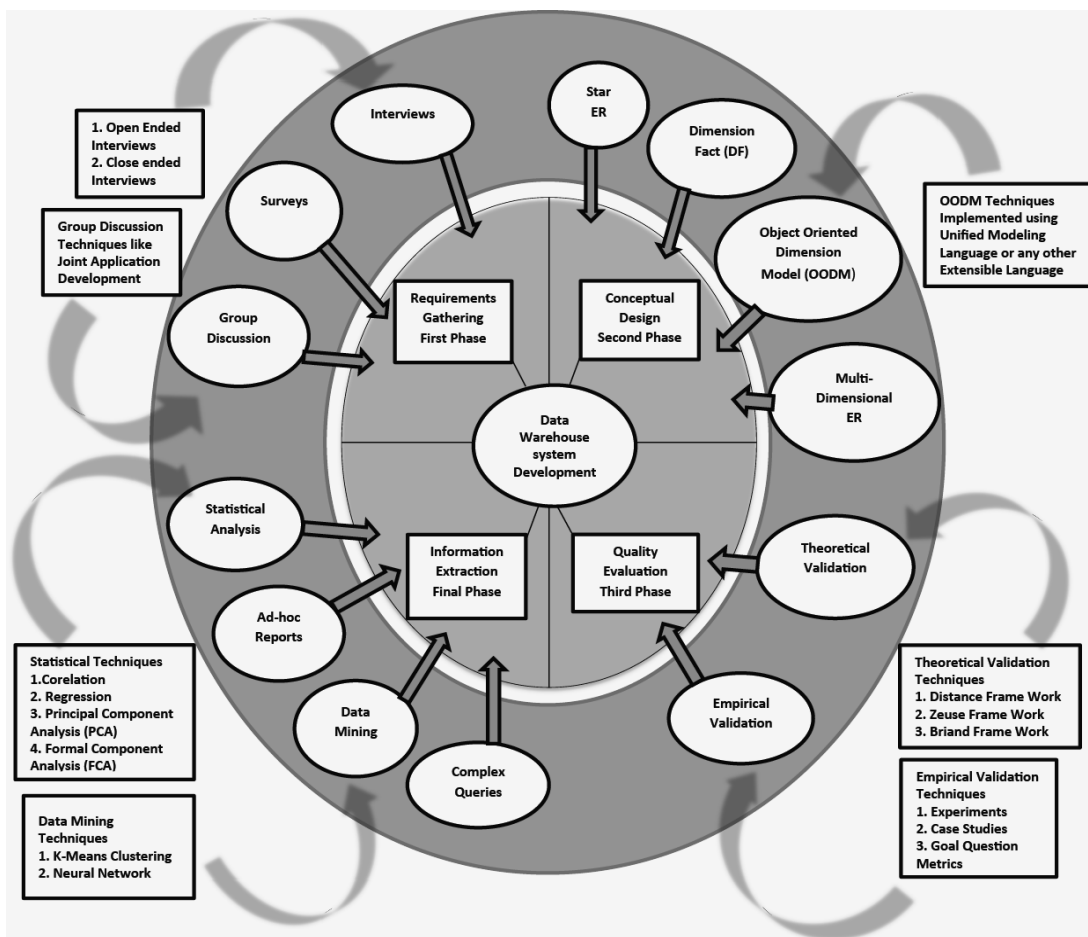


Figure 2.1 Classification framework for efficient data warehouse design and development

The classification for efficient data warehouse design and development is as follows:

- Requirement gathering phase
- Design phase (conceptual, logical, physical design models)
 - Data warehouse Security/ Design of secure Data warehouse
- Quality evaluation phase
- Data warehouse Testing
 - Information Extraction phase

As discussed by Annoni et al. [24] the requirement gathering forms the initial phase of data warehouse development. The users primarily business executives, managers, data base administrators are interviewed and joint application sessions conducted for collection of initial requirements to build a data warehouse. Successively, conceptual designing techniques such as StarER, dimension fact, multidimensional ER, OODM [25] are applied to data collected in requirement gathering phase[26] to build a conceptual model. The conceptual model built can be further extended to logical design model and the physical design model[27, 28, 29]. The significance of conceptual design model towards development of an efficient information delivery system has been discussed in the introductory chapter.

The quality evaluation of conceptual models built in design phase is performed to get the best possible design configuration. The quality of conceptual models can be evaluated by quality metrics that are based on size and structural complexities. Serrano et al. [11] discussed two methods for validation of quality metrics namely theoretical and empirical validation[30]. The best validated model, as obtained from previous phase, is subjected to various techniques for information extraction such as data mining [31], complex querying and statistical analysis techniques as stated by Pighin and Ieronettiet [32].

A tabular classification showing in depth the various sources, approaches, background information and methods in each phase of data warehouse development identified from the cited references is shown by Table 2.1.

Table 2.1 Distribution of articles according to proposed Classification Model

| Phases | Source | Approaches | Background Information | Methods | References | |
|-----------------------|---|---|---|---|--|---|
| Requirement Gathering | Users a)Senior Executives b)Department Managers c)Business Analysts d)Operational DBA's | Nterviews | History and current structure of business unit, No. of employees, Their roles and responsibilities, Location of users, Primary and secondary purpose of business unit, Relationship of business unit to other units, Contribution of business unit to revenue and costs | Select and train project team members, assign specific roles to each team member, prepare list of users to be interviewed, complete pre-interview research, prepare interview questions | [24] Annoni et al.(2006) [33] Haigh(2010) [34] Villarroel et al.(2006) [35] Rodriguez et al.(2006) [36] Solar et al.(2008) [37] Verbo et al.(2007) | |
| | | | Group Sessions (Joint Application Development) | Same as above and current information sources, subject areas, critical business performance metrics | JAD consists of five phases a)Project definition b)Research c)Preparation d)JAD Sessions e)Final documents | [38] Duggan and Thachenkary(2004) [16] Serrano et al.(2007) [39] Munoz et al.(2010) |
| | | | StarER | Facts, entities, relationship between facts and entities, attributes | Combination of star structure with constructs of ER model to build a starER model | [40]Norberto et al.(2009) |
| | | | | Dimension Facts | Facts, attributes, dimensions, hierarchies in dimensions | A tree structured approach to build dimension fact model |
| Conceptual Design | Information Packages (Requirement Definition Document) | Multi-dimensional ER (ME/R) | Same as starER, a special entity set, a special n-ary fact relationship. Special binary rolls up relationship | The approach follows removal of relationship from star ER and addition of specific constructs | [45] Zhang et al.(2011) [46] Hendawi and Sappagh(2012) [47] Cuzzocrea(2006) [48] Simitsis and vassiliadis(2008) | |
| | | | Object oriented Dimensional Modeling (ODM) | Fact class, dimensional class, properties such as inheritance, generalization, specialization, polymorphism | Object oriented approach using unified modeling language (UML), Extensible markup language(XML) | [34] Villarroel et al.(2006) [49] Medina et al.(2007) [50] Genero et al.(2008) [51] Villaroel et al.(2005) |
| | | Theoretical Validation (helps to know when, how to apply metrics) | Identification of goals of metrics, formulation of hypothesis, definition of metrics, characteristics of system, designer's experience | a) Distance framework b) Zuse Framework (Both based on measurement theory) c) Briand framework based on axiomatic approach | [52] Schuff et al.(2011) [53] Ramamurthy et al.(2008) [54] Batini et al.(2009) [55] Moody(2005) [16] Serrano et al.(2007) [56] Khurram and Mustafa (2010) | |
| | | | Empirical validation (help us to prove the practical utility of proposed metric) | Same as background information for theoretical validation | a)Experiments b) Case studies c)Surveys | [57] Kpodjedo e al.(2011) [58] Verbo et al.(2009) [15] Caballero et al.(2009) |
| Quality Evaluation | Conceptual, Logical, Physical Dimensional Model | | | | | |

| Phases | Sources | Approaches | Background Information | Methods | References |
|------------------------|--|---|--|--|--------------------------------|
| Information Extraction | Data Warehouse Multi-dimensional Model | Ad-hoc reports, complex queries, Data mining, statistical analysis, | Knowledge of domain, characteristics of system, familiarization of basic operations on data warehouse system | Operations like rollup, drill down, drill across, slicing, diasing, correlation, regression, PCA's Formal concept analysis, K Means algorithm, Neural Networks | [59] Mojaveri et al(2010) |
| | | | | | [60] Rahman and Harding (2012) |
| | | | | | [61] Bhamra et al.(2011) |
| | | | | | [62] Bobby and Lee(2009) |
| | | | | | [63] Nedjar et al.(2009) |

Another view presents the number of research publications related to each phase of data warehouse development for the period 2005-12 based on the study of existing literature, shown by Figure 2.2.

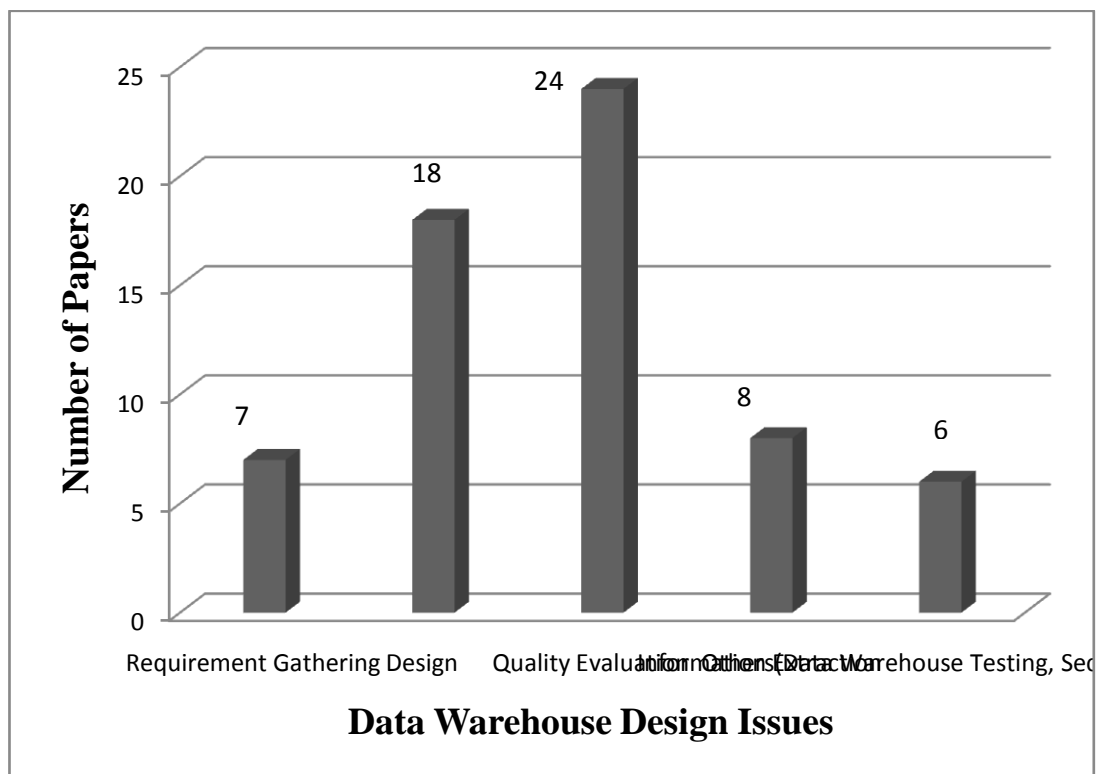


Figure 2.2 Bar plot.
Methodology adapted from Source: Ngai and Wat (2002) [64]

As can be seen from bar graph of Figure 2.2, the trend of research work is more concentrated on quality evaluation phase followed by design, information extraction and requirement gathering.

Based on the study of the literature, the various identified techniques used in each phase of efficient data warehouse development along with the number of related published papers was summarized and presented by Table 2.2. The techniques used in each phase of data warehouse development namely requirements gathering, design, quality evaluation and information extraction is shown in detail along with number of papers published for each of the techniques giving suitable references.

Table 2.2 Techniques for data warehouse development
Methodology adapted from Source: Ngai and Gunasekaran (2007) [65]

| Development Phases | Techniques | Number of Papers | References |
|-------------------------------|--|------------------|---|
| Requirement Gathering | 1) Interview Techniques like open and closed ended interviews | 3 | [24] Annoni et al.(2006), [38] Duggan and Thachenkary(2004), [25] Hofman(2011), [33] Haigh(2010), [66] Mellado et al.(2010), [35] Rodriguez et al.(2006), [36] Solar et al.(2008), [37] Verbo et al.(2007) |
| | 2) Group discussions | 3 | |
| | 3) Joint Application Development | 2 | |
| Design | 1) Star ER Model | 2 | [67] Bara et al.(2009), [47] Cuzzocrea(2006), [68] Genero et al.(2007), [69] Haider and Kumar(2011), [46] Hendawi and Sappagh(2012), [42] Rifaie et al.(2009), [70] Rjagan et al.(2005), [28, 29] Blanco et al.(2009a,2009b), [49] Medina et al.(2007), [44, 39] Munoz et al.(2009,2010), [71] Riberio et al.(2011), [72] Simitsis et al.(2008), [41] Trai et al.(2012), [34] Villarroel et al.(2006) |
| | 2) Dimension Fact Model | 1 | |
| | 3) Multidimensional ER Model | 2 | |
| | 4) Object Oriented Dimensional Modeling | 12 | |
| | 5) Extraction, Transformation, Loading Techniques | 3 | |
| Quality Evaluation | 1) Theoretical Validation Techniques like Distance Framework, Zuse Framework Approach | 6 | [54] Batini et al.(2009), [15] Caballero et al.(2009), [30] Caro et al.(2007), [50] Genero et al.(2008), [73, 74] Gosain et al.(2011,2012), [75] Kefi and Koppel(2011), [57] Kpodjedo et al.(2011), [27] Blanco et al.(2008), [55] Moody(2005), [16, 11] Serrano et al.(2007,2008), [76] Smith(2011), [58] Verbo et al.(2009), [26] Even and Shankararayanan(2007) |
| | 2) Empirical Validation Techniques like Surveys, Experiments, Questionnaires | 15 | |
| | 3) Statistical Techniques like Correlation, Regression, Principal Component Analysis, Formal Concept Analysis, Fuzzy Logic | 14 | |
| Information Extraction | 1) Data Mining Techniques like K-Means clustering, Neural Network based approaches | 10 | [77] Aggarwal et al.(2012), [62] Bobby and Lee(2009), [61] Bhamra et al.(2011), [59] Mojaveri et al.(2010), [63] Nejdard et al.(2009), [31] Pabreja and Datta(2012), [60] Rahman and Harding(2012), [3] Cannoly and Begg(2012), [2] Han and Kamber(2012), [1] Inmon(2010), [4] Pooniah(2010), [78] Thuraisingham et al.(2007), [75] Kefi and Koppel(2011) |
| | 2) Querying | 6 | |
| | 3) Statistical Techniques like Correlation, Regression, Principal Component Analysis, Formal Concept Analysis, Fuzzy Logic | 14 | |

Another dimension to present the observations of literature review towards development of an efficient information delivery system is shown by line graph of Figure 2.3, which shows a trend of the number of papers published in the period from 2005 to 2012 in the data warehouse domain.

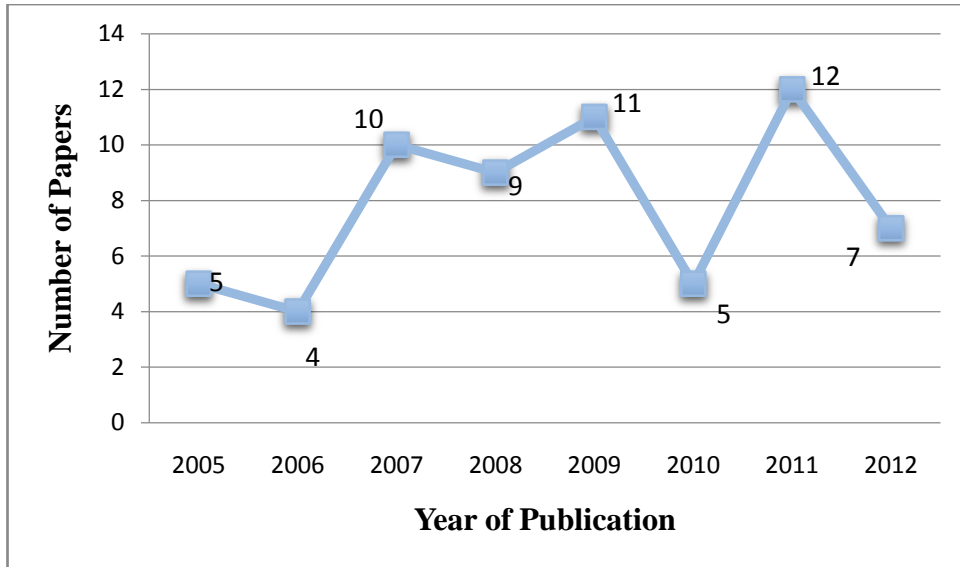


Figure 2.3 Line graph
Methodology adapted from Source: Ngai and Gunasekaran (2007) [65]

The line graph of Figure 2.3 shows the number of publications in data warehouse domain is increasing with time. This shows the continuity of research work in the data warehouse domain.

From the study of articles, it can be stated that data warehouse is the need of modern era which requires ever evolving efforts from the researchers for future research. The analysis of literature also shows that the dominant area of research in data warehouse domain continues to be quality evaluation. Further study of literature related to quality evaluation of conceptual design models was conducted and is presented in next section.

2.3 QUALITY EVALUATION OF CONCEPTUAL DATA WAREHOUSE MODELS

Data warehouse system development follows conceptual phase followed by logical and physical design phases. The significance of conceptual design phase in data warehouse development towards building of an efficient information delivery system has already been discussed and made a part of study for review of literature. Further study of literature revealed that quality evaluation of conceptual models can be measured using certain quality parameters called quality metrics. Several quality

metrics have been proposed by researchers to evaluate the quality of conceptual models.

Basili et al. [79] discussed the importance and need of experimentation in empirical validation. The main focus was on defining goal and hypothesis, experimentation and validity threats to experiments.

Calero et al. [14] proposed various metrics for different configurations of data warehouse models.

- Table metrics: NA, NFK
- Star metrics: NDT, NT, NADT, NAFT, NA, NFK, RSA, RFK
- Model metrics: NFT, NDT, NSDT, NT, NAFT, NADT, NASDT, NA, NFK, RSDT, RT, RFK, RSDTA

Also these metrics were theoretically validated to prove their relevance towards quality evaluation of data warehouse conceptual models.

Serrano et al. [80] has thrown light on various issues involved in experimental validation of multi- dimensional data model metrics.

Moody [55] discussed various theoretical and practical issues in evaluating the quality of conceptual models with special emphasis on experimental techniques.

Serrano et al. [16] has proposed a set of quality metrics for evaluation of conceptual models. The proposed metrics have been theoretically and empirically validated to prove their utility. The proposed metrics for a conceptual model S are as follows:

- NDC(S) Number of dimension classes (equal to the number of aggregation relationships)
- NBC(S) Number of base classes
- NFC(S) Number of fact classes
- NC(S) Total number of classes, $NC(S) = NDC(S) + NBC(S) + 1$
- RBC(S) Ratio of base classes. Number of base classes per dimension class
- NAFC(S) Number of FA (fact attributes) attributes of the fact class
- NADC(S) Number of dimension attributes of the dimension classes

- NABC(S) Number of dimension attributes of the base classes
- NA(S) Total number of fact and dimension attributes, $NA(S) = NAFC(S) + NADC(S) + NABC(S)$
- NH(S) Number of hierarchy relationships
- DHP(S) Maximum depth of the hierarchy relationships
- RSA(S) Ratio of attributes. Number of fact attributes divided by the number of dimension attributes

Empirical validation has been applied to prove that several of the metrics seem to be practical indicators of conceptual model understandability and play a major role towards quality evaluation of conceptual models.

Genero et al. [68] proposed 3 size metrics and 8 structural metrics for conceptual models as follows:

Size metrics

- Number of Classes (NC) The total number of classes in a class diagram
- Number of Attributes (NA) The number of attributes defined across all classes in a class diagram (not including inherited attributes or attributes defined within methods).
- Number of Methods (NM) The total number of methods defined across all classes in a class diagram, not including inherited methods.

Structural metrics

- Number of Associations (NAssoc) The total number of association relationships in a class diagram
- Number of Aggregations (NAgg) The total number of aggregation relationships (each “wholepart” pair in an aggregation relationship).
- Number of Dependencies (NDep) The total number of dependency relationships.
- Number of Generalizations (NGen) The total number of generalization relationships (each “parent-child” pair in a generalization relationship).
- Number of Generalization Hierarchies (NGenH) The total number of generalization hierarchies, i.e. it counts the total number of structures with generalization relationships.

- Number on Generalization Hierarchies (N_{AggH}) The total number of aggregation hierarchies, i.e. it counts the total numbers of “whole-part” structures within a class diagram.
- Maximum DIT (MaxDIT). The maximum DIT value obtained for each class of the class diagram. The DIT value for a class within a generalization hierarchy is the longest path from the class to the root of the hierarchy
- Maximum H_{Agg}(MaxH_{Agg}) The maximum H_{Agg} value obtained for each class of the class diagram. The H_{Agg} value for a class within an aggregation hierarchy is the longest path from the class to the leaves.

It was explored through experimentation that some of these metrics were good predictors of maintainability of class diagrams i.e. understandability and modifiability. A new concept of PCA (Principal component analysis) was also incorporated to identify principal component metrics capable of explaining the model without loss of any significant information.

Serrano et al. [11] proposed a set of structural metrics for quality evaluation of logical models and carried out an empirical study with the aim of investigating these metrics towards measurement of understandability of logical models. The proposed metrics are as follows:

- NFT(Sc). Number of fact tables in the model
- NDT(Sc). Number of dimension tables in the model
- NFK(Sc). Number of foreign keys in all the fact tables of the model
- NMFT(Sc). Number of facts in the fact tables of the model

These metrics were found to be good indicators of data warehouse quality.

Shull et al. [81] defined the role of replications in empirical study. Replications were categorized in two types:

- Exact replication
- Conceptual replication

Exact replication was further classified as dependent and independent replication. Goals, benefits, limitations of each have been discussed. Also role of documentation was further emphasized.

Lucia et al. [82] conducted three sets of controlled experiments aimed at analyzing whether UML class diagrams are more comprehensible than ER diagrams during data models maintenance. The results indicated that UML class diagram subjects achieved better comprehension levels.

Haigh [33] conducted an empirical study involving survey of more than 300 current and just graduated students asking them to rate the importance of 13 quality attributes related to software. The results showed differences in some but agreement in many areas.

Gosain et al. [73] conducted a replica study to explore a correlation between understandability and metrics proposed by Serrano et al. [16]. The results show that NFT, NDT, NFK have significant role towards predicting understandability of logical models, while NMFT was not found to be correlated to understandability. Also the combined effect of different combinations of metrics using univariate and multivariate regression was carried out.

Hofman [25] conducted an empirical validation to analyze the ‘history effect’ in software quality evaluation process. A simplified method was proposed to manipulate observed quality level for a product, thereby making it possible to conduct research. The results showed significant negative influence of negative experience of users on final opinion about software quality regardless of its actual level.

Kpodjedo et al. [57] performed an investigation to find the usefulness of elementary design evolution metrics to identify defective classes. It was shown that design evolution metrics make significantly better predictions of defect density than other metrics and thus help in reducing testing effort by focusing test activity on reduced volume of code.

The study of literature provided motivation to investigate and propose new metric that can be a good predictor of quality of conceptual models and prove its utility by theoretical validation and empirical validation by making use of several statistical techniques, towards development of an efficient information delivery system.

Next the study of literature was conducted related to existing approaches for classification and ordering of quality metrics which is discussed in next section.

2.4 CLASSIFICATION AND ORDERING OF QUALITY METRICS

The previous section presented a literature review related to various quality metrics proposed by researchers for quality evaluation of each type of conceptual design techniques. The quality metrics were found to be based on size and structural complexity of conceptual data warehouse models. The study of literature also revealed that there exist multiple criteria like understandability, efficiency, effectiveness along which quality of conceptual models can be measured using quality metrics. Researchers have conducted controlled empirical experiments to prove the effect of quality metrics on multiple criteria like understandability, efficiency, effectiveness of conceptual data warehouse models. The effect of each metric and hence its relative importance towards predicting the quality of conceptual models can be one of the major considerations during design of conceptual data warehouse models. The need for a methodology for precise ordering of quality metrics towards building of good quality conceptual models was felt. The literature related to classification and ordering of metrics is as follows:

Johnson and Yu [83] proposed a software quality model, based on Bayesian Belief Network (BBN), to predict software reliability through analysis of software metrics.

Dyba [84] identified and ranked key factors involved in software process based on expert opinion.

Briand et al. [85] proposed an approach based on expert opinion to estimate cost effectiveness of software model.

Zhang and Pham [86] conducted an empirical research on data collected from managers, system engineers, programmers and testers of top 13 companies. Based on collected data 32 factors were identified that were involved in every phase of software development. Two techniques namely relative weight method and analysis of variance (ANOVA) were used to analyze and rank the identified factors affecting software reliability.

Ming and Carol [87] conducted a study on thirty identified potential factors affecting software design and reliability. The ranking score for each factor was elicited based on expert opinion to identify and rank the factors in terms of their potential significance.

Garg et al. [20] proposed an approach for ranking of software metrics based on fuzzy logic along certain identified criteria.

In all of the proposed techniques (except [20]) algebraic aggregation has been used to quantify scores of expert opinion with no consideration of uncertainties, ambiguities and biases in human thought process. Ordering of metrics along variable criteria (understandability, efficiency and effectiveness) can lead to multiple-criteria decision making problem. The criteria are defined qualitatively and the significance of quality metrics along the criteria vary according to user requirements, situations and expert opinion. Thus the need of a fuzzy based system was felt that can deal with imprecise and qualitative (non-numeric) data based on actual human (expert) decision making. Further the study of basics of fuzzy logic was incorporated using research references [88, 89, 90, 91, 92, 93, 94, 95]. Also for multi criteria analysis we searched and referenced various research references [96, 97, 98]. The study of literature towards classification and ordering motivated to rank quality metrics for conceptual models along various identified criteria using fuzzy approach.

During the literature review, a study of the research work by Ali and Gosain [22] was made in which fuzzy logic has been used to model non-linear relationship between metrics and understandability of conceptual models. Based on the idea adapted from source Ali and Gosain [22], the need for research work to prepare a fuzzy rule base for predicting the understandability of conceptual models based on the values of their quality metrics was also identified.

Apart from various quality issues related to quality evaluation of conceptual models, the need of an efficient data warehouse system that can efficiently handle and provide strategic information in response to the queries thrown on it was also felt and worked upon. The study of literature related to various techniques for improving the response time of complex queries thrown and provide strategic information efficiently was

made. In next section the related literature in the domain of information extraction efficiently from a data warehouse is discussed and presented.

2.5 EFFICIENT INFORMATION EXTRACTION

This section presents the review of literature conducted towards information extraction in an efficient manner from big data warehouses. Efficient information extraction has also been identified as one of the major issues towards development of an efficient data warehouse system.

A study on decision Support Systems [23, 99] was conducted that showed that decision support systems are supported by huge data warehouses at back end. Complex queries run on data warehouses and results achieved. Query response time was identified as one of the major factors affecting the quality of data warehouses. One of the major factors affecting query optimization is optimal selection of materialized views [100]. Many solutions to the view selection problem have been proposed and analyzed [23, 101, 102].

The basic framework for view selection using greedy approach was proposed by Harinarayan et al [23]. He discusses lattice framework, cost model, benefit metric and greedy approach for materialized view selection. A comparison was also made for greedy view selection and optimal view selection by Harinarayan et al [23].

A pick aggregates algorithm for view selection based on greedy approach was proposed by Shukla et al [103]. The algorithm selects aggregate of views based on pre computed benefits following greedy approach. Many researchers have used A* algorithm [104] based approach to materialize view indexes [105].

Dhote and Ali [101] presented analysis of various methodologies for materialized view selection in data warehouse systems. The solutions to materialized view selection [106] were categorized along various dimensions like: frameworks and resource constraints.

Mami and Bellahsene [106] categorized various algorithms employed to perform view selection as: deterministic algorithms, randomized algorithms, hybrid algorithms or

constraint programming. These algorithms differ in their approach to solve materialized view selection problem and so differs their time and space complexities.

Research work was carried on materialized view selection starting from the very base approach proposed by Harinarayan et al [23] by including some more parameters like space constraint, cost benefits towards development of an efficient information delivery system.

2.6 PROBLEM DEFINITION: REVISED

The problem definition has been introduced in Chapter I. The purpose of research, aims and objectives to be achieved during the course of research have been discussed in problem definition itself. Based on the literature review, the problem definition can again be discussed in the vicinity of research work performed by researchers in the related efficient delivery systems domain. The discussion is as follows:

- Researchers have conducted research work in the domain related to development of an efficient information delivery system [24-29]. Based on literature review from various journals [24-63] like springer, IEEE, Elsevier, wiley, sage, inderscience, IGI global, ACM, we have proposed a classification framework for data warehouse development. The various methodologies and techniques used in various phase of data warehouse development have been identified from various research papers and presented in tables (Table 2.1, Table 2.2) discussed in above sections. Next, we have identified that current hot domain of research in data warehouse development is quality evaluation phase (Figure 2.2, Figure 2.3). Further literature study was made in quality evaluation of data warehouse domain and it was found that quality of a data warehouse conceptual model depends on quality metrics [11, 14, 16, 25, 33]. Several quality metrics have been proposed to evaluate the quality of conceptual data warehouse models [55, 57, 68, 73, 79, 80, 81, 82]. Based on the study of quality metrics, we found that still more quality metrics can be proposed and validated that can have significant affect towards quality evaluation of conceptual data warehouse models. To achieve the objective of improving the quality of data warehouse, one step identified by us is proposal of new quality metrics.

- Based on the literature review, it was found that relevance and utility of proposed metric can be checked by applying first theoretical validation and then empirical validation techniques. The theoretical validation [11, 16] can be performed using any of the existing theoretical validation techniques like DISTANCE framework [17], Zuse framework [19], Briand et al. framework [18]. Theoretical validation is performed to check whether a proposed metric can measure some defined quality attribute numerically or not. Next empirical validation [55, 57, 80] of metric can be performed with a view to check its practical significance and utility in real world. For empirical validation, a controlled experiment is carried out. Initially data warehouse conceptual models from variable domains are prepared. Hypothesis are developed, independent and dependent variables are identified, questionnaire based on structural and size complexity of each conceptual model are prepared and participants are identified. Data is collected from participants who respond to questionnaire of each model and time to respond to questionnaire is recorded for each participant. After data collection, various statistical techniques like correlation, regression, principal component analysis, nearest neighbour analysis are applied to collected data to test the hypothesis [11, 16]. This is the second step towards achievement of set objectives i.e. improving the quality of data warehouse towards building of an efficient information delivery system.
- Based on the literature review, it was found that there exist several techniques for ranking of entities [20, 83-86]. None of the techniques have ranked quality metrics for data warehouse conceptual models along parameters namely understandability, efficiency and effectiveness [16]. If individual impact of quality metric on quality evaluation of data warehouse conceptual models can be found, then it could give way to design of good quality conceptual data warehouse models. The metrics having higher rank can be given more emphasis during design of conceptual data warehouse models, to improve the overall design quality. No researcher has used fuzzy based matrix methodology based on expert opinion to order the quality metrics. The concept of fuzziness [87-95] needs to be referenced as the values of criteria are not crisp values and no other but experts can give a reliable opinion

towards ranking of quality metrics. This is the third step towards achievement of set objectives i.e. improving the quality of data warehouse towards building of an efficient information delivery system.

- Also it was identified in the literature review that calculation of understanding time of conceptual data warehouse models is a very tedious task which requires design of conceptual models, identification of subjects, preparation of questionnaires based on structural properties of models, collection of data in the form of time and then further aggregating the data to get understanding time of model [11, 16]. To minimize the efforts involved in prediction of understanding time, the need for a system that could predict the understanding time of conceptual models was felt. The research work can be carried on towards building of a fuzzy rule base system based on ranking of quality metrics and expert opinion that takes as input values of quality metrics and gives understanding time as output [22]. The design of a fuzzy rule base system that could predict the understandability of conceptual data warehouse models is the fourth step towards achievement of set objectives i.e. improving the quality of data warehouse towards building of an efficient information delivery system.
- The existing literature on efficient information extraction systems from a data warehouse was studied. Several techniques for efficient information extraction from a data warehouse have been proposed by researchers [23, 99, 101, 102, 106]. During the study of literature, we applied our focus onto whether the existing approaches can have scope of further enrichment. We started with the study of a base approach for materialized view selection as proposed by Harinarayan et al [23]. The refinement of base approach for efficient materialized view selection towards accomplishment of second objective of efficient information extraction from data warehouse was identified as the fifth step towards building of an efficient information delivery system.

Based on the literature review, a research framework has been proposed for conducting further research work towards development of an efficient information delivery system. The research framework for current research work is presented by Figure 2.4 which shows that research work starts with the review of literature on

efficient information delivery systems followed by identification of two areas. One dealing with quality evaluation and other dealing with efficient information extraction towards building of efficient information delivery systems. The detailed research work carried out on the basis of literature review and formulated research framework is presented in successive chapters.

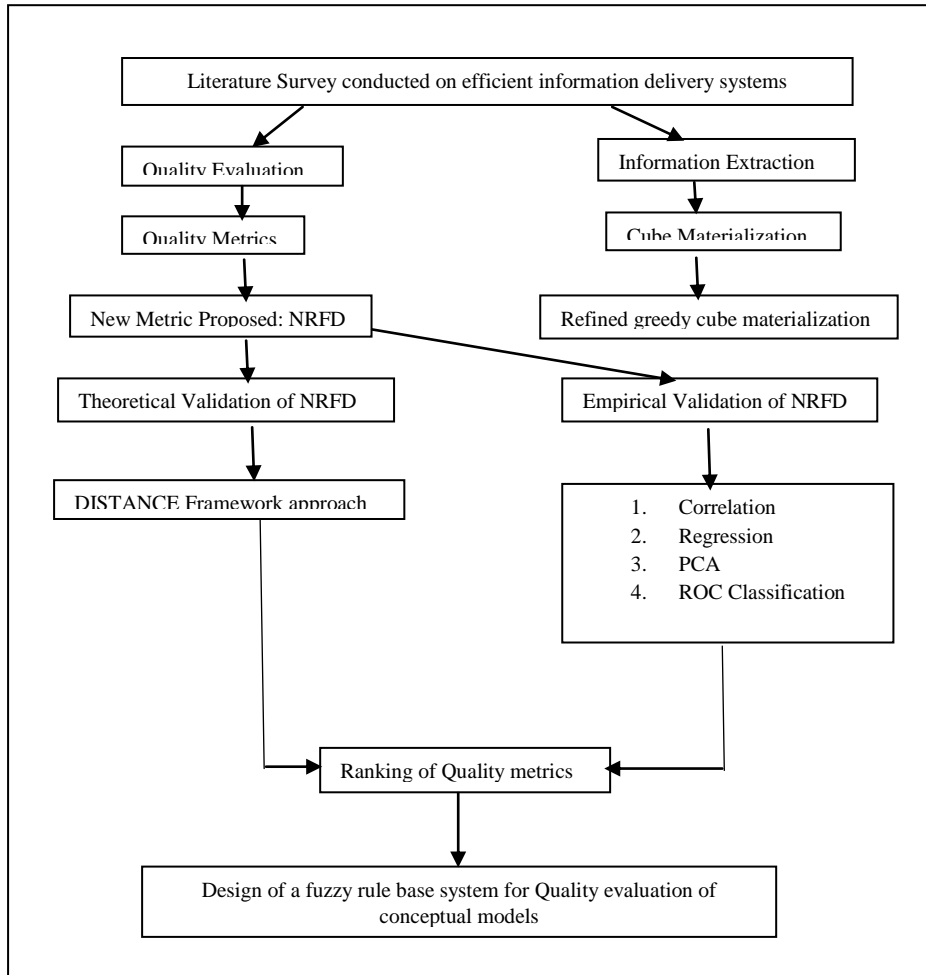


Figure 2.4 Research Framework

2.7 DATA COLLECTION AND ANALYSIS

The data needs to be collected for empirical validation of proposed quality metric. The goal of empirical validation is to analyse the quality metrics and evaluate them for predicting the understandability of conceptual data warehouse models. To investigate the role of quality metrics in data warehouse a controlled experiment was conducted. A total of 80 students, 22 conceptual models and 13 quality metrics were included in the experiment. Twenty-two conceptual models were used to perform the experiment. The models depict data from real world applications of different domains

like banking, airlines, universities, insurance, medical, railways. The model selection was based on structural complexity of conceptual models rather than domain categorization to predict the understandability because the values of metrics are dependent on structural attributes of models and are domain independent.

Eighty students pursuing B.Tech. in the institute, where one of the research scholars is employed, participated as volunteers in the experiment. The experiment was conducted in April, 2014. These students were in third year of their degree course in CSE and IT streams and all were in similar age group of 20-21 years. All the volunteers had adequate knowledge of data warehousing and UML concepts because they were studying the subject 'Data Warehousing and Data Mining' as a part of their course curriculum in third year. The participation of the subjects was taken up voluntarily and it was apart from their course curriculum. The experiment was conducted in two separate rooms with strength of 40 students in each room. A supervisor was appointed for monitoring the students in each room. Before the start of experiment the students were given a description of the tasks to be performed and a tutorial to brush up their related concepts. A sample model was taken, calculation of values of metrics from model was shown, how the questions were to be answered for the sample, where and in what format the answers were to be placed, where and in what format the starting time/ending time of the tasks were to be recorded. A wall clock was installed in each room for recording the start and end time of each task. The time taken by each participant for answering the tasks of each model was recorded and gathered. From the collected data average time for each model is calculated, which acts as dependent variable for further processing. Secondly, for ordering of quality metrics a fuzzy based matrix methodology has been used. The fuzzy based matrix methodology is based on expert opinion. Five experts were identified, out of which 3 were from academics and 2 were from industrial background. All of the five have rich experience of about 15-20 years in the data warehouse domain. The opinion of experts was recorded in a pre-defined format in the form of fuzzy linguistic variables for further processing.

2.8 TOOLS

For accomplishment of objectives set with an aim of building an efficient information delivery system, several software tools to be used are as follows:

- IBM SPSS Statistics has been used for performing various statistical operations for empirical validation of metrics.
- MATLAB has been used for ordering quality metrics and design of a fuzzy inference system for predicting quality of conceptual models.
- Java has been used for implementing refined greedy approach for efficient materialized view selection.

2.9 UNITSUSED FOR ANALYSIS

For accomplishment of objectives towards development of an efficient information delivery system, data is to be collected for further processing. Data collection is done separately for two purposes.

- First data collection is for empirical validation of proposed metric in which the time taken in seconds by each of the participant for each conceptual model is recorded in a designed questionnaire.
- Second expert opinion in the form of fuzzy linguistic variables is recorded for ordering of quality metrics and building of a fuzzy inference system to predict the quality of conceptual data warehouse models.

2.10 THESIS OUTLINE

The research work in the thesis is carried out with an aim towards building of an efficient information delivery system. The focus is on quality evaluation and information extraction aspects of information delivery system. To achieve the objectives and aims set towards building of an efficient information delivery system research work has been carried in the following areas:

- Proposal of a new quality metric
- Check its utility and relevance by carrying out theoretical and empirical validation
- Ordering the quality metrics to judge their contribution towards quality evaluation of information delivery systems
- Building inference system for predicting the quality of conceptual models towards development of an efficient information delivery system.
- Developing a refined greedy materialized selection approach for efficient information extraction form an information delivery system

CHAPTER III

THEORETICAL AND EMPIRICAL STUDY TOWARDS BUILDING OF EIDS

3.1 INTRODUCTION

From the literature review conducted in previous chapter, quality evaluation of data warehouse conceptual model was identified as one of the research domains in which maximum research work has been carried on by the researchers towards development of an efficient information delivery system. The importance of conceptual design model in efficient data warehouse development has also been discussed. The chapter presents the importance of quality metrics towards prediction of quality of conceptual data warehouse models. A new metric has been proposed and its theoretical as well as empirical validation has been carried out to evaluate the significance of proposed metric towards quality evaluation of conceptual data warehouse model. A step by step approach towards achievement of set objectives and aims forms the core of this chapter. The successive section discuss the research framework proposed towards development of an efficient conceptual data warehouse system.

3.2 CONCEPTUAL FRAMEWORK FOR EFFICIENT DATA WAREHOUSE SYSTEM

A framework has been proposed that follows a layered approach towards development of an efficient information delivery system. The proposed research framework is shown in Figure 3.1.

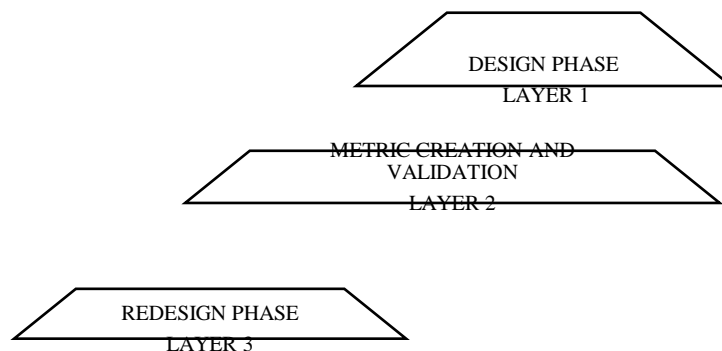


Figure 3.1 Layered Approach for development of an efficient data warehouse system

The framework starts with design of conceptual model in Layer 1 followed by creation and validation of quality metrics based on structural properties of conceptual model in Layer 2 and finally redesign of conceptual model taking in consideration quality metrics in Layer 3. A brief overview of each layer is presented as follows:

- Layer 1 presents the importance of information quality in the design of conceptual models of data warehouse. The output of layer 1 is a conceptual model built using object oriented dimensional modelling technique. The model does not take into consideration the quality factors based on size and structural complexity of conceptual model.
- Layer 2 describes the whole process starting from the proposal and creation of quality metrics till their application to real dynamic environment. The quality metrics measure the quality of data warehouse conceptual model. The output of layer 2 is a set of valid metrics defined and validated based on the structural properties of conceptual model that provide a quality measures for evaluating the quality of conceptual models towards development of efficient information delivery system.
- Layer 3 shows that a good quality conceptual model built by taking into consideration quality metrics, leading to the development of an efficient data warehouse system because conceptual phase lays the foundation of development of data warehouse. The conceptual model built in layer 1 is redesigned taking into consideration the valid quality metrics defined and validated in layer 2.

The detailed framework for Figure 3.1 is shown in Figure 3.2. The detailed analysis of each layer is enumerated as follows:

- **Layer 1** forms the initial phase in the data warehouse development. As seen from Figure 3.2, the information quality depends on data quality and data presentation quality. The data warehouse quality depends on data quality, data model quality and data base management system quality [11, 16]. Data model quality depends on conceptual model quality, logical model quality and physical model quality. The conceptual design phase is the initial design phase in data warehouse development, so its quality will affect the overall design quality of successive phases in data warehouse development. The conceptual

model can be designed using various techniques. A brief overview of each technique is presented as follows:

- Star ER conceptual model: It consists of a single fact set around which dimension sets are arranged using relationship sets (one to many relationship between fact and dimensions). The fact, dimensions and relationship sets consist of respective attributes as discussed by Mishra et al [10].
- Dimension Fact model: The dimension fact [10] modelling approach is suitable for representing hierarchies. It consist of a structure in the form of a tree whose elements are facts, dimensions, attributes and hierarchy relationships.
- Multidimensional ER model: This modelling approach is an extension of entity relationship model. It shows hierarchy in dimensions and operation ‘rolls up’ that can be applied to dimension hierarchy levels as specified in Mishra et al. [10].
- Object Oriented Dimensional model: The technique uses object oriented features such as specialization/generalization and the concept of objects, classes to show dimension/ facts. Languages such as XML, UML [68, 107] which support object oriented features are used for modeling approach. The object oriented dimensional modelling approach is widely used to design conceptual models due to its ability to model objects closer to real world entities. It makes use of inheritance, encapsulation and data hiding. The object oriented dimensional model quality depends on the size and structural complexities of model. The size and structural properties of models are measured in terms of size and structural metrics, as shown by detailed analysis of layer 2. The complexity of conceptual model is measured in terms of attributes namely understandability, efficiency and effectiveness.
- **Layer 2** defines the complete process of quality metrics creation and validation. Based on the size and structural properties of conceptual data warehouse model new quality metrics can be proposed. The proposed metrics can be theoretically validated [16] to test their correctness. Broadly theoretical validation follows two approaches:

- Framework based on axiomatic approach: In this approach a set of formal properties is defined for a domain. The metrics thus created are based on these properties and are used for further classification. Briand et al. framework is based on axiomatic approach.

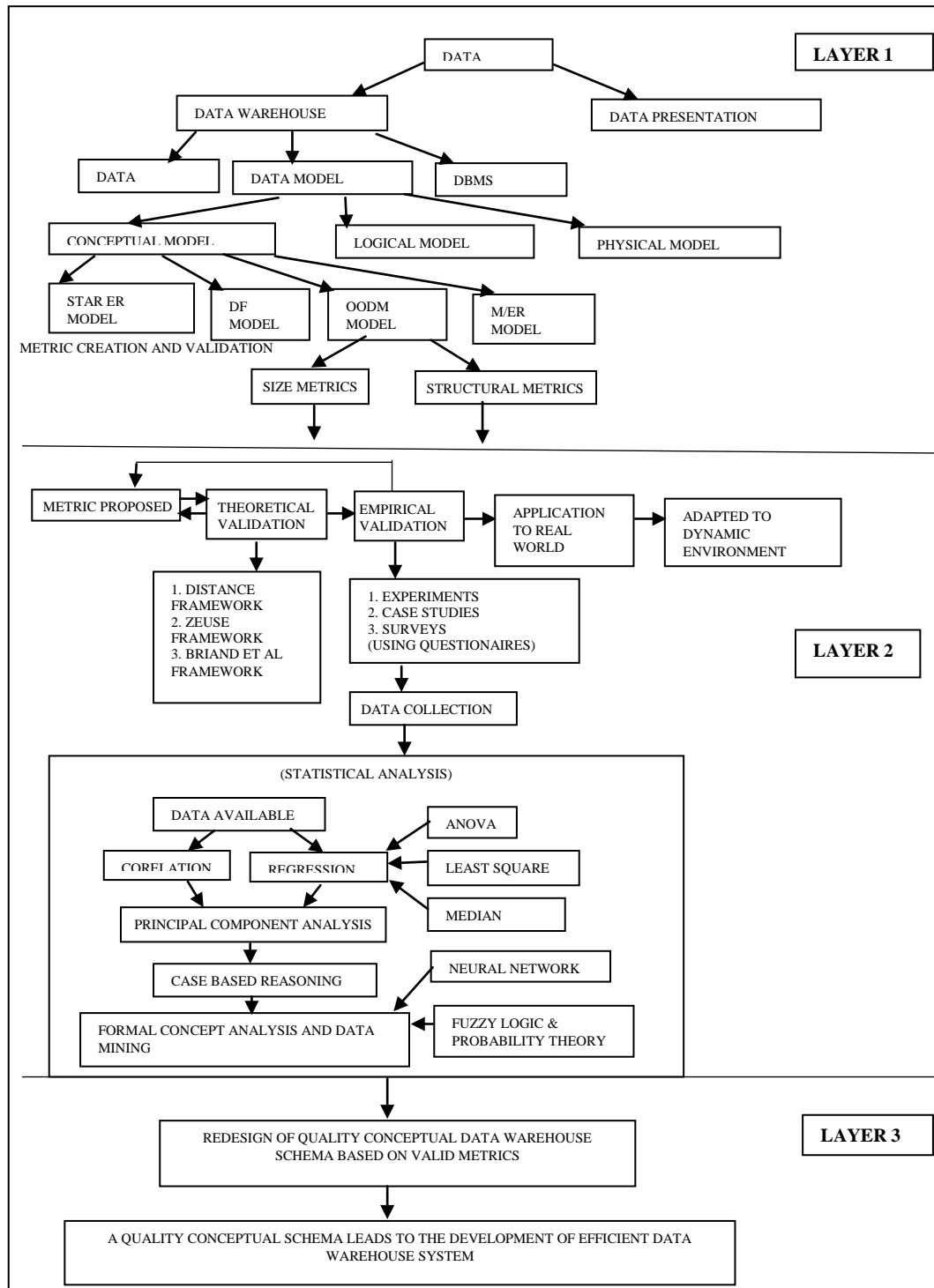


Figure 3.2 Proposed Detailed Research Framework for Development of Efficient Data Warehouse System

- Framework based on measurement theory: This approach is based on ratio scale and this information can be used to define the transformations that can be applied to proposed metrics. DISTANCE and Zuse framework are based on measurement theory.

After theoretical validation, the next step is empirical validation of metrics [11, 16] to prove their utility and relevance. This process is iterative as the proposed metric can be discarded or redefined depending on the results of empirical validation. Empirical validation is carried out using experiments, case studies and surveys. The data collected from the above mentioned techniques is applied to statistical analysis. The various techniques for carrying out statistical analysis are following:

- Correlation is used to check whether there exists a relationship between independent and dependent variables.
- Regression is to determine whether the relationship between constructs (independent and dependent variables) is linear or not. The various regression techniques are ANOVA, least square, median square analysis.
- Principal Component Analysis: This technique finds only the principal components that can be used for explaining the model without loss of significant information.
- Case Base Reasoning: It is used to find the most similar models having same values of metrics.

The empirical validation gives valid set of metrics which can be applied to real world applications. The redundant metrics can be discarded at this stage. The valid metrics can be redefined for adapting to the changing dynamic environment of real world projects.

- **Layer 3** takes as input the valid metrics obtained from Layer 2. Taking into consideration these valid set of metrics the conceptual model can be redesigned. The quality of conceptual model can be measured in terms of understanding time, efficiency and effectiveness. A good quality conceptual data warehouse model leads to the development of a good logical model and physical model. In this way regulating the quality of conceptual model leads to the development of an efficient data warehouse system.

Based on the framework proposed, layer 2 was made focus of further research. The study prompted the way for a new quality metric which has a significant contribution towards quality evaluation of conceptual models. The proposed metric is discussed in detail in next section.

3.3NRFD (NUMBER OF RELATIONS BETWEEN FACTS AND DIMENSIONS), NEW PROPOSED METRIC AND ITS THEORETICAL VALIDATION

Quality metrics proposed by various researchers for different configurations of data warehouse systems namely table metrics, star metrics and schema metrics were studied. Most recent studies focused on schema metrics. It was identified that Manuel Serrano has been consistently working in the conceptual data warehouse domain and has mainly concentrated his work on quality evaluation. To add to his credit, he has published his research contributions in various journals of repute. The quality metrics proposed and validated by Manuel Serrano were taken as base for further research study.

Serrano et al [11, 16] discuss metrics based on structural properties of conceptual models. These metrics are defined below:

- NDC: Number of dimension classes of the model.
- NFC: Number of fact classes of the model.
- NBC: Number of base classes of model.
- NC: Total number of classes of the model which includes fact classes, dimension classes and base classes.
- RBC: Ratio of base classes. Number of base classes per dimension class of model.
- NAFC: Number of attributes of the fact class of the model.
- NADC: Number of dimension attributes of the dimension classes.
- NABC: Number of dimension attributes of the base classes of the model.
- NA: Total number of attributes of the model which includes fact class attributes, dimension class attributes and base class attributes.
- NH: Number of hierarchy relationships of the model.
- DHP: Maximum depth of the hierarchy relationships of the model.

- RSA: Ratio of attributes of the model. Number of fact attributes divided by the number dimension attributes.

These metrics play a significant role towards quality evaluation of conceptual models.

3.3.1 How the Idea Generated

The study of quality metrics at the conceptual level gave necessary motivation to propose a new metric based on structural properties of model. During the study, it was found that a situation might arise when one has to find best possible configuration from existing conceptual models for same domain. All of the proposed metrics [16] may have the same values for several existing conceptual models for same domain. This tempted a thought that some other quality factors may exist that can affect the understandability of models for which values of all the proposed metrics mentioned above is same. Two scenarios might arise. In the first scenario, the models may have same values of quality metrics discussed above; the understanding times of the models might be same. A second scenario might exist where the understanding time of the models may vary. This situation made the need for further in-depth study of several models and find the variations in their structural properties. A rigorous study of the existing literature lead towards proposal of a new quality metric that affects the understanding times of given conceptual models, other than the quality metrics already proposed. The metric namely number of relations existing between fact classes and dimension classes within a conceptual model is proposed and discussed in detail in successive sections. The metric is proposed keeping in mind that as the number of relations between fact class and dimension classes increases, it leads to increase in structural complexity and understanding time of the conceptual models. Figure 3.3 shows the proposed metric (NRFD) along with existing metrics.

Figure 3.3 shows the quality metrics (discussed above) proposed by Serrano et al [16], in oval boxes. The new proposed quality metric namely NRFD (Number of relations between fact class and dimension class of conceptual data warehouse model) is shown in box with hexagonal shape. At the conceptual level, designer considers the entity names and their relationships while details regarding primary and foreign keys are discussed [11] at the higher level which is logical level of design. The concept of keys leads to complexities at conceptual level considering the fact that, when a fact and

dimension table is joined, the primary keys of dimension table become the foreign keys in fact table.

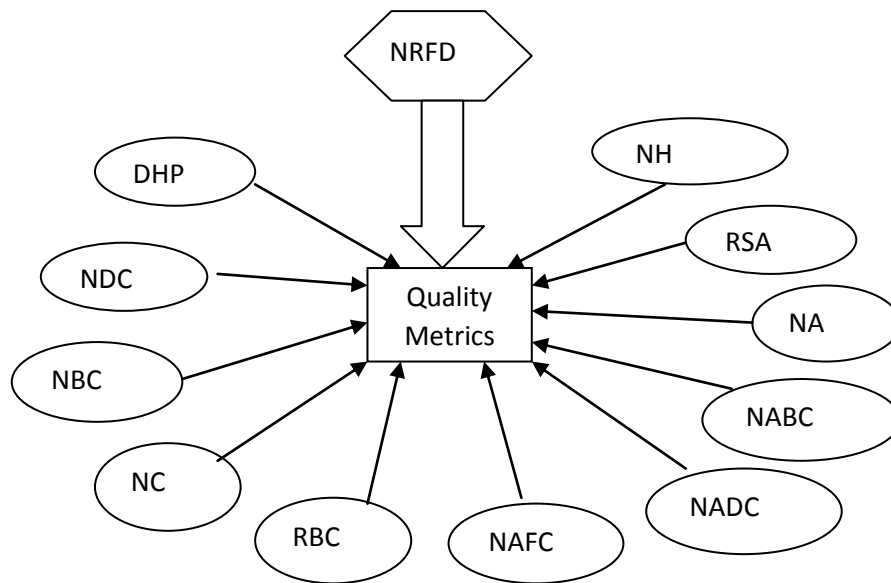


Figure 3.3 Quality metrics along with proposed metric
Adapted from Serrano et al., 2007 [16]

A situation might arise where the attribute for a record in dimension table changes with time. To maintain the past and current values of the particular record, duplication of primary key for particular record takes place. This gives rise to surrogate keys. The surrogate keys now act as primary keys. This change must also be reflected in fact tables to which the dimension tables are linked. This procedure creates a lot of complexities. A relationship between facts and dimensions is a better approach to deal with structural complexities at conceptual level than the concept of keys which is the prime reason of proposing a metrics at conceptual level.

3.3.2 Importance of Proposed Metric

The section presents the various issues that can be solved using proposed metric in respect of conceptual data warehouse models. The conceptual model [108] consists of fact classes and related dimension classes. The fact class contains subjects important from point of view of business executives, managers, customers and users. The facts are analysed along various multiple dimensions. Dimensions provide a way to measure facts. A number of relationships exist between facts and dimensions. The increase in the number of relations between dimension classes and facts classes, leads

to an increase in the structural complexity of data warehouse models. With the increase in structural complexity [54], it is predicted that understanding time of the model also increases in accordance. Understanding time is directly proportional to structural complexity of data warehouse conceptual model. The proposed metric aims to investigate following issues based on its importance:

- Is there any difference between the structural complexities of data warehouse conceptual models from a single domain that have same values for all the metrics mentioned in Serrano et al [16]?
- Do there exists any quality factors that can measure the difference in structural complexity and hence understandability of several conceptual models from a single domain that have same values for all the metrics mentioned in Serrano et al [16]?
- What are the techniques that can prove the correctness of quality factors measuring the difference in structural complexity of several conceptual models from a single domain?
- Are the quality factors other than those mentioned in Serrano et al [16], playing any significant role in making an efficient data warehouse system?

3.3.3 Metric Creation

The section, to describe the various characteristics of multidimensional model at conceptual level using UML [39, 41]. Figure 3.4 and 3.5 show the structural properties by means of a class diagram in which sales and store billing of certain items in a store are represented by means of fact classes and related dimension classes. The fact contains store_billing_fact and sales fact class. The facts are measured along main dimensions such as employees, store, supply, date, orders, product, customer, geography and store_bill. Each of the fact and dimension class contains respective attributes, OID (Object id's), primary keys and foreign keys. The facts and dimension classes are associated with each other with relationships of different cardinalities. The dimension classes are associated with base classes namely department, store_region, suppliestype, customer type, address, region type, brand and category via relationships depicting hierarchies. The only difference in Figure 3.4 and 3.5 is the number of relations between dimension classes and fact classes. The

relations are labeled in the Figure 3.4 and 3.5 as E1, E2 ... corresponding to edge1, edge2. The number of fact and dimension classes is same in both the figures.

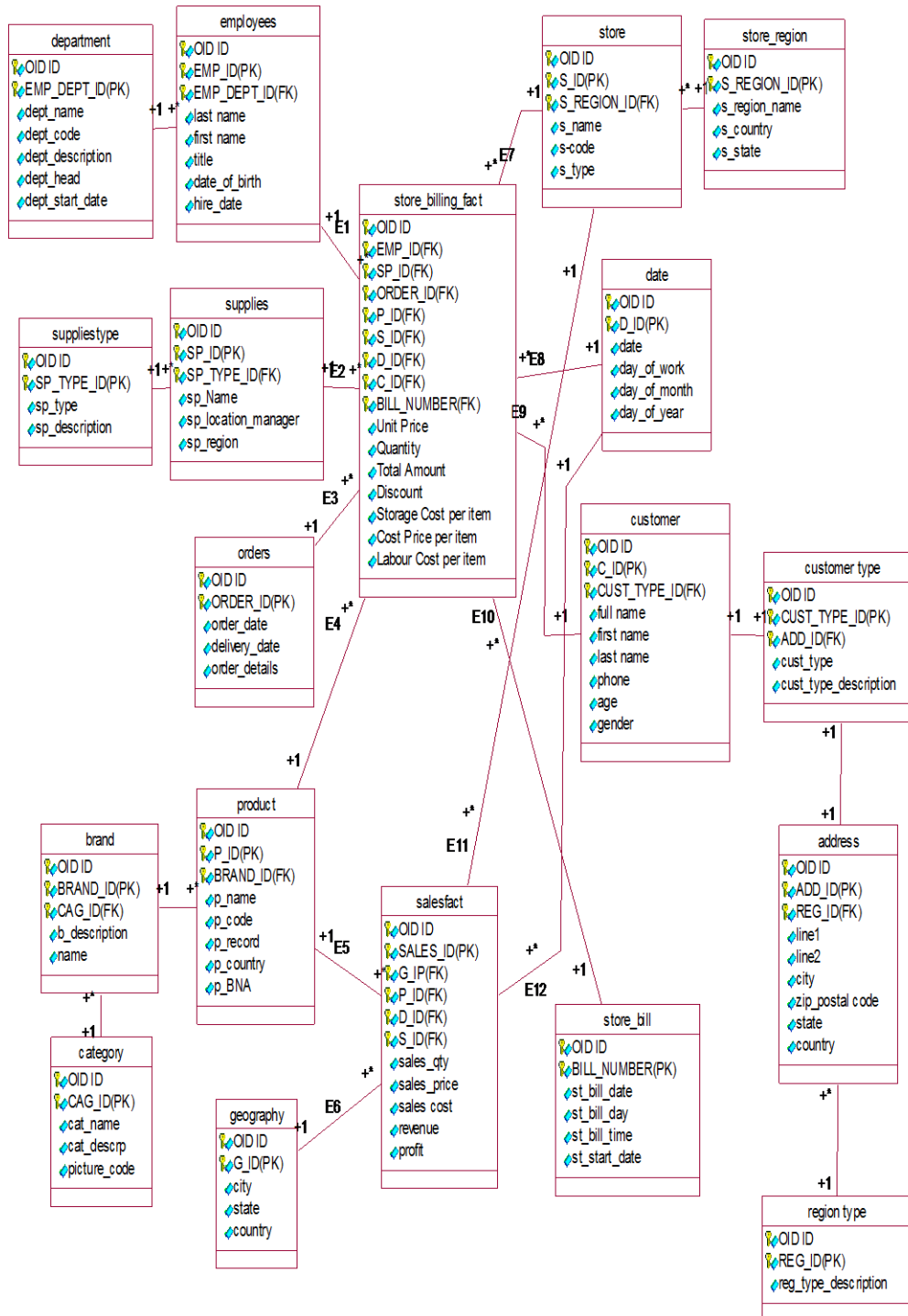


Figure 3.4 A conceptual model showing sales of items of a store
Idea adapted from source: inmoncif.com [109]

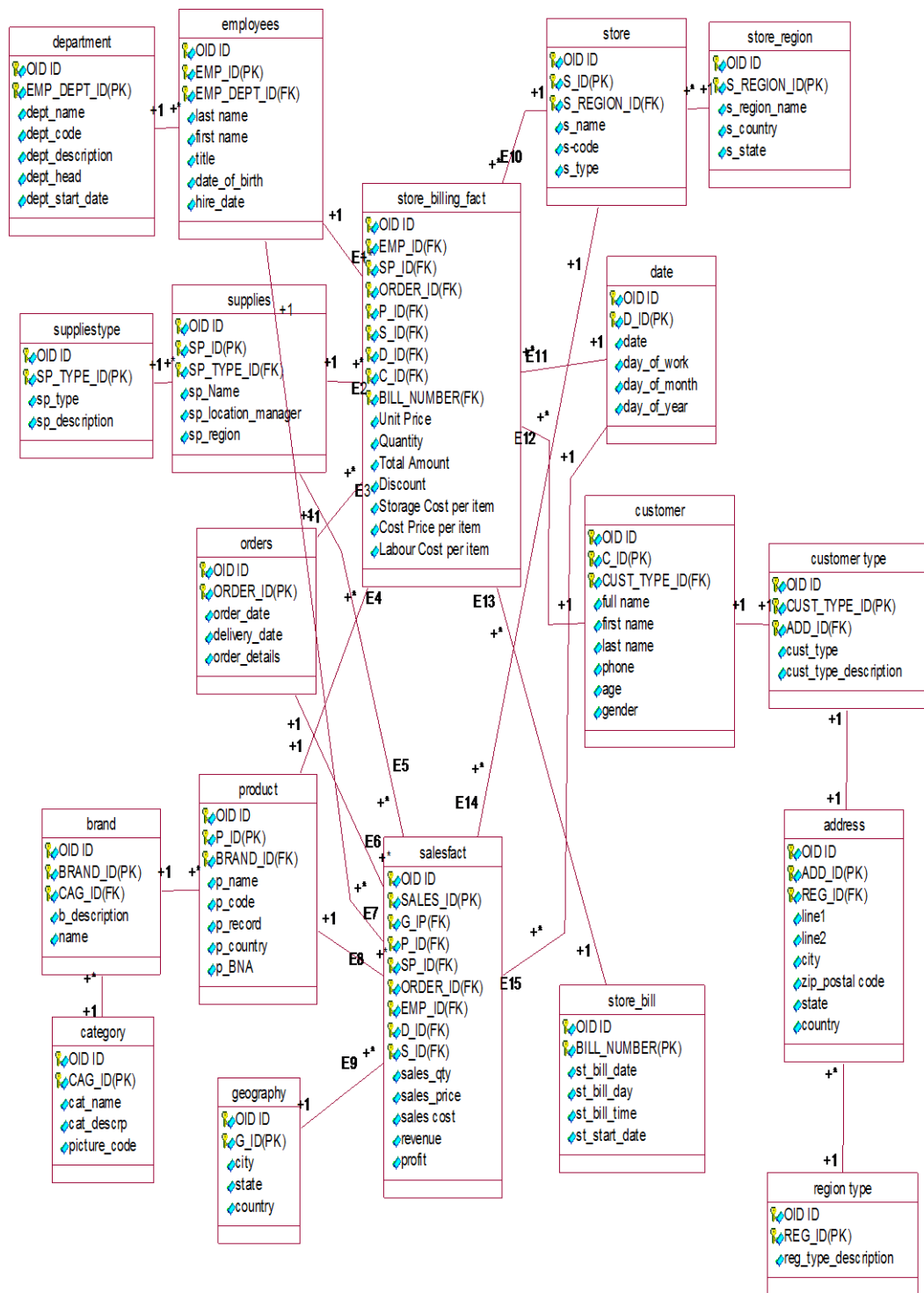


Figure 3.5 A conceptual model showing sales of items of a store
Idea adapted from source: inmoncif.com [109]

3.3.4 Theoretical Validation

The proposed metric validation is carried out using the DISTANCE framework approach. DISTANCE [16] framework guarantees that the metrics defined and validated using the framework are in a ratio scale. The theoretical validation of metric using DISTANCE framework aims to provide answer to following questions:

- Does the proposed metric provide a measure (numeric value) to evaluate specific attribute of data warehouse conceptual model? The specific attribute measured in terms of proposed metric is understandability based on the structural complexity of model.
- Is the metric capable enough to transform one configuration of data warehouse conceptual model to other on application of finite sequence of elementary transformations?

This distance based measure construction process as discussed by Serrano et al [16] consists of five steps:

- Step 1. Find a measurement abstraction: The step aims to map data warehouse conceptual model onto its set of relationships between facts and dimensions. The output of this step is a set of measurement abstractions M containing existing relationships between facts and dimensions.
- Step 2. Model distances between measurement abstractions: The step outputs a set of elementary transformations that can transform relationship sets of one model to relationship sets of other model. The input is taken in the form of measurement abstractions obtained in previous step.
- Step 3. Quantify distances between measurement abstractions: This step aims to give a count of shortest possible elementary transformations to transform relationship set of one model to relationship set of other model.
- Step 4. Find a reference abstraction: To generalize the approach that can be applicable to any number of conceptual data warehouse models a base case having lowest possible value of proposed metric is identified and is output of this step.
- Step 5. Define the software measure: This step gives the numerical count of the proposed metric measured with respect to base case as identified in previous step for any data warehouse conceptual model.

The Number of Relationship between fact and dimension classes (NRFD) is defined as the total number of relations/edges between fact and dimension classes within a data warehouse conceptual model. The models given in Figure 3.4 and 3.5 are used for reference while validating this new metric. There is equal number of fact and dimension classes and all the metrics discussed earlier (Figure 3.3) have same values for both the Figures 3.4 and 3.5.

- Step 1 Find a measurement abstraction

We define the set of software entities P as the Universe of data warehouse conceptual models (UDCM) that is relevant for some Universe of Discourse (UoD) and p is a Data warehouse Conceptual Model (DCM) where $p \in \text{UDCM}$.

Let URC be the Universe of Relations between facts and dimensions relevant to the UoD. The set of relations/edges between fact and dimension classes within a DCM, called SR (DCM) is a subset of URC. All the sets of relations between Fact and dimension classes within the DCM of UDCM are elements of the power set of URC, denoted by $\wp(\text{URC})$. The set of measurement abstractions M can be equated to $\wp(\text{URC})$ and the abstraction function can be defined as:

$$\text{ab}_{\text{NRFD}} : \text{UDCM} \rightarrow \wp(\text{URC}) : \text{DCM} \rightarrow \text{SR}(\text{DCM})$$

This function (Serrano et al., 2007) maps the DCM onto a set of the relations/edges between fact and dimension classes. The set of relations between fact and dimension classes for DCM A (from Figure 3.4) and DCM B (from Figure 3.5) is given as:

$$\begin{aligned} \text{ab}_{\text{NRFD}}(\text{DCM A}) &= \text{SR}(\text{DCM A}) = \{E1, E2, E3, E4, E5, E6, E7, E8, E9, E10, \\ &\quad E11, E12\} \\ \text{ab}_{\text{NRFD}}(\text{DCM B}) &= \text{SR}(\text{DCM B}) \\ &= \{E1, E2, E3, E4, E5, E6, E7, E8, E9, E10, E11, E \\ &\quad 12, E13, E14, E15\}. \end{aligned}$$

similar abstraction functions are referred by Serrano et al [16].

- Step 2 Model distances between measurement abstractions

The next step is to model distances between the elements of DCM. There is a need to find a set of elementary transformation types as stated by Poels and Dedene [17] for the set of measurement abstractions $\wp(\text{URC})$ such that any set of relations between

fact and dimension classes can be transformed into another set of relations between fact and dimension by means of a finite sequence of elementary transformations. Since the elements of p (URC) are relations between all the classes of the model, 'Te' must only contain two types of elementary transformations: one for adding a relation/edge between classes to a set and one for removing a relation/edge between classes from a set. Given two sets of relations between classes:

$s1 \in p$ (URC) and $s2 \in p$ (URC)

$s1$ can always be transformed into $s2$ by removing first all the relations from $s1$ that are not in $s2$, and adding all the relations to $s1$ that are in $s2$, but were not in the original $s1$ [16]. Formally, $Te = \{t_{0-NRFD}, t_{1-NRFD}\}$, where t_{0-NRFD} and t_{1-NRFD} are defined as:

$t_{0-NRFD} : p$ (URC) \rightarrow p (URC) : $s \cup \{a\}$, with $a \in$ URC

$t_{1-NRFD} : p$ (URC) \rightarrow p (URC) : $s - \{a\}$, with $a \in$ URC

- Step 3 Quantify distances between measurement abstractions

In this step the count of the shortest possible sequence of elementary transformations that can transform one set of relationship sets to other is generated as output. A function \bar{G}_{NRFD} that quantifies these distances is the metric (in the mathematical sense) that is defined by the symmetric difference model.

$\bar{G}_{NRFD} : p$ (URC) * p (URC) \rightarrow $R : (s, s') \rightarrow |s - s'| + |s' - s|$

Mathematically, using Figure 3.4 and 3.5 as references, the function can be calculated as follows:

$$\begin{aligned} & \bar{G}_{NRFD}(\text{abs}_{NRFD}(\text{DCM A}), \text{abs}_{NRFD}(\text{DCM B})) \\ &= | \{E1, E2, E3, E4, E5, E6, E7, E8, E9, E10, E11, E12\} - \\ & \quad \{E1, E2, E3, E4, E5, E6, E7, E8, E9, E10, E11, E12, E13, E14, E15\} | + \\ & \quad | \{E1, E2, E3, E4, E5, E6, E7, E8, E9, E10, E11, E12, E13, E14, E15\} - \\ & \quad \{E1, E2, E3, E4, E5, E6, E7, E8, E9, E10, E11, E12\} | \\ &= | \{ \} | + | \{E13, E14, E15\} | \\ &= 3 \end{aligned}$$

- Step 4 Find a reference abstraction

To generalize the application to any number of data warehouse conceptual models a base case or reference is identified. The reference point for measurement is the empty

set of dimension classes which contains no relation between fact and dimension classes. A DCM without dimension classes will have the lowest possible value for the NRFD measure. We define the following function:

$$\text{ref}_{\text{NRFD}} : \text{UDCM} \rightarrow \text{p(URC)} : \text{DCM} \rightarrow \emptyset$$

- Step 5 Define the software measure

This step aims to provide a numerical count of the proposed metric in reference to the base case identified in previous step. The count is calculated as the distance between its set of relations $\text{SR}(\text{DCM})$ and the empty set of dimension classes \emptyset with no relation between facts and dimensions. Hence, the NRFD measure can be defined as a function that returns for any $\text{DCM} \in \text{UDCM}$ the value of the metric G_{NRFD} for the pair of sets $\text{SR}(\text{DCM})$ and \emptyset :

$$\begin{aligned} \text{DCM} \in \text{UDCM}: \text{NRFD}(\text{DCM}) &= \text{G}_{\text{NRFD}}(\text{SR}(\text{DCM}), \emptyset) \\ &= |\text{SR}(\text{DCM}) - \emptyset| + |\emptyset - \text{SR}(\text{DCM})| \\ &= |\text{SR}(\text{DCM})| \end{aligned}$$

The measure $\text{SR}(\text{DCM})$ returns measurable mathematical value of the total number of relations between fact and dimension classes in a data warehouse conceptual model. We aimed to define the metric, number of relations between facts and dimensions, in measurable mathematical form and to check the capability of proposed metric in transforming one conceptual model to other. This proves the validity of proposed metric NRFD theoretically as the metric based on structural complexity of conceptual model can be measured in terms of mathematical values and has the capability to transform one set of relationships between facts-dimensions to other set.

The next step following theoretical validation is empirical validation of proposed metric along with existing metrics, as proposed by Serrano et al [16], to check their practical utility in quality evaluation of conceptual data warehouses. The following section discusses empirical validation in detail.

3.4 EMPIRICAL VALIDATION

This phase succeeds theoretical validation phase, where the metrics proved to be theoretically correct are subjected to empirical validation techniques. Empirical validation techniques involve case studies, surveys and experimental techniques. The

output of empirical validation decides the acceptance, redefinition or rejection of proposed metric. A controlled experiment was conducted with 80 students, who worked on 22 conceptual models, aiming at evaluating 13 quality metrics with regard to their power in predicting the understandability of conceptual models. Questionnaire was designed for each conceptual model and the time taken by each student in answering the questions correctly was recorded. The role played by each of the quality metric in predicting the understanding time of conceptual models was analysed using statistical techniques. The details of controlled experiment conducted for empirical validation are presented in successive sections.

3.4.1 Preliminaries

This section provides a demonstration to calculate values of metrics, defined in section 3.2, for a conceptual data warehouse model for analysing manufactured part of cars along multiple dimensions namely plant, supplier and package shown in Figure 3.6.

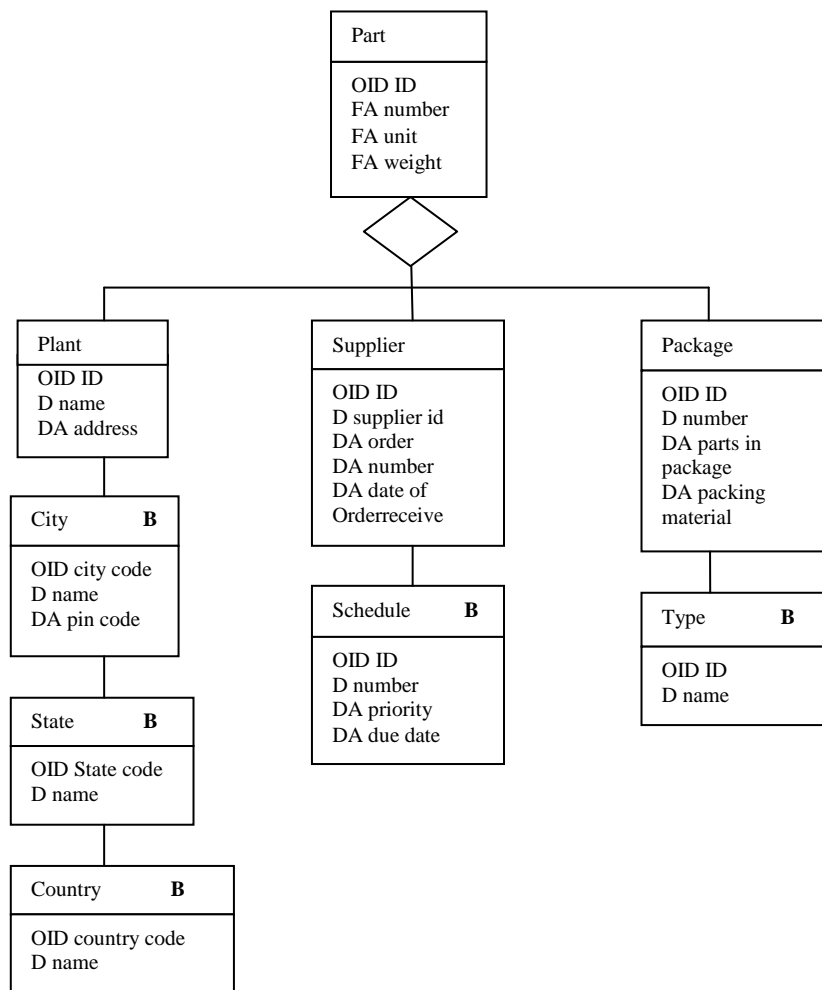


Figure 3.6 UML Class Diagram for manufacturing parts

Part is fact class which contains specific measures called fact attributes namely number, unit and weight of a part to be analysed. Plant, supplier and package are dimension classes. City, state, country, schedule, type are base classes for the model shown in Figure 3.6.

The values of quality metrics for UML class diagram of manufacturing parts is shown in Table 3.1.

Table 3.1 Metric Values

| Metrics | NFC | NDC | NBC | NC | RBC | NAFC | NADC | NABC | NA | NH | DHP | RSA | NRFD |
|---------|-----|-----|-----|----|-----|------|------|------|----|----|-----|-----|------|
| Values | 1 | 3 | 5 | 9 | 1.6 | 3 | 9 | 8 | 20 | 3 | 3 | .18 | 3 |

The value of NFC is 1 corresponding to single fact class part. The value of NDC is 3 corresponding to 3 dimension classes plant, supplier and package. The value of NBC is 5 corresponding to 5 base classes city, state, country, schedule and type. The value of NC is 9 as it is total of NFC, NDC and NBC. The value of RBC is 1.6 as it is number of base class per dimension class i.e. $5/3 = 1.6$. The value of NAFC is 3 as there are 3 attributes of fact class part namely number, unit and weight. The value of NADC is 9 corresponding to all the D (descriptor) and DA (dimension attributes) for the dimension classes plant, supplier and package. The value of NABC is 8 corresponding to all the D (descriptor) and DA (dimension attributes) for the base classes city, state, country, schedule and type. The value of NA is 20 corresponding to a total sum of attributes of fact, dimension and base classes. The base classes associated with dimension classes contribute towards NH. The value of NH for model is 3 as base classes are associated with each of the three dimension classes. The value of DHP is 3 as the maximum number of base classes (city, state, country) associated with a dimension class (plant) is 3. The value of RSA is 0.18. The value of NRFD is 3 as the fact class (part) is associated with three dimension classes (plant, supplier, package).

3.5 EXPERIMENTAL SETUP

To investigate the role of quality metrics in data warehouse design a controlled experiment was conducted. A total of 80 students, 22 conceptual models and 13 quality metrics were included in the experiment. This section incorporates the

necessary details like goal, models, subjects and hypothesis for testing the validity of experimental study.

3.5.1 Goal

The goal of empirical validation is to analyse the quality metrics and evaluate them for predicting the understandability of conceptual data warehouse models.

3.5.2 Model

Twenty-two conceptual models were used to perform the experiment. The models depict data from real world applications of different domains like banking, airlines, universities, insurance, medical, railways. Metric values for all the 22 models are given in Table 3.2. The model selection was based on structural complexity of conceptual models rather than domain categorization to predict the understandability because the values of metrics are dependent on structural attributes of models and are domain independent.

Table 3.2 Table of metrics for models

| | NFC | NDC | NBC | NC | RBC | NAFC | NADC | NABC | NA | NH | DHP | RSA | NRFD |
|-----|-----|-----|-----|----|------|------|------|------|----|----|-----|-----|------|
| S01 | 1 | 4 | 9 | 14 | 2.25 | 4 | 12 | 14 | 30 | 4 | 3 | .15 | 4 |
| S02 | 1 | 5 | 13 | 19 | 2.60 | 3 | 14 | 14 | 31 | 5 | 4 | .10 | 5 |
| S03 | 1 | 5 | 11 | 17 | 2.20 | 3 | 13 | 12 | 28 | 5 | 3 | .12 | 5 |
| S04 | 1 | 3 | 7 | 11 | 2.23 | 2 | 8 | 9 | 19 | 3 | 3 | .11 | 3 |
| S05 | 1 | 4 | 8 | 13 | 2.00 | 4 | 11 | 9 | 24 | 4 | 3 | .20 | 4 |
| S06 | 1 | 4 | 10 | 15 | 2.50 | 3 | 13 | 10 | 26 | 4 | 3 | .13 | 4 |
| S07 | 1 | 3 | 5 | 9 | 1.67 | 3 | 9 | 8 | 20 | 3 | 3 | .17 | 3 |
| S08 | 1 | 4 | 7 | 12 | 1.75 | 4 | 13 | 8 | 25 | 4 | 3 | .19 | 4 |
| S09 | 1 | 6 | 6 | 13 | 1.00 | 2 | 15 | 13 | 30 | 4 | 3 | .07 | 6 |
| S10 | 1 | 4 | 8 | 13 | 2.00 | 2 | 10 | 8 | 20 | 4 | 3 | .11 | 4 |
| S11 | 2 | 9 | 6 | 17 | .66 | 10 | 36 | 20 | 66 | 4 | 3 | .18 | 13 |
| S12 | 2 | 9 | 6 | 17 | .66 | 10 | 36 | 20 | 66 | 4 | 3 | .18 | 15 |
| S13 | 2 | 9 | 8 | 19 | .88 | 12 | 36 | 23 | 71 | 5 | 3 | .20 | 12 |
| S14 | 2 | 9 | 8 | 19 | .88 | 12 | 36 | 23 | 71 | 5 | 3 | .20 | 15 |
| S15 | 1 | 22 | 0 | 23 | 0 | 5 | 30 | 0 | 35 | 0 | 0 | .15 | 22 |
| S16 | 1 | 3 | 0 | 4 | 0 | 4 | 25 | 0 | 29 | 0 | 0 | .16 | 3 |
| S17 | 1 | 5 | 1 | 7 | .20 | 12 | 10 | 3 | 25 | 1 | 1 | .92 | 5 |
| S18 | 1 | 6 | 0 | 7 | 0 | 1 | 20 | 0 | 21 | 0 | 0 | .05 | 6 |
| S19 | 1 | 6 | 0 | 7 | 0 | 6 | 23 | 0 | 29 | 0 | 0 | .26 | 6 |
| S20 | 1 | 4 | 0 | 5 | 0 | 4 | 14 | 0 | 18 | 0 | 0 | .14 | 4 |
| S21 | 3 | 6 | 0 | 9 | 0 | 11 | 47 | 0 | 58 | 0 | 0 | .23 | 11 |
| S22 | 1 | 5 | 0 | 6 | 0 | 3 | 33 | 0 | 36 | 0 | 0 | .09 | 5 |

3.5.3 Subjects

Eighty students pursuing B.Tech. in the institute participated as volunteers in the experiment. These students were in third year of their degree course in CSE and IT streams and all were in similar age group of 20-21 years. All the volunteers had adequate knowledge of data warehousing and UML concepts because they were studying the subject 'Data warehousing and data mining' as a part of their course curriculum in third year. The participation of the subjects was taken up voluntarily and it was apart from their course curriculum. The experiment was conducted in two separate rooms with strength of 40 students in each room. A supervisor was appointed for monitoring the students in each room. Before the start of experiment the students were given a description of the tasks to be performed and a tutorial to brush up their related concepts. A sample model was taken, calculation of values of metrics from model was shown, how the questions were to be answered for the sample, where and in what format the answers were to be placed, where and in what format the starting time/ending time of the tasks were to be recorded. The students were given the printed hard copy of samples along with questionnaires in a variable order. The seating arrangement in each room was such that every two consecutive student had different set of models to be answered. Each of the alternate students was given a set 10 models first and the other was given other 12 models. After 1 hour, the set of models were interchanged. A wall clock was installed in each room for recording the start and end time of each task.

3.5.4 Hypothesis

The main hypothesis of experiment was classified into following sub hypothesis:

- Null Hypothesis H_{01} : Quality metrics have no impact/contribution towards prediction of understandability of conceptual data warehouse models.
- Null Hypothesis H_{02} : All the principal components of the model summary are significant to predict the understandability of models.
- Null Hypothesis H_{03} : The models having similar values of quality metrics do not have any relation in respect of their understanding times.
- Alternate Hypothesis H_{01} : Quality metrics have significant effect/contribution towards prediction of understandability of conceptual data warehouse models.
- Alternate Hypothesis H_{02} : Not all the principal components of the model

summary are significant to predict the understandability of models.

- Alternate Hypothesis H_{03} : The models having similar values of quality metrics have significant relation in respect of their understanding times.

3.6 EMPIRICAL DATA COLLECTION

Two types of variables are used in the study namely dependent and independent variables.

3.6.1 Independent Variables

Independent variables are those variables whose values do not change throughout the course of experiment. These are the variables for which effect is evaluated. The impact of independent variables on dependent variables is studied in this experiment. In the study, the 13 quality metrics in Table 3.2 form the independent variables [79].

3.6.2 Dependent Variables

Dependent variables [79] are the variables whose values vary according to change in values of independent variables. Understanding time, the time taken by each subject to understand and answer the questions of each model correctly, is the dependent variable in our study. Small values of understanding time predict better understandability and large values of understanding time predict non-understandability.

3.6.3 Data Validation

A set of 4 questions based on each model was given to each one of the 80 subjects (participants). The participants had to analyse each model and answer specific questions for the particular model. The questions were designed for each model keeping in view that they were based on quality measures to predict the understandability and not on the domain of models. Domains of all models were familiar and known to subjects. Each of the 4 questions for a model was based on different level of understandability and therefore time taken to answer each question was different. The selected models had different structural complexities, so the time taken to answer questions varied from one model to another. The set of 4 questions for model given in Figure 3.6 is as follows:

Record Starting Time: (ss).....

1. Enumerate which classes are required for knowing the packing material of a part?
2. Enumerate which classes are required for knowing the priority of one schedule?
3. Enumerate which classes are required for knowing the address of one plant?
4. If we want to increase the manufacturing of a part then in which class do we add the increased raw material?

Record ending time: (ss).....

The starting time and ending time to answer the questions for each model was noted down in seconds. The difference of starting and ending time gives the time taken to answer questions for a particular model. This time is the respective understanding time for a model. Out of 80 students 11 students were unable to complete all the tasks, 6 students gave incorrect answers for nearly 80 percent of the models and 3 students recorded exceptional times for completing the tasks. The collected average understanding time of 22 models for 60 subjects is shown in Table 3.3. The descriptive statistics showing the minimum, maximum and average understanding time along with standard deviation is shown in Table 3.3. The values under column Minimum gives the minimum time taken by some subject to understand a model and the values under column Maximum gives the maximum time taken by some subject to understand a model. As seen from Table 3.3 the lowest minimum time taken by subject to understand is 19 seconds for model 6 and highest minimum time taken by subject to understand is 139 seconds for model 12. Similarly, the lowest maximum time taken by subject to understand is 121 seconds for model 20 and the lowest maximum time taken by subject to understand is 412 seconds for model 14. The Mean column gives the average understanding time of each model. It is lowest for model 20 and highest for model 14. The Std. Deviation column gives the average deviation from the mean. It is lowest for model 1 and highest for model 14.

3.7 RESULT ANALYSIS

In this section, details of the statistical analysis performed on collected data to test the hypothesis is given. Statistical techniques like Correlation, Regression, PCA, Nearest Neighbour Analysis and ROC have been used to analyse the conceptual models. The descriptive statistics of the metrics and understanding time are already given in Table 3.2 and Table 3.3. The following section explains the analysis results for each technique.

Table 3.3 Descriptive Statistics for Understanding Time

| | Minimum | Maximum | Mean | Std. Deviation |
|-----|---------|---------|------|----------------|
| S01 | 65 | 143 | 103 | 17 |
| S02 | 75 | 183 | 121 | 24 |
| S03 | 45 | 160 | 106 | 28 |
| S04 | 37 | 170 | 96 | 31 |
| S05 | 20 | 325 | 94 | 42 |
| S06 | 19 | 270 | 96 | 36 |
| S07 | 39 | 154 | 88 | 25 |
| S08 | 28 | 325 | 98 | 38 |
| S09 | 40 | 195 | 95 | 30 |
| S10 | 43 | 260 | 99 | 33 |
| S11 | 110 | 305 | 209 | 34 |
| S12 | 139 | 395 | 272 | 39 |
| S13 | 98 | 343 | 224 | 56 |
| S14 | 75 | 412 | 295 | 61 |
| S15 | 57 | 180 | 97 | 23 |
| S16 | 43 | 180 | 104 | 33 |
| S17 | 39 | 231 | 126 | 36 |
| S18 | 36 | 169 | 91 | 31 |
| S19 | 45 | 170 | 100 | 27 |
| S20 | 35 | 121 | 82 | 21 |
| S21 | 93 | 256 | 164 | 36 |
| S22 | 43 | 148 | 96 | 28 |

3.7.1 Correlation Analysis

Pearson correlation was applied to find whether any correlation exists between independent variables (metrics) and dependent variables (understanding time). The level of significance α (alpha) was taken to be 0.05. Pearson correlation coefficient is obtained by dividing the covariance of the two variables by the product of their standard deviations. The correlation coefficient $\rho_{X,Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:

$$\rho(x, y) = \text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sigma(x) * \sigma(y)} = \frac{E[(x - \mu(x))(y - \mu(y))]}{\sigma(x) * \sigma(y)}$$

Where E is the expected value operator, cov means covariance and, corr is notation for Pearson's correlation. The closer the coefficient is to either -1 or 1, the stronger

the correlation between the variables. If the variables are independent, Pearson's correlation coefficient is 0.

The correlation of each individual metric with understanding time was calculated. The results of correlation are given in Table 3.4.

Table 3.4 Pearson correlation analysis

| Metrics | NDC | NBC | NC | RBC | NAFC | NADC | NABC | NA | NH | DHP | RSA | NRFD | NFC |
|-------------|-------------|-------------|------|-------------|------|------|------|------|------|-------------|-------------|------|------|
| Correlation | .323 | .186 | .474 | -.168 | .808 | .687 | .710 | .928 | .752 | .307 | .058 | .647 | .764 |
| Sig. | .143 | .407 | .026 | .454 | .000 | .000 | .000 | .000 | .000 | .165 | .799 | .001 | .000 |

It can be seen from Table 3.4 that metrics NC, NAFC, NADC, NABC, NA, NH, NRFD and NFC are correlated with understanding time. Out of these 8 metrics 5 metrics NAFC, NADC, NABC, NH and NFC are showing high degree of correlation with understanding time with significance value of 0.000. The Sig. greater than 0.05 are bold in Table 3.4, showing that the metrics NDC, NBC, RBC, DHP and RSA are not correlated with understanding time. The results of Table 3.4 supports alternate hypothesis H_{01} proving that some of the quality metrics are significantly correlated with understanding time and rejects null hypothesis H_{01} .

3.7.2 Regression Analysis

Linear regression was performed for analysing the effect of each individual metric (independent variable) on understanding time (dependent variable) at level of significance $\alpha=0.05$. The equation used in regression analysis is:

$$Y_j = a + X_{1j}b_1 + X_{2j}b_2 + \dots + X_{kj}b_k$$

- Where,
- a, b_1, b_2, \dots, b_k = regression coefficients.
 - $X_{1j}, X_{2j}, \dots, X_{kj}$ = independent variables.
 - K = number of independent variables
 - Y_j = predicted value of dependent variable.

Using the above equation regression coefficient was computed. If the regression coefficient is close to one, it means that independent variables have significant role towards prediction of dependent variable and values close to 0 defines a lower degree of significance of independent variables towards prediction of dependent variables.

Table 3.5 Univariate Linear Regression

| Model | Sum of Squares | Degree of freedom | Mean Square | F | Sig. |
|-------------------|----------------|-------------------|-------------|--------------|------|
| NDC | | | | | |
| Regression | 8470.645 | 1 | 8470.645 | 2.322 | .143 |
| Residual | 72956.394 | 20 | 3647.820 | | |
| Total | 81427.039 | 21 | | | |
| NBC | | | | | |
| Regression | 2824.214 | 1 | 2824.214 | .719 | .407 |
| Total | 81427.039 | 21 | | | |
| NC | | | | | |
| Regression | 18257.830 | 1 | 18257.830 | 5.781 | .026 |
| Residual | 63169.209 | 20 | 3158.460 | | |
| Total | 81427.039 | 21 | | | |
| RBC | | | | | |
| Regression | 2309.584 | 1 | 2309.584 | .584 | .454 |
| Residual | 79117.454 | 20 | 3955.873 | | |
| Total | 81427.039 | 21 | | | |
| NAFC | | | | | |
| Regression | 53164.177 | 1 | 53164.177 | 37.621 | .000 |
| Residual | 28262.862 | 20 | 1413.143 | | |
| Total | 81427.039 | 21 | | | |
| NADC | | | | | |
| Regression | 38484.798 | 1 | 38484.798 | 17.924 | .000 |
| Residual | 42942.241 | 20 | 2147.112 | | |
| Total | 81427.039 | 21 | | | |
| NABC | | | | | |
| Regression | 41061.953 | 1 | 41061.953 | 20.345 | .000 |
| Residual | 40365.085 | 20 | 2018.254 | | |
| Total | 81427.039 | 21 | | | |
| NA | | | | | |
| Regression | 70059.505 | 1 | 70059.505 | 123.262 | .000 |
| Residual | 11367.533 | 20 | 568.377 | | |
| Total | 81427.039 | 21 | | | |
| NH | | | | | |
| Regression | 46072.141 | 1 | 46072.141 | 26.063 | .000 |
| Residual | 35354.898 | 20 | 1767.745 | | |
| Total | 81427.039 | 21 | | | |
| DHP | | | | | |
| Regression | 7672.750 | 1 | 7672.750 | 2.081 | .165 |
| Residual | 73754.289 | 20 | 3687.714 | | |
| Total | 81427.039 | 21 | | | |
| RSA | | | | | |
| Regression | 269.362 | 1 | 269.362 | .066 | .799 |
| Residual | 81157.676 | 20 | 4057.884 | | |
| Total | 81427.039 | 21 | | | |
| NRFD | | | | | |
| Regression | 34119.201 | 1 | 34119.201 | 14.424 | .001 |
| Residual | 47307.838 | 20 | 2365.392 | | |
| Total | 81427.039 | 21 | | | |
| NFC | | | | | |
| Regression | 47494.866 | 1 | 47494.866 | 27.994 | .000 |
| Residual | 33932.173 | 20 | 1696.609 | | |
| Total | 81427.039 | 21 | | | |

ANOVA (analysis of variance) and comparison of F ratio value was used to test the stated hypothesis. F ratio value can be obtained from F ratio table by proceeding along x columns right and y rows down. The point of intersection of x and y coordinates in F ratio table is the critical F value at level of significance 0.05. If the experimental value of F is greater than table value of F, then the results are significant at that level of probability. The value of $F_{1,20}$ is 4.35 at $\alpha = 0.05$ in the F ratio table. The results of univariate regression analysis are given in Table 3.5.

The results of Table 3.5 show that there exist a significant relationship between metrics NC, NAFC, NADC, NABC, NA, NH, NRFD, NFC and understanding time as $F_{1,20}$ (experimental) > $F_{1,20}$ (tabulated) at $\alpha = 0.05$. The metric NA has the highest F value of 123.2 showing its greatest impact on understanding time. The metrics that do not significantly affect understanding time are bold, as can be seen from Table 3.5. These metrics are NDC, NBC, RBC, DHP and RSA. It can be seen that the results of univariate linear regression are similar to the results of correlation analysis. Thus alternate hypothesis H_{01} is a valid hypothesis as quality metrics significantly affect the understanding time of models.

The results of model summary for univariate linear regression are presented in Table 3.6.

Table 3.6 Model Summary of Univariate Linear Regression

| Model | R | R Square |
|-------|------|-------------|
| NDC | .323 | .104 |
| NBC | .186 | .035 |
| NC | .474 | .224 |
| RBC | .168 | .028 |
| NAFC | .808 | .653 |
| NADC | .687 | .473 |
| NABC | .710 | .504 |
| NA | .928 | .860 |
| NH | .752 | .566 |
| DHP | .307 | .094 |
| RSA | .058 | .003 |
| NRFD | .647 | .419 |
| NFC | .764 | .583 |

Table 3.6 shows the values of R and R² for each of the individual metrics in relation with understanding time. R is sample coefficient and its values are same as given by correlation coefficient in Table 3.4. R² is the coefficient of determination and provides a measure of how well a regression line fits the data points. A unit value of R² shows that regression line perfectly fits data points. It can be seen from Table 3.6 that NA accounts for 86.0% variance followed by NAFC with 65.3% variance. A non-significant variance (less than 10 percent) is shown by metrics NDC, NBC, RBC, DHP and RSA. The higher the value of R² for a metric, greater is its contribution towards predicting understanding time.

Table 3.7 Multiple Regression

| Model | Sum of Squares | Degree of freedom | Mean Square | F | Sig. |
|---|----------------|-------------------|-------------|---------|-------------------|
| NA Regression | 69911.753 | 1 | 69911.753 | 123.306 | .000 |
| Residual | 11339.520 | 20 | 566.976 | | |
| Total | 81251.273 | 21 | | | |
| NA, NAFC Regression | 70701.194 | 2 | 35350.597 | 63.664 | .000 |
| Residual | 10550.079 | 19 | 555.267 | | |
| Total | 81251.273 | 21 | | | |
| NA, NAFC, NFC Regression | 71220.942 | 3 | 23740.314 | 42.603 | .000 ^a |
| Residual | 10030.331 | 18 | 557.241 | | |
| Total | 81251.273 | 21 | | | |
| NA, NAFC, NFC, NH Regression | 73044.416 | 4 | 18261.104 | 37.827 | .000 ^a |
| Residual | 8206.856 | 17 | 482.756 | | |
| Total | 81251.273 | 21 | | | |
| NA, NAFC, NFC, NH, NABC Regression | 74401.300 | 5 | 14880.260 | 34.757 | .000 ^a |
| Residual | 6849.973 | 16 | 428.123 | | |
| Total | 81251.273 | 21 | | | |
| NA, NAFC, NFC, NH, NABC, NADC Regression | 74401.300 | 6 | 14880.260 | 34.757 | .000 ^a |
| Residual | 6849.973 | 15 | 428.123 | | |
| Total | 81251.273 | 21 | | | |
| NA, NAFC, NFC, NH, NABC, NRFD Regression | 74446.393 | 7 | 12407.732 | 27.350 | .000 ^a |
| Residual | 6804.880 | 14 | 453.659 | | |
| Total | 81251.273 | 21 | | | |
| NA, NAFC, NFC, NH, NABC, NRFD, NC Regression | 75937.759 | 8 | 10848.251 | 28.583 | .000 ^a |
| Residual | 5313.514 | 13 | 379.537 | | |
| Total | 81251.273 | 21 | | | |
| NA, NAFC, NFC, NH, NABC, NRFD, NC, NDC Regression | 80114.668 | 9 | 10014.333 | 114.540 | .000 ^a |
| Residual | 1136.605 | 12 | 87.431 | | |
| Total | 81251.273 | 21 | | | |

Also multiple regression was performed to find the effect of various combinations of independent variables (metrics) on understandability. Table 3.7 shows the multiple regression between various combinations of the metrics and understandability.

If F (Obtained) is greater than F (Tabulated) at significance level 0.05 then the alternate hypothesis H_{01} is a valid hypothesis. Various F values are $F_{1,20}$ (Tabulated) = 4.35, $F_{2,19}$ (Tabulated) = 3.52, $F_{3,18}$ (Tabulated) = 3.15, $F_{4,17}$ (Tabulated) = 2.96, $F_{5,16}$ (Tabulated) = 2.85, $F_{6,15}$ (Tabulated) = 2.79, $F_{7,14}$ (Tabulated) = 2.76, $F_{8,13}$ (Tabulated) = 2.76, $F_{9,12}$ (Tabulated) = 2.79, $F_{10,11}$ (Tabulated) = 2.85, $F_{11,10}$ (Tabulated) = 2.91, $F_{12,9}$ (Tabulated) = 3.00, $F_{13,8}$ (Tabulated) = 3.26.

According to the descending F value, combination of metrics is shown in Table 3.7. In this table, F value of all the combination of the metrics is greater than F (tabulated). The results of Table 3.6 and Table 3.7 further strongly supports the results of Table 3.5 and thus prove the validity of alternate hypothesis H_{01} .

Table 3.8 Model Summary of Understandability

| Model | R | R ² | Adj R ² |
|--|------|----------------|--------------------|
| NA | .928 | .860 | .853 |
| NA,NAFC | .933 | .870 | .856 |
| NA,NAFC, NFC | .936 | .877 | .856 |
| NA,NAFC, NFC, NH | .948 | .899 | .875 |
| NA,NAFC, NFC, NH, NABC | .957 | .916 | .889 |
| NA,NAFC, NFC, NH, NABC, NRFD | .956 | .916 | .883 |
| NA,NAFC, NFC, NH, NABC, NRFD, NC | .967 | .935 | .902 |
| NA,NAFC, NFC, NH, NABC, NRFD, NC, NDC | .993 | .986 | .977 |
| NA,NAFC, NFC, NH, NABC, NRFD, NC, NDC, DHP | .993 | .987 | .977 |
| NA,NAFC, NFC, NH, NABC, NRFD, NC, NDC, DHP, RBC | .994 | .988 | .976 |
| NA,NAFC, NFC, NH, NABC, NRFD, NC, NDC, DHP, RBC, RSA | .994 | .988 | .974 |

For all the combinations of the metrics shown in Table 3.7, value of R , R^2 and adjusted R^2 are calculated and shown in Table 3.8. Here, R gives linear regression coefficient. R^2 provides information about the model fitness. If $R^2 = 1.0$, it says regression line correctly fits the real data. Adjusted R^2 indicates the adjustment

of R^2 when a new variable is added to the model. Increase in adjusted R^2 indicates the improvement in the model on adding a new variable. Always $\text{adj } R^2$ is less than or equal to the value of R^2 . Model summary shows the change in variance on combining the various independent variables. The results in Table 3.8 show a variance of 98.8 in the dependent variable (understandability) on the final combination of the independent variables which indicate a good model.

3.7.3 Principal Component Analysis (PCA)

PCA was applied to the collected data to find the principal components that are important in the sense that they account for explaining the model without losing any significant information. The threshold value for considering a component as principal is assumed to be 1. The inputs to the PCA technique are the tables of metrics for all the 22 models. The results obtained after application of PCA are given in Table 3.9.

Table 3.9 Results of PCA

| Component | Initial Eigenvalues | | |
|-----------|---------------------|---------------|---------------|
| | Total | % of Variance | Cumulative % |
| 1 | 5.382 | 41.403 | 41.403 |
| 2 | 4.388 | 33.751 | 75.154 |
| 3 | 1.555 | 11.958 | 87.112 |
| 4 | 1.105 | 8.497 | 95.609 |
| 5 | .338 | 2.596 | 98.205 |
| 6 | .111 | .858 | 99.063 |
| 7 | .050 | .386 | 99.449 |
| 8 | .032 | .243 | 99.692 |
| 9 | .019 | .148 | 99.840 |
| 10 | .012 | .093 | 99.933 |
| 11 | .009 | .067 | 100.000 |
| 12 | 2.854E-018 | 2.195E-017 | 100.000 |
| 13 | -3.566E-017 | -2.743E-016 | 100.000 |

Table 3.9 shows that the first 4 components (highlighted as bold) account for a variance of 95.6%. This means that the identified 4 components are capable enough to explain 95.6% of the model summary and the rest of the components contribute 4.4% towards the explanation of the model summary. These facts are further supported by the scree plot in Figure 3.7.

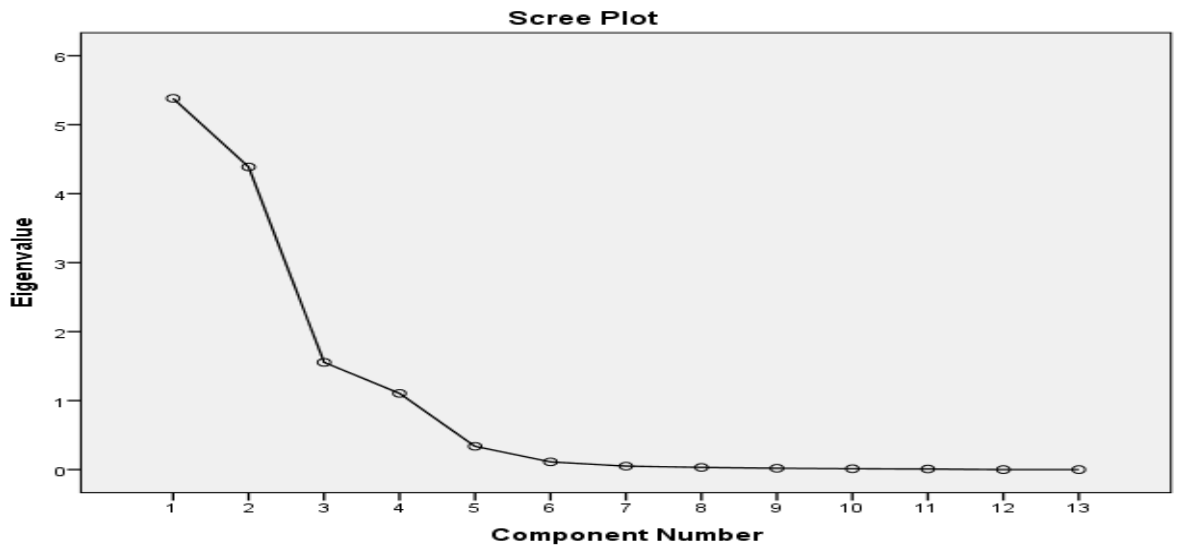


Figure 3.7 Scree plot of PCA

Figure 3.7 shows that first four components have eigen values greater than threshold value of 1 and these components account for 95.6% variance of data. Table 3.10 presents a rotated component matrix that identifies the principal components of the data.

Table 3.10 Rotated Component Matrix

| | Component | | | |
|-------------|-------------|-------------|-------------|-------------|
| | 1 | 2 | 3 | 4 |
| NDC | | | .972 | |
| NBC | .955 | | | |
| NC | .666 | | .712 | |
| RBC | .810 | -.437 | | |
| NAFC | | .790 | | .556 |
| NADC | | .883 | | |
| NABC | .880 | .389 | | |
| NA | | .910 | | |
| NH | .844 | .433 | | |
| DHP | .973 | | | |
| RSA | | | | .982 |
| NRFD | | .526 | .842 | |
| NFC | | .931 | | |

To identify the principal components, the metric with highest value in each column 1, 2, 3 and 4 is selected as highlighted in Table 3.10. The identified components are DHP, NFC, NDC and RSA. These selected components are the principal components in explanation of model summary. Looking at the results of correlation analysis, regression analysis and PCA, we find that one metric NFC is having a significant role

in predicting the understanding time and it is also the principal component of the models. The results of PCA further show that it is not necessary that all the principle components have a significant effect on understandability of model. Thus alternate hypothesis H_{02} is valid hypothesis and hypothesis H_{02} is rejected.

3.7.4 Nearest Neighbour Analysis

This technique was used to identify models with similar structures. After identification, the understanding time of similar models were compared to check the validity of alternate hypothesis H_{03} . The inputs were the values of metrics for all the 22 models along with their understanding times. A total of 13 predictors (metrics) were used to calculate the nearest neighbour of each model. The model pairs having minimum Euclidean distance between them are nearest to each other. The Euclidean distance is calculated between metric values of models. A model is assumed as a point in n-dimensional space and coordinates corresponding to metrics m_j . The distance between pair of metrics is calculated as:

$$Dis(s, \hat{s}) = \sum_{j=1}^n \beta_j dis(m_j(s_i), m_j(\hat{s}_i))$$

Where β_j = weight of the metric

$dis(m_j(s_i), m_j(\hat{s}_i))$ = dissimilarity with respect to metric m_j .

$dis(m_j(s_i), m_j(\hat{s}_i)) = |m_j(s) - m_j(\hat{s})| / dom(m_j)$

Where $dom(m_j)$ = maximal difference of two values in domain m_j .

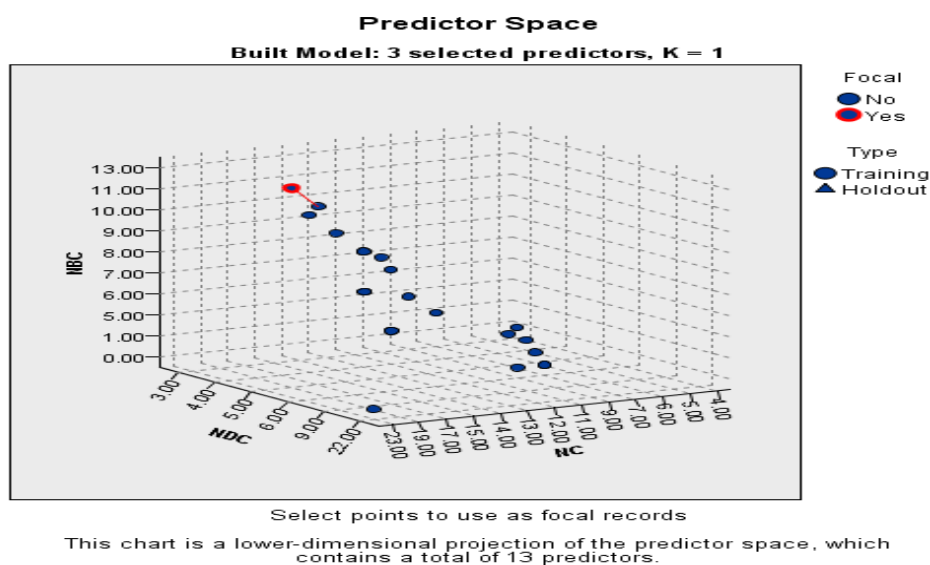


Figure 3.8 Predictor space for selected model 2

Figure 3.8 shows the snapshot results of predictor space for selected model 2 (shown by red color). The red line connects the selected model to its nearest neighbour.

For a more detailed analysis, in terms of individual metrics peer charts were analysed. The snapshot of peer chart, as shown in Figure 3.9, shows the metric values of the nearest neighbour for selected model 2. It can be seen from peer chart that the model for which the nearest neighbour is to be calculated is shown in red along with model number which is model 2. The nearest neighbour is shown in blue along with its number which is model 3. Each rectangle in the peer chart show the value of individual metric for selected model and its nearest neighbour. It can be seen from peer chart that the value of NDC is 5 for model 2 and model 3. Similar interpretations hold for other metrics in peer chart.

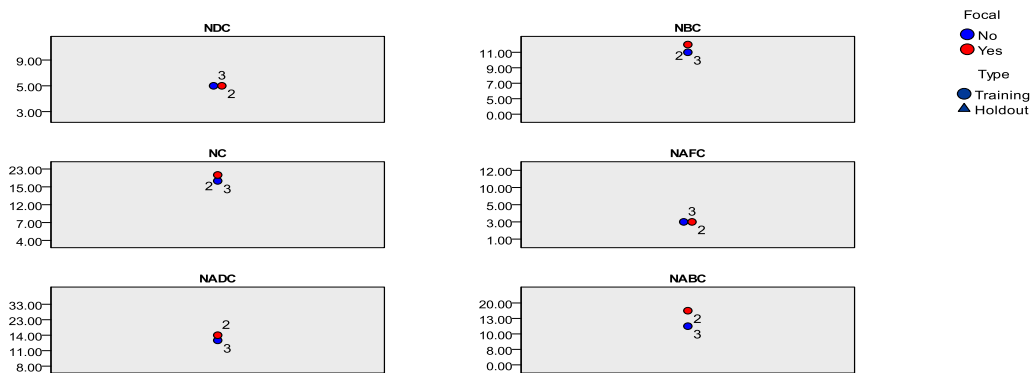


Figure 3.9 Peer chart for selected model 2

The nearest neighbours of each model are given in Table 3.11. The first column of Table 3.11 presents the model number, the second column shows the model that is nearest to model in first column, the third column shows the average understanding time for model in first column, the fourth column shows the average understanding time for model in second column and fifth column shows the time interval in which the understanding times fall. Time interval in fifth column of Table 3.11 is categorized as follows:

Category T[1] with understanding time less than average understanding time (130 seconds)

Category T[2] with understanding time more than average understanding time (130 seconds)

Time 130 seconds is the average understanding time of all the models.

The time interval column of Table 3.11 has entry T[1] if the understanding times of that particular row are below 130 seconds and T[2] if the understanding times of that particular row are above 130 seconds. It can be seen from Table 3.11 that all the nearest neighbours fall in same understanding time category T[1] or T[2] except for model 21 (with model 19). This exception can be explained for large distance (3.347) between S21 and S19. Analysing the results of nearest neighbour analysis, it was found that the models having similar values of quality metrics have significant relation in respect of their understanding times or similar structures leads to similar understandability. Thus alternate hypothesis H_{03} is valid hypothesis.

Table 3.11 Nearest Neighbour Analysis

| Schema | Nearest Neighbor (NN) | Understanding Time Schema | Understanding Time NN | Time Interval |
|--------|-----------------------|---------------------------|-----------------------|---------------|
| S01 | S05 | 103 | 94 | T [1] |
| S02 | S03 | 121 | 106 | T [1] |
| S03 | S02 | 106 | 121 | T [1] |
| S04 | S07 | 96 | 88 | T [1] |
| S05 | S08 | 94 | 98 | T [1] |
| S06 | S08 | 96 | 98 | T [1] |
| S07 | S08 | 88 | 98 | T [1] |
| S08 | S05 | 98 | 94 | T [1] |
| S09 | S10 | 95 | 99 | T [1] |
| S10 | S05 | 99 | 94 | T [1] |
| S11 | S12 | 209 | 272 | T [2] |
| S12 | S11 | 272 | 272 | T [2] |
| S13 | S14 | 224 | 295 | T [2] |
| S14 | S13 | 295 | 224 | T [2] |
| S15 | S22 | 97 | 96 | T [1] |
| S16 | S20 | 104 | 82 | T [1] |
| S17 | S22 | 126 | 96 | T [1] |
| S18 | S19 | 91 | 100 | T [1] |
| S19 | S18 | 100 | 91 | T [1] |
| S20 | S16 | 82 | 104 | T [1] |
| S21 | S19 | 164 | 100 | exception |
| S22 | S15 | 96 | 97 | T [1] |

3.7.5 ROC Classification

ROC classification was used to find the individual contributions of metrics on understanding time. Receiver Operating Characteristic (ROC) curves are helpful in interpreting sensitivity and specificity levels and determination of related cut scores.

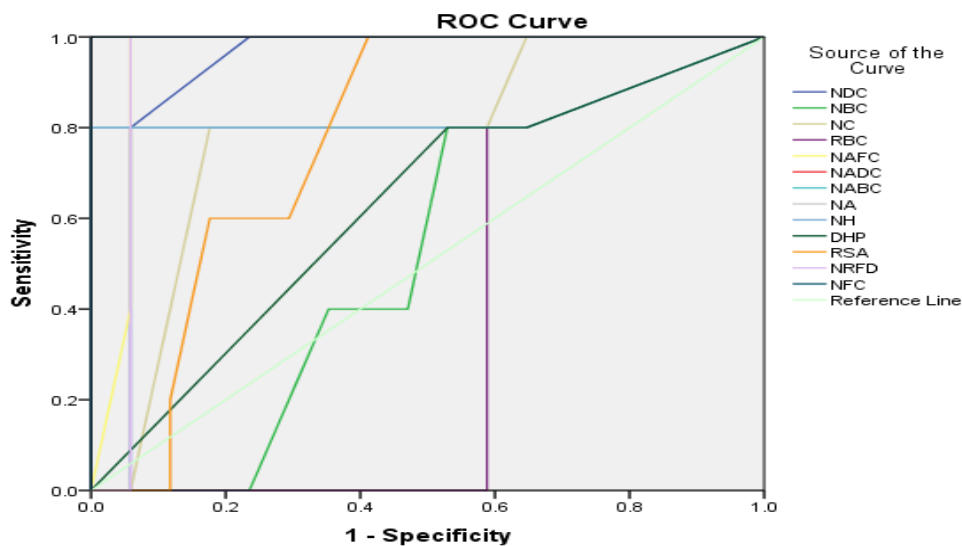
ROC analyses provide a common scale for comparing different predictors that are measured in different units.

The dependent variable (understanding time) was categorized into two categories.

Category T[1] with understanding time less than average understanding time (130 seconds) [Understandable models]

Category T[2] with understanding time more than average understanding time (130 seconds) [Non-understandable models]

Understanding time category was used as state variable with value 1 corresponding to T[1] and 0 corresponding to T[2]. The ROC curve plot with sensitivity on y-axis and 1-specificity on x-axis is shown in Figure 3.10. Sensitivity and specificity both measure the correctness of predicted models/models. Sensitivity is defined as the number of models correctly predicted as understandable to total number of actual understandable models. Specificity is defined as the number of models correctly predicted as non-understandable to total number of actual non- understandable models.



Diagonal segments are produced by ties.

Figure 3.10 ROC Curve plot

The summary analysis of ROC curve plot of Figure 3.10 is shown in Table 3.12.

Table 3.12 Summary results of ROC

| Test Result Variable(s) | Area |
|-------------------------|--------------|
| NDC | .924 |
| NBC | .518 |
| NC | .782 |
| RBC | .365 |
| NAFC | .953 |
| NADC | 1.000 |
| NABC | .835 |
| NA | 1.000 |
| NH | .835 |
| DHP | .624 |
| RSA | .776 |
| NRFD | .941 |
| NFC | 1.000 |

The measure of interest in Table 3.12 is Area. The area under ROC curve gives a measure of accuracy of predicted model. Accuracy is defined as the number of models correctly predicted as understandable to the number of models predicted as non-understandable. As can be seen from Table 3.12, the highlighted metrics NADC, NA and NFC occupy highest area under ROC curve plot. The contribution of these metrics is significant in predicting the understandability of models having understanding time less than or equal to 130 seconds. Similarly, RBC has least contribution in predicting the understandability of given set of models. The results of analysis of ROC classification support the validity of alternate hypothesis H_{01} .

3.8 THREATS TO VALIDITY AND LIMITATIONS

The following section presents the threats to construct, internal, external and conclusion validity that might affect experimental results and challenge the generalization of results. Following threats were identified during the course of experiment.

- Construct validity [11] takes into consideration relationship between theoretical concepts and actual experimental observation. An assumption was made that it was not the domain of set of models but the structural complexity of models that caused variation in analyzing and answering the questions. After the experiment, a interaction with the volunteers was made to know whether the questions designed for each of the models were capable enough in measuring the understandability based on structural complexities of models.

The majority of participants responded satisfactorily in favor of designed questions. More experiments with models from different domains along with varying answers can help reduce threats to construct validity.

- Internal validity[11] explains the cause-effect relation between independent and dependent variables. It measures the causal effect of independent variables on dependent variables. The possible, identified threats to internal validity are discussed as follows:
 - *Differences among subjects*: Experiments conducted from within the subjects reduce variability. All the students participating in the experiment were of same age group, experience and stream.
 - *Differences among models*: Models from different domains can affect the results of experiments. The models used in the experiment were from generalized domains so that the subjects did not find any difficulty with domain understandability.
 - *Time recorded for completing the tasks*: The subjects recorded the starting and ending time of tasks. It is understandable that subjects could introduce impression while recording the time for completion of tasks. A supervisor was constantly monitoring the student activities so that actual facts could be recorded for analysis.
 - *Learning effect*: Assignment of tasks to subjects followed a variable order to reduce learning effects.
 - *Fatigue effects*: The average time for task completion was around 130 seconds. This much time hardly introduces fatigue effects. The variable order of tasks further minimized this effect.
 - *Subject motivation*: The subjects of the experiments were volunteers and this experiment was apart from their course curriculum. The subjects were motivated enough to participate in the experiment.
 - *Subjects influence*: During the complete duration of experiment a supervisor monitored the subjects so that they do not talk with or influence each other.
- External validity[11] specifies the extent to which the experimental results can be generalized. The possible threats to external validity are discussed as follows:

- *Models and questionnaire used:* Domains of all models were familiar and known to subjects to avoid any problems with understandability of domain. Each of the questions for a model was based on different level of understandability and therefore time taken to answer each question was different. The selected models had different structural complexities, so the time taken to answer questions varied from one model to another. More number of experiments with complex models needs to be conducted.
- *Nature of Subjects:* The subjects of our experiment were students who were having adequate knowledge of tasks to be performed. More experiments with industrial subjects could be carried out.
- Conclusion validity [11] explains the extent to which the results of experiment are statistically valid. The factor limiting the results of conclusion validity is sample size (22) and subject size (80). Experimentation with big sample data might give more positive results.

3.9 SUMMARY

In this chapter, a new quality metric is proposed. The metric is theoretically and empirically validated, as suggested by many researchers, along with existing metrics to prove its validity towards quality evaluation of conceptual models. The next step in the research study is to rank the quality metrics in accordance to their significance in evaluation of quality of conceptual data warehouse models. In the next chapter, research work related to ranking of conceptual data warehouse models is performed and presented towards building of an efficient information delivery system.

CHAPTER IV

QUALITY EVALUATION BASED ON RANKING, INFERENCE APPROACH TOWARDS BUILDING OF EIDS

4.1 INTRODUCTION

In the previous chapters, research was focused on the role played by quality metrics towards prediction of understandability of data warehouse conceptual models. There exist several other criteria such as efficiency and effectiveness, in addition to understandability, along which quality of conceptual models can be evaluated using quality metrics. The criteria are defined qualitatively and the significance of quality metrics along the criteria varies according to user requirements, situations and expert opinion. To measure the quality of conceptual metrics along multiple criteria, the need for ranking of quality metrics was felt. The rank of metrics can be one of the major considerations during design of conceptual data warehouse models. From the study of literature, the need of a systematic ranking approach that considers uncertainties, ambiguities, biases involved in human thought process and takes into account all possible interdependencies of attributes involved was identified. A fuzzy based ranking system was evolved to deal with imprecise and qualitative (non-numeric) data based on actual human (expert) decision making. The successive sections discuss the research work carried towards ranking of quality metrics along multiple criteria using fuzzy methodology.

4.3 PRELIMINARIES

Before discussing the precise methodology based on fuzzy logic and matrix operations to rank quality metrics of conceptual data warehouse models some basics needs to be presented. The basic of fuzzy logic, linguistic variables and matrix functions are present subsequently.

4.2.1 Introduction to Fuzzy Sets

The basic dictionary meaning of fuzzy is blurred, indistinct. The concept of fuzziness shows uncertainty, imprecision, ambiguity, inconsistency, vagueness of situations.

Zadeh [21] introduced a theory whose objects are the sets with no precise boundaries. The concept of fuzziness was developed to solve the problems in which description of observations was imprecise, ambiguous, uncertain or vague. A fuzzy set is a class of objects, where each object is associated with membership grades varying from 0-1. The fuzzy sets show gradual transition from membership to non-membership and vice-versa. The range of membership functions is the unit interval [0, 1]. The membership function of a fuzzy set A is denoted by μ_A ,

$$\mu_A : X \rightarrow [0,1], \text{ where } X \text{ is a universal set [110]}$$

Degree of membership is 0 when the element is not in set, degree of membership is 1 when the element is in the set. A value between 0-1 shows the ambiguity of membership. Various operations like union, intersection, addition, subtraction, multiplication, division can be applied to fuzzy sets.

4.2.2 Triangular Fuzzy Membership Functions

Zadeh [21] defined several fuzzy membership functions like Γ - functions (increasing membership functions with straight lines), L- functions (decreasing membership functions with straight lines), S- functions, Bell shaped functions, \wedge - functions (triangular functions). The use and application of the membership functions depends on the scenario to which it is applied. The most commonly used is the triangular function due to its ease of use and calculations. The triangular fuzzy membership function [110], denoted as $\wedge: X \rightarrow [0, 1]$ is defined as follows:

$$\wedge(x;a,b,c) = \begin{cases} (x - a)/(b - a), & a \leq x \leq b \\ (c - x)/(c - b), & b \leq x \leq c \\ 0, & \text{otherwise} \end{cases}$$

Where a, b, c are real numbers $a \leq b \leq c$. The triangular fuzzy function can be represented graphically as shown in Figure 4.1

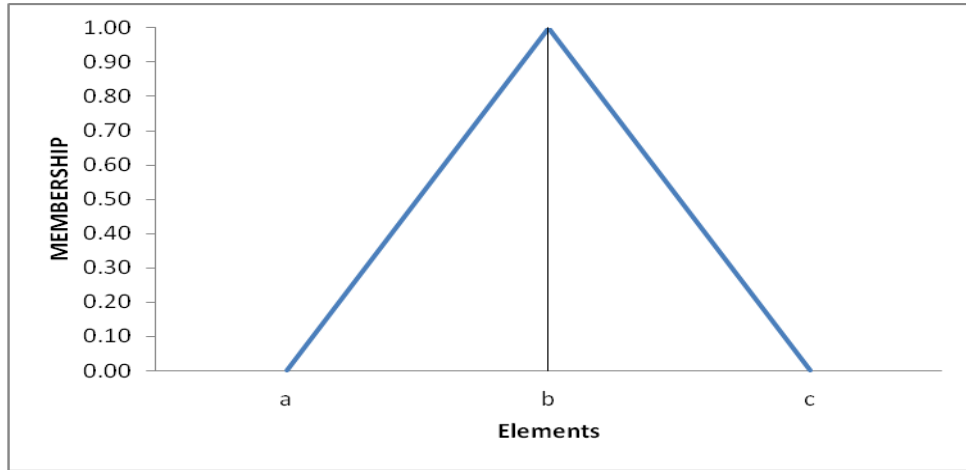


Figure 4.1 Triangular fuzzy membership function graph

Addition and multiplication operations on fuzzy numbers have been made use of in the research study. Given two fuzzy numbers defined in terms of triangular membership functions as $A_1 = (a_1, b_1, c_1)$ and $A_2 = (a_2, b_2, c_2)$. The addition and multiplication operation can be expressed as:

Addition: Let \oplus denotes addition

$$A_1 \oplus A_2 = (a_1, b_1, c_1) \oplus (a_2, b_2, c_2) = (a_1 + a_2, b_1 + b_2, c_1 + c_2)$$

Multiplication: Let \otimes denotes multiplication

$$A_1 \otimes A_2 = (a_1, b_1, c_1) \otimes (a_2, b_2, c_2) = (a_1 \times a_2, b_1 \times b_2, c_1 \times c_2)$$

4.2.3 Fuzzy Linguistic Terms and Variables

Fuzzy logic theory involves the uncertainty and ambiguity in human thought process and quantify it in terms of lingual terms. The natural linguistic terms used in common usage are closer to human perceptions and thoughts than crisp numeric values. A linguistic term is a variable whose values are not numbers but words or sentences used in natural language. A linguistic variable is some non-numeric syllable/term used in natural usage. Various linguistic variables to weight the criteria and rate the metrics have been made use of in the research study.

The weights assigned to specified criteria are evaluated in terms of linguistic variables High(H), Medium(M), Low(L). The membership values for each of the linguistic variables is expressed as $H(0.5, 0.8, 1)$, $M(0.3, 0.5, 0.8)$, $L(0, 0.3, 0.5)$ as shown in Table 4.1 and the corresponding membership graph is shown in Figure 4.2.

Table 4.1 Fuzzy membership values for weights assigned to criteria

| Linguistic Variable | High(H) | Medium(M) | Low(L) |
|---------------------|--------------|---------------|-------------|
| Fuzzy membership | (0.5,0.8, 1) | (0.3,0.5,0.8) | (0,0.3,0.5) |

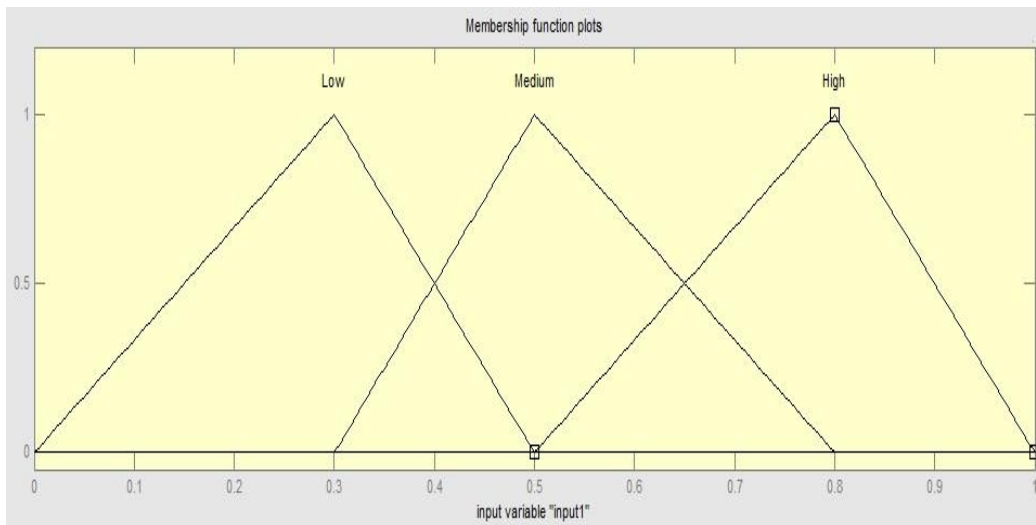


Figure 4.2 Fuzzy membership graph for weighting criteria

Similarly, the ratings assigned to quality metrics are expressed in terms of linguistic variables Very Good(VG), Good(G), Medium(M), Poor(P), Very Poor(VP). The triangular fuzzy membership values are assigned to the variables as shown in Table 4.2 and the corresponding membership graph is shown by Figure 4.3.

Table 4.2 Fuzzy membership values for rating assigned to quality metrics

| Linguistic Variable | Very Good(VG) | Good(G) | Medium(M) | Poor(P) | Very Poor(VP) |
|---------------------|---------------|-------------|---------------|-------------|---------------|
| Fuzzy membership | (0.7,1,1) | (0.5,0.7,1) | (0.2,0.5,0.7) | (0,0.3,0.5) | (0,0,0.3) |

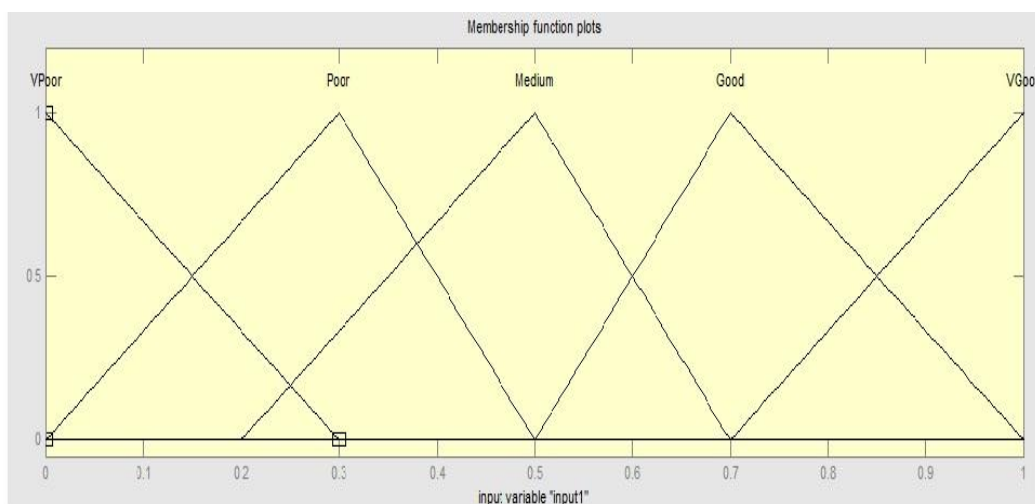


Figure 4.3 Fuzzy membership graph for rating quality metrics

4.2.4 Quality Metrics Ranking Problem and Fuzzy Solution

The section defines the problem for ranking quality metrics of conceptual data warehouse model using fuzzy based approach and multi-criteria analysis.

The quality metrics ranking problem [20] and its multi criteria fuzzy solution can be defined as:

A team of n experts ($E_1, E_2, E_3, \dots, E_n$), has to analyse and grant weights to k criteria ($C_1, C_2, C_3, \dots, C_k$) and the ratings to m quality metrics ($Q_1, Q_2, Q_3, \dots, Q_m$) for each of the k criteria. Let W_{ij} ($i=1,2,3,\dots,k; j=1,2,3,\dots,n$) be the weight assigned to criteria C_i by expert E_j . Let R_{ijt} ($i=1,2,3,\dots,m; j=1,2,3,\dots,n; t=1,2,3,\dots,k$) be the rating given to metric Q_i by expert E_j under criteria C_t .

$$W_i = 1/n \otimes (W_{i1} \oplus W_{i2} \oplus \dots \oplus W_{in})$$

$$R_{ij} = 1/n \otimes (R_{ij1} \oplus R_{ij2} \oplus \dots \oplus R_{ijn})$$

Where W_i is the average weight of criteria and R_{ij} is the aggregated rating of quality metric Q_i under criteria C_j . Mean has been used to aggregate the opinions of expert, as it is most commonly and widely used operator in common practice. Defuzzification [97] (conversion of fuzzy aggregations to crisp scores) have been carried out using area of centroid method due to ease of application and usage.

4.2.5 Criteria Matrix

Each of the quality metrics has multiple rating scores corresponding to expert evaluation along several criteria. The multiple scores for each metric needs to be converted into a single index score to rank the metrics based on their relative significance and impact towards quality evaluation of conceptual data warehouse models. The crisp scores for each metric are achieved using Criteria matrix. A criteria matrix [20] is aggregation of metric rating along multiple criteria and aggregated relative weights of each criteria. The order of criteria matrix is $n \times n$, where n is the number of criteria for metric evaluation. The diagonal elements of criteria matrix show the aggregated rating of a metric along multiple criteria and the off diagonal elements represent the relative aggregated weights of multiple criteria. Thus, a criteria matrix is a combination of two matrix. One is metric rating matrix and other is relative weight matrix.

- Metric rating matrix: This is a diagonal matrix is a $n \times n$ matrix, whose elements are the aggregated rankings of a metric evaluated along multiple criteria.

$$\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & a_{nn} \end{bmatrix}$$

- Relative weight matrix: This is a $n \times n$ matrix, whose diagonal elements are all 0's, and whose off diagonal elements gives the aggregated relative weights of criteria. In mathematical terms, an element a_{ij} of the relative weight matrix equals weight of criteria j divided by weight of criteria i .

$$a_{ij} = \frac{\text{weight of criteria } j}{\text{weight of criteria } i}$$

$$\begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ a_{21} & 0 & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix}$$

Thus the criteria matrix which is a combination of metric rating matrix and relative weight matrix is as follows:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

4.2.6 Permanent of Matrix

Permanent [98] of a matrix is an important technique for ranking of systems based on multi-criteria evaluation. The permanent is similar to determinant with the only difference that no negative term appears in calculation of permanent. In mathematical terms, permanent [98] is given as:

For a square matrix M (order n) = $[m_{ij}]_{1 \leq i, j \leq n}$

$$\text{perm}(M) = \sum_{\pi \in S} \prod_{i \in I} M_{i\pi(i)}$$

Where S consists of the group of symmetric elements S_n .

4.2.7 Expert Opinion Ranking Methodology

The results of proposed fuzzy methodology have been compared with aggregations of expert opinion [87]. The input given for calculation of expert opinion is algebraic aggregation of linguistic membership functional data collected from experts. Ranking problem and its expert opinion solution can be stated as:

A team of m experts ($E_1, E_2, E_3, \dots, E_m$), has to analyse and grant weights to l criteria ($C_1, C_2, C_3, \dots, C_l$) and the ratings to n quality metrics ($Q_1, Q_2, Q_3, \dots, Q_n$) for each of the l criteria. Let $r(i,j,k)$ be the rating given to metric Q_i by expert E_k under criteria C_j and $w(j)$ be the weight evaluated by experts for criteria C_j . The ratings and weights given as input are the mean algebraic aggregation of linguistic membership functional data collected from experts. The aggregated rating $R(i)$ for metric Q_i is calculated using the aggregation mean value function [87] defined as follows:

$$R(i) = \frac{1}{ml} * \sum_{j=1}^l \sum_{k=1}^m r(i, j, k) * w(j)$$

4.3 RESEARCH METHODOLOGY

The research methodology followed during the current research is shown in Figure 4.4. The stepwise detail of research methodology shown in Figure 4.4 is given in the following sub-sections.

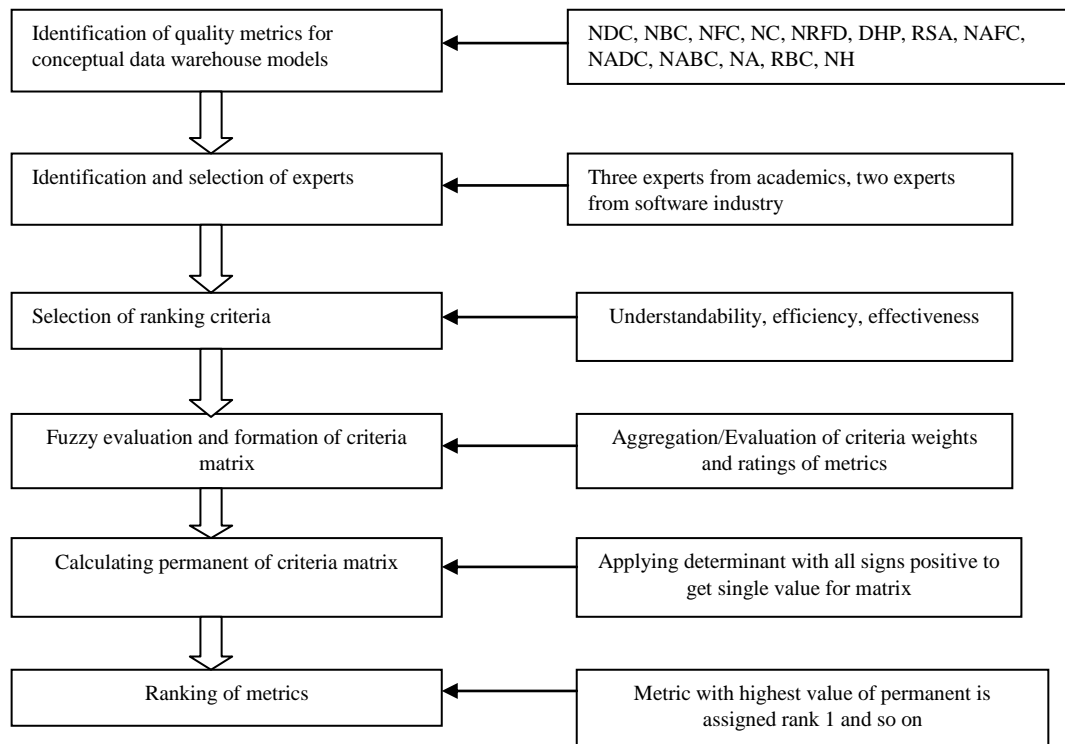


Figure 4.4 Research methodology

4.3.1 Identification of Quality Metrics for Conceptual Data Warehouse Models

The quality of conceptual data warehouse [111, 112, 113] can be predicted using quality metrics based on size and structural complexity of models and the same has been discussed in previous chapter. A total of 13 quality metrics have been identified for ranking including the metrics proposed by Serrano et al [16] and one proposed. The metrics have already been discussed in detail chapter III and shown in Figure 4.4.

4.3.2 Identification and Selection of Experts

One common data collection technique is to prepare questionnaires and conduct a survey based on questionnaire. Various statistical techniques can be applied to data collected. Due to blind nature of statistics, the results may vary from one survey to another and cannot be generalized. In the questionnaire survey various possible threats to validity exist like fatigue effects, biased results, motivation effects, learning effects that cannot be avoided. So expert's opinion was identified as the best feasible approach for data collection. Certain factors [86] were kept in mind for expert selection which are as follows:

- Experts should have wide publications, good practical hands on experience and should be capable enough to handle or address diverse research issues related to domain under consideration.
- Experts should have vast experience in the related issues in college/universities, industries/consultancy firms and public sector/government agencies.
- Experts should be volunteer and willing to part of the methodology under study.

In this study five experts from data warehouse domain having up to date knowledge of technological advances and rich practical hands on experience, with more than 10-20 years of experience were selected and approached. Out of five experts three were from academics and two were from software industry. The academic experts were chosen as they have good experimental knowledge and are well acquainted with up to date technical advancements. The industrial experts have good insights into issues related to cost and benefits.

4.3.3 Selection of Ranking Criteria

The quality metrics can be evaluated in terms of several parameters termed as ranking criteria. The metrics have been evaluated along previously identified parameters as specified by Serrano et al [16]. The identified parameters [16] are as follows:

- *Understandability*: It is defined as the time taken to understand a conceptual models and perform tasks (answer questions) based on understanding of the conceptual models.
- *Efficiency*: It is defined as the number of correct tasks performed per unit time based on the understandability of conceptual models.
- *Effectiveness*: It is defined as the number of correct tasks performed per total number of tasks based on size and structural complexity of conceptual models.

Each of the identified experts was to fill requisite Performa for assigning weights and ranking of metrics based on their experience and opinion. In consultation with the experts, forms for linguistic variables and membership functions to weight the criteria/rate the metrics was prepared and filled by each of the experts. The Performa 1 is shown by Figure 4.5.

| | | |
|-------------------|----------------|--|
| Criteria | Expert Opinion | <p>How do you weight the criteria towards quality evaluation of conceptual data warehouse models in terms of linguistic variables $H(0.5,0.8,1)$, $M(0.3,0.5,0.8)$, $L(0,0.3,0.5)$.</p> <p>High(H), Medium(M), Low(L).</p> |
| Understandability | | |
| Efficiency | | |
| Effectiveness | | |

Figure 4.5 Performa 1

| | | | | |
|---------|-------------------|------------|---------------|---|
| Metrics | Understandability | Efficiency | Effectiveness | <p>How do you rank the metrics along criteria of understandability , efficiency, effectiveness towards quality evaluation of conceptual data warehouse models in terms of linguistic variables $VG(0.7,1,1)$, $G(0.5,0.7,1)$, $M(0.2,0.5,0.7)$, $P(0,0.3,0.5)$, $VP(0,0,0.3)$.</p> <p>Very Good (VG), Good (G), Medium (M), Poor (P), Very Poor (VP)</p> |
| NA | | | | |
| NADC | | | | |
| NRFD | | | | |
| NBC | | | | |
| NC | | | | |
| RBC | | | | |
| NH | | | | |
| NDC | | | | |
| NFC | | | | |
| DHP | | | | |
| NABC | | | | |
| NAFC | | | | |
| RSA | | | | |

Figure 4.6 Performa 2

Each expert was to fill the above Performa to assign weights to specified criteria in terms of linguistic variables High(H), Medium(M), Low(L). The membership values for each of the linguistic variables was agreed upon as $H(0.5,0.8, 1)$, $M(0.3,0.5,0.8)$, $L(0,0.3,0.5)$. Performa 2 assigning rating to metrics versus criteria required to be filled by each of the experts is shown in Figure 4.6.

Each expert was to assign linguistic variables Very Good (VG), Good (G), Medium (M), Poor (P), Very Poor (VP) to rate the metrics in relation to specific criteria. The membership values for each of the linguistic variables is $VG(0.7,1,1)$, $G(0.5,0.7,1)$, $M(0.2,0.5,0.7)$, $P(0,0.3,0.5)$, $VP(0,0,0.3)$.

4.3.4 Fuzzy Evaluation and Formation of Criteria Matrix

Fuzzy evaluation of expert's opinion follows an incremental stepwise approach. Firstly, experts evaluate weight of each identified criteria and give ratings to metrics versus criteria in terms of fuzzy linguistic variables. Then the weights and ratings are aggregated. The aggregations are then converted to crisp scores to form a criteria matrix for each of the metrics (combination of weights and ratings).

4.3.5 Calculating Permanent of Criteria Matrix

A permanent function (determinant with all signs positive) is calculated for each of the criteria matrix build up in the previous step. A permanent gives the single value for the entire criteria matrix.

4.3.6 Ranking of Metrics

The matrix with the highest value of permanent calculated in previous step is ranked to number 1 and subsequently to number 2, 3 and so on.

4.4 PRACTICAL APPLICATION

An example is presented to illustrate the application of fuzzy based methodology discussed in the above sections.

Table 4.3 Fuzzy membership values and linguistic representation for ranking

| Criteria | E1 | E2 | E3 | E4 | E5 |
|-------------------|------------------|-----------------|------------------|-----------------|------------------|
| Understandability | $H(0.5,0.8, 1)$ | $H(0.5,0.8, 1)$ | $H(0.5,0.8, 1)$ | $H(0.5,0.8, 1)$ | $H(0.5,0.8, 1)$ |
| Efficiency | $H(0.5,0.8, 1)$ | $H(0.5,0.8, 1)$ | $M(0.3,0.5,0.8)$ | $H(0.5,0.8, 1)$ | $M(0.3,0.5,0.8)$ |
| Effectiveness | $M(0.3,0.5,0.8)$ | $H(0.5,0.8, 1)$ | $M(0.3,0.5,0.8)$ | $L(0,0.3,0.5)$ | $L(0,0.3,0.5)$ |

Table 4.4 Fuzzy membership values and linguistic representation for quality metrics

| Metrics | | Understandability | Efficiency | Effectiveness |
|---------|----|-------------------|----------------|----------------|
| NA | E1 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | G(0.5,0.7, 1) |
| | E2 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | G(0.5,0.7, 1) |
| | E3 | G(0.5,0.7, 1) | G(0.5,0.7, 1) | M(0.2,0.5,0.7) |
| | E4 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | P(0,0.3,0.5) |
| | E5 | G(0.5,0.7, 1) | P(0,0.3,0.5) | P(0,0.3,0.5) |
| NADC | E1 | G(0.5,0.7, 1) | M(0.2,0.5,0.7) | P(0,0.3,0.5) |
| | E2 | G(0.5,0.7, 1) | P(0,0.3,0.5) | P(0,0.3,0.5) |
| | E3 | M(0.2,0.5,0.7) | P(0,0.3,0.5) | VP(0,0,0.3) |
| | E4 | P(0,0.3,0.5) | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) |
| | E5 | M(0.2,0.5,0.7) | P(0,0.3,0.5) | P(0,0.3,0.5) |
| NRFD | E1 | G(0.5,0.7, 1) | G(0.5,0.7, 1) | G(0.5,0.7, 1) |
| | E2 | VG(0.7,1,1) | G(0.5,0.7, 1) | M(0.2,0.5,0.7) |
| | E3 | G(0.5,0.7, 1) | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) |
| | E4 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | G(0.5,0.7, 1) |
| | E5 | G(0.5,0.7, 1) | G(0.5,0.7, 1) | M(0.2,0.5,0.7) |
| NBC | E1 | VG(0.7,1,1) | G(0.5,0.7, 1) | G(0.5,0.7, 1) |
| | E2 | VG(0.7,1,1) | VG(0.7,1,1) | VG(0.7,1,1) |
| | E3 | G(0.5,0.7, 1) | VG(0.7,1,1) | VG(0.7,1,1) |
| | E4 | VG(0.7,1,1) | G(0.5,0.7, 1) | G(0.5,0.7, 1) |
| | E5 | VG(0.7,1,1) | VG(0.7,1,1) | VG(0.7,1,1) |
| NC | E1 | VG(0.7,1,1) | VG(0.7,1,1) | VG(0.7,1,1) |
| | E2 | VG(0.7,1,1) | G(0.5,0.7, 1) | VG(0.7,1,1) |
| | E3 | VG(0.7,1,1) | VG(0.7,1,1) | G(0.5,0.7, 1) |
| | E4 | VG(0.7,1,1) | G(0.5,0.7, 1) | VG(0.7,1,1) |
| | E5 | VG(0.7,1,1) | VG(0.7,1,1) | G(0.5,0.7, 1) |
| RBC | E1 | G(0.5,0.7, 1) | G(0.5,0.7, 1) | G(0.5,0.7, 1) |
| | E2 | G(0.5,0.7, 1) | M(0.2,0.5,0.7) | P(0,0.3,0.5) |
| | E3 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) |
| | E4 | M(0.2,0.5,0.7) | G(0.5,0.7, 1) | P(0,0.3,0.5) |
| | E5 | G(0.5,0.7, 1) | M(0.2,0.5,0.7) | G(0.5,0.7, 1) |
| NH | E1 | VG(0.7,1,1) | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) |
| | E2 | G(0.5,0.7, 1) | P(0,0.3,0.5) | P(0,0.3,0.5) |
| | E3 | M(0.2,0.5,0.7) | G(0.5,0.7, 1) | G(0.5,0.7, 1) |
| | E4 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | P(0,0.3,0.5) |
| | E5 | M(0.2,0.5,0.7) | P(0,0.3,0.5) | M(0.2,0.5,0.7) |
| NDC | E1 | G(0.5,0.7, 1) | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) |
| | E2 | M(0.2,0.5,0.7) | P(0,0.3,0.5) | P(0,0.3,0.5) |
| | E3 | P(0,0.3,0.5) | M(0.2,0.5,0.7) | P(0,0.3,0.5) |
| | E4 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | P(0,0.3,0.5) |
| | E5 | M(0.2,0.5,0.7) | G(0.5,0.7, 1) | M(0.2,0.5,0.7) |
| NFC | E1 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | G(0.5,0.7, 1) |
| | E2 | M(0.2,0.5,0.7) | P(0,0.3,0.5) | M(0.2,0.5,0.7) |
| | E3 | P(0,0.3,0.5) | G(0.5,0.7, 1) | P(0,0.3,0.5) |
| | E4 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | P(0,0.3,0.5) |
| | E5 | P(0,0.3,0.5) | P(0,0.3,0.5) | P(0,0.3,0.5) |
| DHP | E1 | M(0.2,0.5,0.7) | P(0,0.3,0.5) | P(0,0.3,0.5) |
| | E2 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | P(0,0.3,0.5) |
| | E3 | G(0.5,0.7, 1) | G(0.5,0.7, 1) | M(0.2,0.5,0.7) |
| | E4 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) |
| | E5 | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) |
| NABC | E1 | VP(0,0,0.3) | M(0.2,0.5,0.7) | M(0.2,0.5,0.7) |
| | E2 | P(0,0.3,0.5) | P(0,0.3,0.5) | P(0,0.3,0.5) |
| | E3 | M(0.2,0.5,0.7) | P(0,0.3,0.5) | VP(0,0,0.3) |
| | E4 | P(0,0.3,0.5) | VP(0,0,0.3) | P(0,0.3,0.5) |
| | E5 | P(0,0.3,0.5) | M(0.2,0.5,0.7) | P(0,0.3,0.5) |
| NAFC | E1 | VP(0,0,0.3) | M(0.2,0.5,0.7) | VP(0,0,0.3) |
| | E2 | VP(0,0,0.3) | VP(0,0,0.3) | M(0.2,0.5,0.7) |
| | E3 | P(0,0.3,0.5) | P(0,0.3,0.5) | P(0,0.3,0.5) |
| | E4 | VP(0,0,0.3) | P(0,0.3,0.5) | P(0,0.3,0.5) |
| | E5 | VP(0,0,0.3) | P(0,0.3,0.5) | M(0.2,0.5,0.7) |
| RSA | E1 | VP(0,0,0.3) | M(0.2,0.5,0.7) | VP(0,0,0.3) |
| | E2 | VP(0,0,0.3) | VP(0,0,0.3) | P(0,0.3,0.5) |
| | E3 | VP(0,0,0.3) | VP(0,0,0.3) | P(0,0.3,0.5) |
| | E4 | P(0,0.3,0.5) | P(0,0.3,0.5) | VP(0,0,0.3) |
| | E5 | VP(0,0,0.3) | VP(0,0,0.3) | VP(0,0,0.3) |

Thirteen quality metrics namely NDC, NFC, NBC, NC, NRFD, DHP, RSA, NH, NAFC, NADC, NABC, NA, RBC have been used for ranking based on three ranking criteria namely understandability, efficiency and effectiveness.

The weights assigned to three ranking criteria and ratings of thirteen quality metrics versus each ranking criteria are assigned in terms of linguistic fuzzy variables by each of the five experts presented in Table 4.3 and Table 4.4.

Using fuzzy triangular aggregation, the aggregated weights (W_i) and aggregate ratings (R_{it}) of A_i metric under criteria C_t were calculated as shown in Table 4.5 and Table 4.6. For example, the aggregated weight of criteria C_1 (understandability) was calculated as:

$$C_1 = 1/5 \otimes [(0.5, 0.8, 1) \oplus (0.5, 0.8, 1) \oplus (0.5, 0.8, 1) \oplus (0.5, 0.8, 1) \oplus (0.5, 0.8, 1)] \\ = 1/5(2.5, 4.0, 5) = (0.5, 0.8, 1)$$

Likewise the aggregated rating of metric A_1 (NA) under criteria C_1 (understandability) was calculated as:

$$A_{11} = 1/5 \otimes [(0.2, 0.5, 0.7) \oplus (0.2, 0.5, 0.7) \oplus (0.5, 0.7, 1) \oplus (0.2, 0.5, 0.7) \oplus (0.5, 0.7, 1)] \\ = 1/5(1.6, 2.9, 4.1) = (0.32, 0.58, 0.82)$$

The crisp scores of these aggregated values are then calculated using methods described in section above and shown in Table 4.7.

Table 4.5 Aggregated weights for criteria ranking

| Criteria | Understandability | Efficiency | Effectiveness |
|----------|-------------------|----------------|----------------|
| W_t | 0.5,0.8,1 | 0.42,0.68,0.88 | 0.22,0.48,0.72 |

Table 4.6 Aggregated rating for quality metrics

| Metrics | Understandability | Efficiency | Effectiveness |
|---------|-------------------|----------------|----------------|
| NA | 0.32,0.58,0.82 | 0.22,0.5,0.72 | 0.24,0.5,0.74 |
| NADC | 0.28,0.54,0.78 | 0.08,0.38,0.58 | 0.04,0.28,0.5 |
| NRFD | 0.48,0.72,0.94 | 0.38,0.62,0.88 | 0.32,0.58,0.82 |
| NBC | 0.66,0.94,1 | 0.62,0.88,1 | 0.62,0.88,1 |
| NC | 0.7,1,1 | 0.62,0.88,1 | 0.62,0.88,1 |
| RBC | 0.38,0.62,0.88 | 0.32,0.58,0.82 | 0.24,0.5,0.74 |
| NH | 0.36,0.64,0.82 | 0.18,0.46,0.68 | 0.18,0.46,0.68 |
| NDC | 0.22,0.5,0.72 | 0.22,0.5,0.72 | 0.08,0.38,0.58 |
| NFC | 0.12,0.42,0.62 | 0.18,0.46,0.68 | 0.12,0.42,0.62 |
| DHP | 0.26,0.54,0.76 | 0.22,0.5,0.72 | 0.12,0.42,0.62 |
| NABC | 0.04,0.28,0.5 | 0.08,0.32,0.54 | 0.04,0.28,0.5 |
| NAFC | 0.0,0.06,0.34 | 0.04,0.28,0.5 | 0.08,0.32,0.54 |
| RSA | 0.0,0.06,0.34 | 0.04,0.16,0.42 | 0.0,0.12,0.38 |

Table 4.7 Values of crisp scores for rating quality metrics

| Metrics | Understandability | Efficiency | Effectiveness |
|----------|-------------------|------------|---------------|
| NA | 0.5733 | 0.48 | 0.4933 |
| NADC | 0.5333 | 0.3466 | 0.2733 |
| NRFD | 0.7133 | 0.6266 | 0.5733 |
| NBC | 0.8666 | 0.8333 | 0.8333 |
| NC | 0.9 | 0.8333 | 0.8333 |
| RBC | 0.6266 | 0.5733 | 0.4933 |
| NH | 0.6066 | 0.44 | 0.44 |
| NDC | 0.48 | 0.48 | 0.3466 |
| NFC | 0.3866 | 0.44 | 0.3866 |
| DHP | 0.52 | 0.48 | 0.3866 |
| NABC | 0.2733 | 0.3133 | 0.2733 |
| NAFC | 0.1333 | 0.2733 | 0.3133 |
| RSA | 0.1333 | 0.2066 | 0.1666 |
| Criteria | 0.7666 | 0.66 | 0.4733 |

The criteria matrix is formed for each quality metric and the value of permanent for each criteria matrix is calculated. For example the criteria matrix for metric NA is constructed as follows:

$$\begin{bmatrix} 0.5733 & 0.8609 & 0.6174 \\ 1.1615 & 0.48 & 0.7171 \\ 1.6196 & 1.3944 & 0.4933 \end{bmatrix}$$

The value of permanent obtained using criteria matrix is then used to rank the quality metrics. The calculated rank values and rank of each quality metric is shown in Table 4.8 as follows:

Table 4.8 Ranking values and rank of quality metrics

| Metrics | Ranking Values | Rank |
|---------|----------------|------|
| NA | 3.6815 | 6 |
| NADC | 3.0704 | 11 |
| NRFD | 4.1686 | 3 |
| NBC | 5.1340 | 2 |
| NC | 5.1906 | 1 |
| RBC | 3.8696 | 4 |
| NH | 3.6539 | 7 |
| NDC | 3.3875 | 9 |
| NFC | 3.2783 | 10 |
| DHP | 3.4823 | 8 |
| NABC | 3.7627 | 5 |
| NAFC | 2.7307 | 12 |
| RSA | 2.5106 | 13 |

Table 4.9 Comparison and analysis with other technique

| Metrics | Ranking Values based on proposed fuzzy method | Rank | Ranking Values based on expert opinion | Rank | Group |
|-------------|---|-----------|--|-----------|-------|
| NC | 5.1906 | 1 | 0.5447 | 1 | G1 |
| NBC | 5.134 | 2 | 0.5361 | 2 | |
| NRFD | 4.1686 | 3 | 0.4102 | 3 | G2 |
| RBC | 3.8696 | 4 | 0.3637 | 4 | G3 |
| NABC | 3.7627 | 5 | 0.1815 | 11 | |
| NA | 3.6815 | 6 | 0.3295 | 5 | |
| NH | 3.6539 | 7 | 0.3208 | 6 | |
| DHP | 3.4823 | 8 | 0.2990 | 7 | |
| NDC | 3.3875 | 9 | 0.2825 | 8 | |
| NFC | 3.2783 | 10 | 0.2583 | 9 | |
| NADC | 3.0704 | 11 | 0.2552 | 10 | |
| NAFC | 2.7307 | 12 | 0.1433 | 12 | |
| RSA | 2.5106 | 13 | 0.1056 | 13 | |

4.5 RESULT ANALYSIS AND COMPARISON

The quality metrics have been given ranking in accordance to their significance towards predicting the understandability, efficiency and effectiveness of conceptual data warehouse models [114, 115, 116]. As can be seen from Table 4.8, the metrics with higher value of permanent are ranked higher in order. The metric NC has been ranked as first owing to its high score for three criteria namely understandability, efficiency and effectiveness. The metric NC is followed by NBC with a slight difference in values of permanent i.e. 5.1340 for NBC and 5.1906 for NC. The metric NBC is followed by NRFD with a score of 4.1686. The successive metrics in order are RBC, NABC, NA, NH, DHP, NDC, NFC, NADC with a score of 3.8696, 3.7627, 3.6815, 3.6539, 3.4823, 3.3875, 3.2783, 3.0704. The score for permanent of these metrics vary in fractions showing their relatively similar significance on the quality of conceptual models. The metrics NAFC and RSA have lowest ranks due to their low score of permanent.

This way we can categorize the metrics into 4 groups based on their values of permanent.

G1 = [NC, NBC]

G2 = [NRFD]

G3 = [RBC, NABC, NA, NH, DHP, NDC, NFC, NADC]

G4 = [NAFC, RSA]

Table 4.10 Input to rank based on expert opinion

| Metrics | Experts | Understandability | Efficiency | Effectiveness |
|---------|---------|-------------------|------------|---------------|
| NA | E1 | 0.466 | 0.466 | 0.733 |
| | E2 | 0.466 | 0.466 | 0.733 |
| | E3 | 0.733 | 0.733 | 0.466 |
| | E4 | 0.466 | 0.466 | 0.266 |
| | E5 | 0.733 | 0.266 | 0.266 |
| NADC | E1 | 0.733 | 0.466 | 0.266 |
| | E2 | 0.733 | 0.266 | 0.266 |
| | E3 | 0.466 | 0.266 | 0.100 |
| | E4 | 0.266 | 0.466 | 0.466 |
| | E5 | 0.466 | 0.266 | 0.266 |
| NRFD | E1 | 0.733 | 0.733 | 0.733 |
| | E2 | 0.900 | 0.733 | 0.466 |
| | E3 | 0.733 | 0.466 | 0.466 |
| | E4 | 0.466 | 0.466 | 0.733 |
| | E5 | 0.733 | 0.733 | 0.466 |
| NBC | E1 | 0.900 | 0.733 | 0.733 |
| | E2 | 0.900 | 0.900 | 0.900 |
| | E3 | 0.733 | 0.900 | 0.900 |
| | E4 | 0.900 | 0.733 | 0.733 |
| | E5 | 0.900 | 0.900 | 0.900 |
| NC | E1 | 0.900 | 0.900 | 0.900 |
| | E2 | 0.900 | 0.733 | 0.900 |
| | E3 | 0.900 | 0.900 | 0.733 |
| | E4 | 0.900 | 0.733 | 0.900 |
| | E5 | 0.900 | 0.900 | 0.733 |
| RBC | E1 | 0.733 | 0.733 | 0.733 |
| | E2 | 0.733 | 0.466 | 0.266 |
| | E3 | 0.466 | 0.466 | 0.466 |
| | E4 | 0.466 | 0.733 | 0.266 |
| | E5 | 0.733 | 0.466 | 0.733 |
| NH | E1 | 0.900 | 0.466 | 0.466 |
| | E2 | 0.733 | 0.266 | 0.266 |
| | E3 | 0.466 | 0.733 | 0.733 |
| | E4 | 0.466 | 0.466 | 0.266 |
| | E5 | 0.466 | 0.266 | 0.466 |
| NDC | E1 | 0.733 | 0.466 | 0.466 |
| | E2 | 0.466 | 0.266 | 0.266 |
| | E3 | 0.266 | 0.466 | 0.266 |
| | E4 | 0.466 | 0.466 | 0.266 |
| | E5 | 0.466 | 0.733 | 0.466 |
| NFC | E1 | 0.466 | 0.466 | 0.733 |
| | E2 | 0.466 | 0.266 | 0.466 |
| | E3 | 0.266 | 0.733 | 0.266 |
| | E4 | 0.466 | 0.466 | 0.266 |
| | E5 | 0.266 | 0.266 | 0.266 |
| DHP | E1 | 0.466 | 0.266 | 0.266 |
| | E2 | 0.466 | 0.466 | 0.266 |
| | E3 | 0.733 | 0.733 | 0.466 |
| | E4 | 0.466 | 0.466 | 0.466 |
| | E5 | 0.466 | 0.466 | 0.466 |
| NABC | E1 | 0.100 | 0.466 | 0.466 |
| | E2 | 0.266 | 0.266 | 0.266 |
| | E3 | 0.466 | 0.266 | 0.100 |
| | E4 | 0.266 | 0.100 | 0.266 |
| | E5 | 0.266 | 0.466 | 0.266 |
| NAFC | E1 | 0.100 | 0.466 | 0.100 |
| | E2 | 0.100 | 0.100 | 0.466 |
| | E3 | 0.266 | 0.266 | 0.266 |
| | E4 | 0.100 | 0.266 | 0.266 |
| | E5 | 0.100 | 0.266 | 0.466 |
| RSA | E1 | 0.100 | 0.466 | 0.100 |
| | E2 | 0.100 | 0.100 | 0.266 |
| | E3 | 0.100 | 0.100 | 0.266 |
| | E4 | 0.266 | 0.266 | 0.100 |
| | E5 | 0.100 | 0.100 | 0.100 |

The results of proposed fuzzy based approach are compared with the results based on expert opinion [87] can be seen from Table 4.9. The input data given to ranking based on expert opinion is given in Table 4.10. It can be seen that the results of proposed fuzzy methodology are consistent with the results based on expert opinion. The ranking of six highlighted metrics in Table 4.9, namely NRFD, NBC, NC, RBC, NAFC, RSA are exactly same for two approaches. The ranks of NA, NADC, NH, NDC, NFC and DHP differ by one. The ranking for NABC shows variation with a rank of 5 in proposed approach and 11 in expert opinion approach. The number of elements and their ranking is exactly same for groups G1, G2 and G4. The number of elements is same for group G3 with slight differences of ranks owing to fractional differences in the values of their permanent. So the results obtained by proposed methodology are consistent with the results based on expert opinion. The comparison of proposed methodology with expert opinion approach along certain parameters is shown in Table 4.11.

Table 4.11 Comparison based on various parameters

| S.No. | Parameters | Proposed fuzzy methodology | Expert opinion approach |
|-------|--|---|---|
| 1 | Number of computations | Proportional to number of attributes (N) i.e. experts, metrics and criteria | Proportional to number of attributes (N) i.e. experts, metrics and criteria |
| 2 | Weight matrix | Fuzzy aggregation | Algebraic aggregation |
| 3 | Rate matrix | Fuzzy aggregation | Algebraic aggregation |
| 4 | Criteria matrix | Fuzzy aggregation | Algebraic aggregation |
| 5 | Consideration of all possible interdependencies of variables | Yes | No |
| 6 | Rank of metrics | Yes | Yes |
| 7 | Accuracy | More due to fuzzy base approach | Lesser due to algebraic approach |

It can be seen from the comparison table that the results of fuzzy based approach are more reliable, accurate as compared to expert opinion approach due to the consideration of ambiguity, imprecision prevalent in human thought process and consideration of all interdependencies of attributes by the use of permanent function. Proceeding further, the rankings of quality metrics have been used to develop a fuzzy based rule base to predict the understanding time of conceptual models. The detailed study is presented in the following sections.

4.6 FUZZY RULE BASE FOR PREDICTING UNDERSTANDABILITY OF CONCEPTUAL MODELS

To predict the understandability of data warehouse conceptual models, efforts were involved towards design of conceptual models, identification of subjects, preparation of questionnaires based on structural properties of models, collection of data in the form of time and then further aggregating the data to get understanding time of model. To minimize the efforts involved in prediction of understanding time, the need for a system that could predict the understanding time of conceptual models was felt. The research work was carried on towards building of a fuzzy rule base system based on ranking of quality metrics. The values of quality metrics are given as input to the system and understanding time is the output. To measure the efficiency of rule base system, the predicted results were compared with actual understanding time (calculated before). The details are presented in following sections.

4.6.1 Component Classification

The basic components of a fuzzy rule base system for predicting understanding time are given in Figure 4.7.

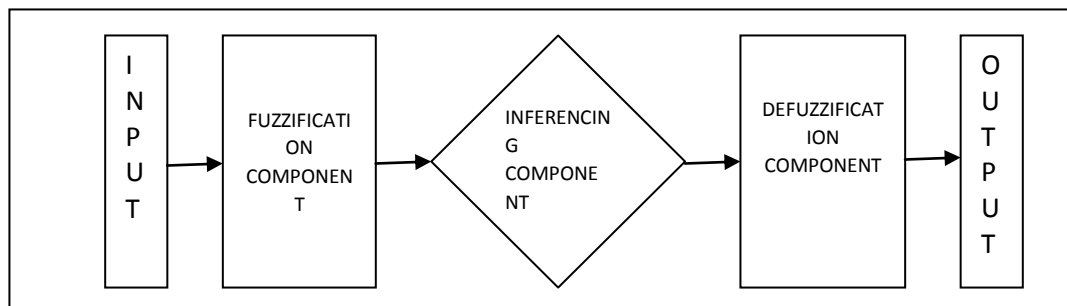


Figure 4.7 Fuzzy inference system [22]

As can be seen from Figure 4.7, there are three basic components of a fuzzy rule base system namely Fuzzification component, Inference component and Defuzzification component. A brief overview of each of these components [22] is defined as follows:

- Fuzzification component: The input given to the system is mapped to fuzzy membership set. The inputs are the values of quality metrics for a conceptual model.
- Inference component: The module makes use of a fuzzy rule base for processing inputs. IF-THEN rules are stored in a fuzzy rule base, which are

referred for input processing. The working of inference system is a two-step process namely antecedent (computing values for IF part of rules) and consequent (computing values for THEN part of the rules).

- Defuzzification component: This component maps fuzzy to crisp output.

To make use of fuzzy rule base system for predicting understanding time of conceptual models, a stepwise approach was followed as presented in the next section.

4.6.2 Stepwise Approach

The section discusses a stepwise approach followed by us for developing a fuzzy rule base system for predicting the understanding time of conceptual models. Matlab has been used to generate and check the validity of results. The type of fuzzy logic system used for inference (fuzzification and defuzzification) is Mamdani system. The approach consists of following steps:

- Identification and ranking of quality metrics: Thirteen quality metrics namely NFC, NDC, NBC, NC, NAFC, NADC, NABC, NA, NH, DHP, RSA, NRFD and RBC (explained in chapter III) were identified for pursuing research towards development of a fuzzy rule base system. The identified metrics were ranked in accordance to their significance towards quality evaluation of conceptual models (discussed in previous sections).
- Minimization of identified metrics: Significant and more relevant metrics were identified, so as to prepare a simple rule base. The increase in number of parameters leads to a large complex rule base. A big rule base increase the time and space complexity of search, match and hit of correct rule towards prediction of understanding time. More relevant metrics were identified out of the metrics towards development of a simple rule base that could give correct results in minimum possible time. The relevant metrics identified were NC, NA, NH, DHP and NRFD. The justifications for the choice are as follows:
 - NC was selected as it is the total sum of all the classes be it number of dimension class, fact class or base class. $NC = NFC + NDC + NBC$.
 - NA was selected as it is total sum of attributes of all the classes be it fact class, dimension class or base class. $NA = NAFC + NADC + NABC$.
 - NH was selected as it provides a measure of total number of hierarchies in the model and hence is significant towards quality evaluation.

- DHP was selected as it provides a measure of path of the longest hierarchy in the model and hence it is significant.
- NRFD was selected as it defines the complexities of models by giving a measure of number of relations between facts and dimension classes and hence it is significant.
- RBC was not selected as its value is between 0-3 most of the times and it makes a marginal/insignificant contribution towards quality evaluation.
- RSA was not selected as its value is always fractional and less than one (99% times) and makes minor contribution towards quality evaluation.

All of the identified metrics have direct proportionality with understanding time. Also for development of fuzzy rule base system metric values of the set of 22 models, whose understanding times have been calculated and discussed in previous chapter, were taken for verification of correctness of developed model with respect to the actual calculated times.

- Construct fuzzy memberships: Based on the opinion of experts, fuzzy membership functions for the identified metrics were created. The fuzzy membership functions for NC, NA, DHP, NH, NRFD and Understanding Time (UT) were defined in terms of three linguistic variables low, medium and high. Also the range of values for each of the linguistic variables and respective fuzzification of each of the metrics is shown by Figures 4.8, 4.9, 4.10, 4.11, 4.12, 4.13.

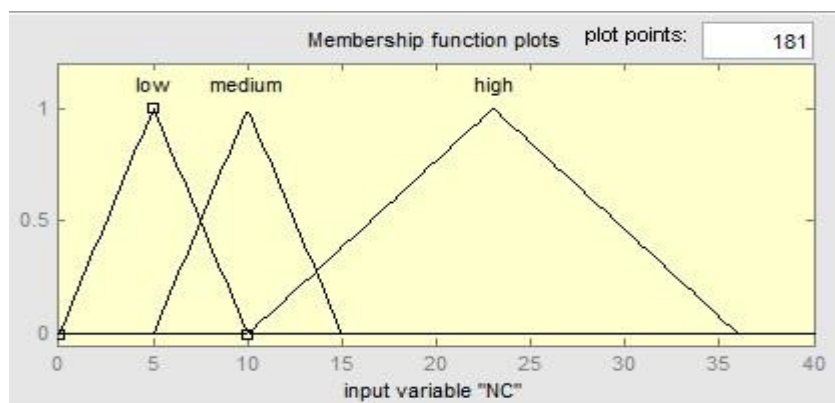


Figure 4.8 Fuzzification of NC

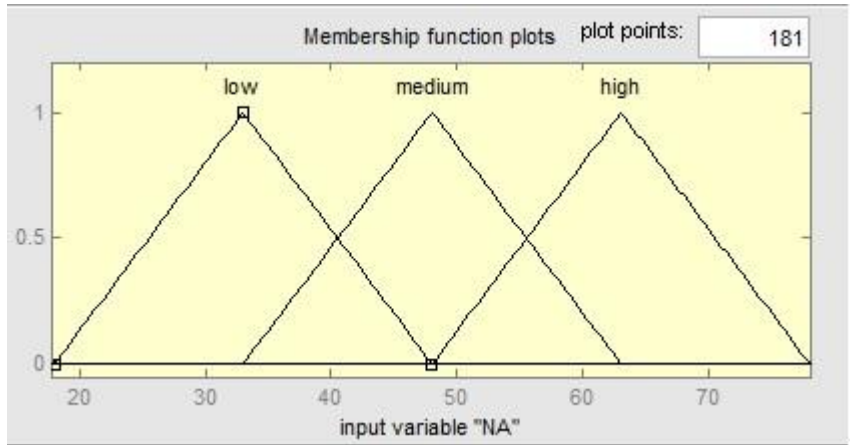


Figure 4.9 Fuzzification of NA

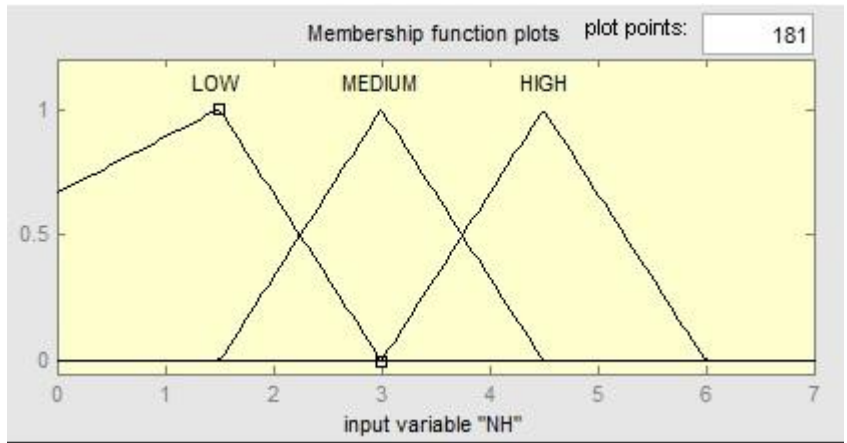


Figure 4.10 Fuzzification of NH

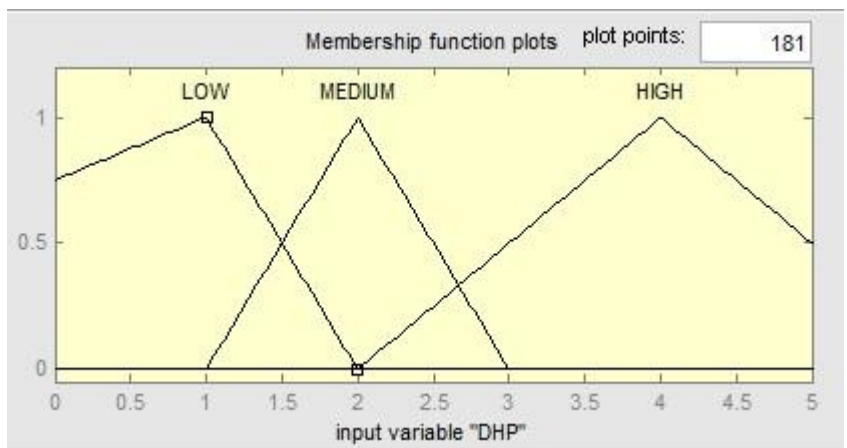


Figure 4.11 Fuzzification of DHP

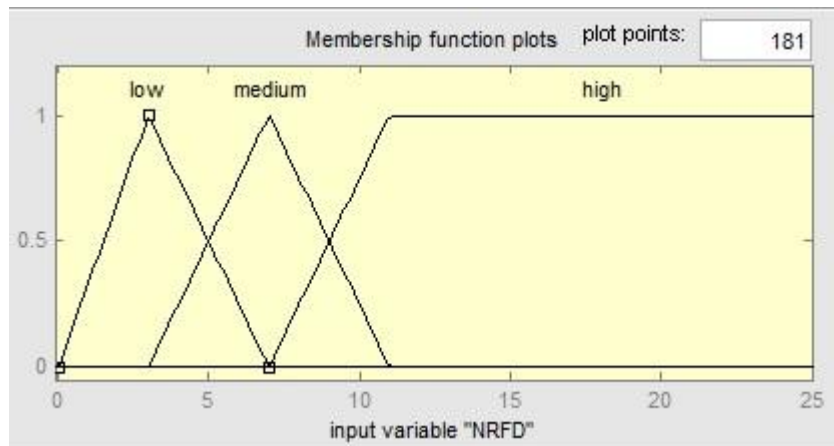


Figure 4.12 Fuzzification of NRFD

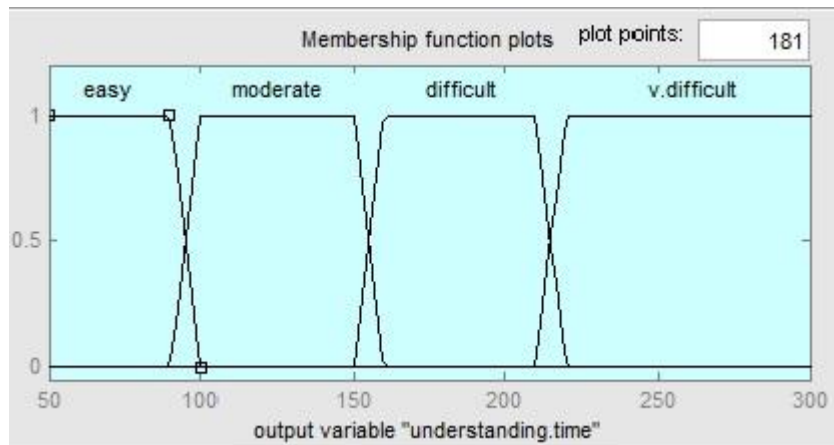


Figure 4.13 Fuzification of Understanding Time

Three linguistic variables (low, medium, high) are defined for NC, NA, DHP, NH, NRFD and four (easy, moderate, difficult, very difficult) linguistic variables were defined for UT. The corresponding range of values for each of the linguistic variables for metrics is shown by Table 4.12.

Table 4.12 Fuzzy linguistic variables for metrics

| Metrics | Low | Medium | High |
|---------|------------|-------------|------------|
| NC | [0 5 10] | [5 10 15] | [10 23 36] |
| NA | [18 33 48] | [33 48 63] | [48 63 78] |
| NH | [-3 1.5 3] | [1.5 3 4.5] | [3 4.5 6] |
| DHP | [-3 1 2] | [1 2 3] | [2 4 6] |
| NRFD | [0 3 7] | [3 7 11] | [7 11 25] |

The corresponding range of values for each of the linguistic variables of understanding time is shown by Table 4.13.

Table 4.13 Fuzzy linguistic variables for understanding time

| Understanding Time | Easy | Moderate | Difficult | Very difficult |
|--------------------|----------------|------------------|-------------------|-------------------|
| UT | [40 50 90 100] | [90 100 150 160] | [150 160 210 220] | [210 220 300 310] |

- Identification of quality attribute: The quality attribute along which the results of the rule base fuzzy system were predicted and validated was identified as Understanding Time (UT).
- Construct rule base: Based on the ranking of quality metrics, discussed in previous sections, heuristic rules were created that mapped the fuzzified values of metrics, given as input to rule base, to value of quality attribute generated as output. A snapshot of the fuzzy rule base system is given in Figure 4.14.

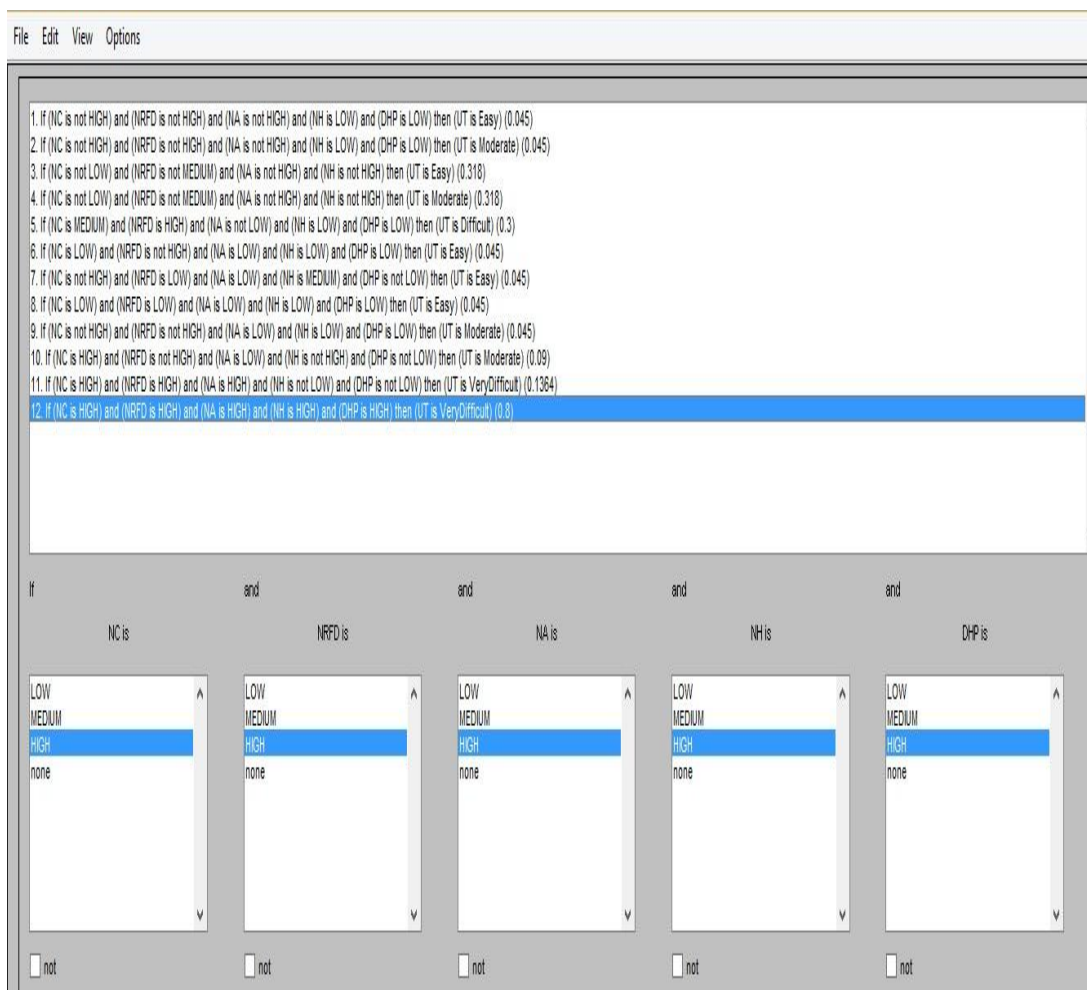


Figure 4.14 Fuzzy rule base

- Defuzzification: Centroid of area method (most widely used) was used for defuzzification of the results produced by fuzzy rule base. The defuzzification gave as output crisp values for the understanding time of each of the given models.

4.7 RESULT ANALYSIS

The predicted results as well as results calculated manually are presented by Table 4.14.

Table 4.14 Predicted vs Calculated Results

| Schema | NC | NA | NH | DHP | NRFD | Predicted Understanding time (results using fuzzy rule base) | UT Category | Average Understanding time (calculated manually) | UT Category |
|--------|----|----|----|-----|------|--|-------------|--|-------------|
| 1 | 14 | 30 | 4 | 3 | 4 | 99.2 | moderate | 103 | moderate |
| 2 | 19 | 31 | 5 | 4 | 5 | 99.2 | moderate | 121 | moderate |
| 3 | 17 | 28 | 5 | 3 | 5 | 99.2 | moderate | 106 | moderate |
| 4 | 11 | 19 | 3 | 3 | 3 | 100 | easy | 96 | easy |
| 5 | 13 | 24 | 4 | 3 | 4 | 99.2 | easy | 94 | easy |
| 6 | 15 | 26 | 4 | 3 | 4 | 99.2 | easy | 96 | easy |
| 7 | 09 | 20 | 3 | 3 | 3 | 99.8 | easy | 88 | easy |
| 8 | 12 | 25 | 4 | 3 | 4 | 99.2 | easy | 98 | easy |
| 9 | 13 | 30 | 4 | 3 | 6 | 99.2 | easy | 95 | easy |
| 10 | 13 | 20 | 4 | 3 | 4 | 99.2 | moderate | 99 | moderate |
| 11 | 17 | 66 | 4 | 3 | 13 | 234 | v.difficult | 209 | difficult |
| 12 | 17 | 66 | 4 | 3 | 15 | 234 | v.difficult | 272 | v.difficult |
| 13 | 19 | 71 | 5 | 3 | 12 | 218 | v.difficult | 224 | v.difficult |
| 14 | 19 | 71 | 5 | 3 | 15 | 218 | v.difficult | 295 | v.difficult |
| 15 | 23 | 35 | 0 | 0 | 22 | 100 | moderate | 97 | moderate |
| 16 | 04 | 29 | 0 | 0 | 3 | 99.2 | moderate | 104 | moderate |
| 17 | 07 | 25 | 1 | 1 | 5 | 99.2 | moderate | 126 | moderate |
| 18 | 07 | 21 | 0 | 0 | 6 | 99.2 | easy | 91 | easy |
| 19 | 07 | 29 | 0 | 0 | 6 | 99.2 | easy | 100 | easy |

The output (understanding time) of the fuzzy rule base system is predicted as follows: Let us consider model no. 1. The inputs for the model are NC=14, NA=30, NH=04, DHP=03, NRFD=04. The input values are fuzzified as NC to medium, NA to low, NH to medium, DHP to medium, NRFD to low. The fuzzy inference system matches the fuzzy inputs to rules to give the output, which is then defuzzified to crisp value i.e. 99.2 for model no. 1. Out of the 22 models, the fuzzy rule base system designed gave predicted results similar to average calculated values for 19 models and gave different results for 03 models. So the efficiency of rule base designed is 86.36% (19/22*100). The dissimilarity in results could be attributed to the contribution of other quality metrics not considered i.e. RBC, RSA.

The results of the fuzzy rule base inference system are achieved using fuzzy tool box in MATLAB. Rule viewer was used to observe predicted values of understanding time. The output of the rule viewer for model no. 1 is presented by snapshot of Figure 4.15.



Figure 4.15 Output of rule viewer for model no. 1

The ranking of quality metrics have been made use of to design a fuzzy rule base for predicting the understanding time of conceptual data warehouse models. The fuzzy prediction model can be enriched by including more number of experts and expanding the domain of conceptual models, to provide more accurate results for usage by data analysts to predict the quality of conceptual data warehouse models. All of the research work presented so far was related to quality evaluation phase of data warehouse development.

The next chapter focus on the research work carried out in next stage of data warehouse development which is information extraction/retrieval (already discussed in literature review). Efficient information extraction from a data warehouse has been another main focus of research which is discussed in detail in the following chapter.

4.8 SUMMARY

The chapter presented research work carried on one of the major aspects towards building of an efficient information delivery system in which a fuzzy inference system was designed for predicting the understanding time of conceptual data warehouse models. For the designing of inference system the quality metrics were first ranked based on the opinion of experts. It has been found that the quality metrics have significant effect towards quality evaluation of models which is proved on the basis of theoretical and empirical validation conducted in the previous chapter.

CHAPTER V

EFFICIENT INFORMATION EXTRACTION TOWARDS BUILDING OF EIDS

5.1 INTRODUCTION

This chapter discusses related research work carried out in information extraction phase of data warehouse development. As discussed in literature review, the information extraction phase was next to quality evaluation phase in terms of research work carried out and published by researchers in the past few years. This chapter presents the research work carried by research scholar in the domain of efficient information extraction from data storehouses towards building of EIDS.

It is well known that efficient information delivery systems [23, 99] are being used by all business enterprises to gain competitive advantage in current market scenario. All the efficient information delivery systems are supported by huge data warehouses at the backend. Complex queries run on data warehouses and results achieved. Query response time is one of the major factors affecting the quality of data warehouses. A number of query optimization [106] techniques exist of which one such base technique involves greedy selection of views. The greedy approach [23] selects the view that has maximum cost benefit (explained in successive sections) among all the views not selected so far.

We have used the greedy view selection approach proposed by Harinarayanan et al [23] as base approach to carry out further research work in the related domain. The research work has been carried out towards enhancement of approach proposed by Harinarayanan et al [23] that lead to a refined greedy selection approach which makes use of forward references to give better materialized view selection. The proposed approach uses lattice framework of data that shows inter dependencies of data. The choice of materialized views using the proposed approach gives a better trade off in terms of space/benefits, which is proved from the experimental results. The refined greedy selection approach is independent of space constraint and depends on number

of passes entered by the user. The view selection is further enhanced by including space constraints to the results of greedy and refined greedy approach using knapsack implementation.

Before proceeding further to subsequent sections, the familiarity with few related terms is necessary. The terms are as follows:

- **View:** A derived relation/result in response to a query. It is defined in terms of base relation and/or combination of attributes [106]. Each cell in multidimensional cubes forms a view.
- **Materialized View:** A view is materialized if its result in response to query is stored in memory [106]. It is the set of materialized views whose optimal selection improves query optimization.
- **View Selection:** It aims at selecting a set of materialized views given some database to optimize query response time [106]. The optimal view selection improves query response time and is one of the main factors affecting query optimization of decision support systems.
- **Data Cubes:** The data in a data warehouse is viewed along multiple dimensions. Multidimensional analysis of data is graphically represented as data cubes [106].
- **Lattice:** A Lattice [103] is a graphical framework used to show dependencies among multiple views of a multi-dimensional data warehouse. A lattice is a ordered collection of views to which view selection approaches are applied to get optimal subset of materialized views. The optimal selection of views enhances query optimization. A Lattice is a graphical framework used to show dependencies among multiple views of a multi-dimensional data warehouse.

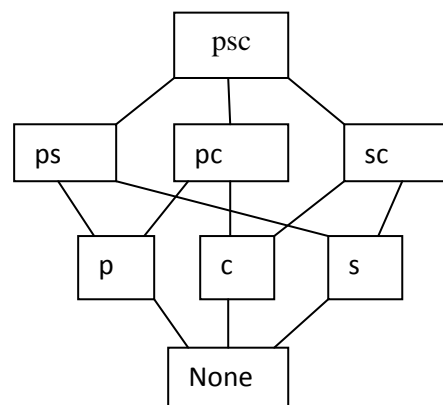


Figure 5.1 Lattice Framework 1 (Source: Harinarayan et al. [23])

Consider the example of a business data warehouse that stores information about various parts bought from suppliers and sold to various customers. The measure/fact is analysed along three dimensions/attributes part(p), supplier(s), customer(c) is sales.

Figure 5.1 shows a lattice framework [23] for business data warehouse having 3 attributes/dimensions. Each rectangular box is a view or node. The node psc shows a view having three dimensions of part, customer and supplier. Likewise interpretation can be made for all other nodes. The hierarchy between nodes is clearly shown in Figure 5.1. The node having larger dimensions is at a higher level than a node with comparative smaller dimensions. The lines connecting the boxes show dependencies. The three lines from psc to ps, pc, sc show that views ps, pc, sc can be generated from psc or are dependent on psc. The same holds for other lines in the lattice framework. The nodes along with number of rows in each view can also be stated as follows. Here M stands for million.

- Part, Supplier, Customer (6 M rows)
- Part, Customer (6 M)
- Part, Supplier (0.8 M)
- Supplier, Customer (6M)
- Part (0.2M)
- Supplier (0.01M)
- Customer (0.1M)
- None (1)

Suppose a user queries for the total sales grouped by supplier. If view 6 is materialized, then only 0.01M rows need to be processed. The same query can be answered using view 4 that needs 6M rows to be processed. It takes more time to process 6M rows compared to 0.01M rows. Optimal selection of materialized views can greatly minimize space and improve query response time. There are few notations that must be known before proceeding further.

Lattice Notation

As seen, the query (supplier) can be answered using query (supplier, customer). This can be shown as

$$(s) \leq (s,c)$$

The operation \leq is a partial order relation. A lattice satisfies the property that any two elements of the lattice must have a least upper bound and greatest lower bound. A lattice L is denoted as (L, \leq) .

5.2 ADVANTAGES OF LATTICE FRAMEWORK

The use of lattice framework for materialized view selection for query optimization offers following advantages to users:

- The lattice framework gives the users a friendly, easy to understand and graphical view of the multidimensional data warehouse. [23]
- Data dependencies and hierarchies are presented in such a way that the users can visualize and understand complete database at a glance. [23]
- By having a formalized view of data warehouse, the users can analyse data quickly and efficiently. [23]

5.3 OPTIMAL GREEDY SELECTION OF MATERIALIZED VIEWS

The focus of this study is to make a optimal selection of views to improve query execution/response time. The view selection deal with following aspects [23]:

- The first aspect is optimization of query execution time. The time taken to respond to a query should be as small as possible.
- The second aspect deals with optimal selection of fixed number of materialized views with no space constraint.
- The third aspect deals with optimal selection of fixed number of materialized views with limited space.

Before discussing the optimal view selection greedy techniques in detail, the following assumptions are the assumptions made for the study:

- A linear cost model has been used to calculate the cost of answering a query. The cost of answering a query is taken equal to the space/number of rows occupied by the view from which a particular query is answered,
 - $T = mS + C$
 - T = Running time of a query on a view of size S
 - S = View size
 - C = Fixed cost

- $m =$ Query execution time/size of view
- View selection approach requires the calculation of size of each view in advance. This is achieved using statistical sampling techniques that select a small representative of raw data. The selected data is then actually materialized. The calculations made for the representative data are used to predict values of the actual raw data.

5.3.1 Greedy Algorithm for View Selection

Given a Lattice (L, \leq) . Each node (v) in the lattice is associated with space cost $C(v)$, which is the number of rows in a given view. The top view is always included in materialized view set as no other view can answer the queries corresponding to top view. k more views are to be selected using greedy approach. Suppose S is the number of views already selected. The benefit [23] of a view v relative to S , denoted as $B(v, S)$, is as follows:

- For each view $w \leq v$, define B_w as.
 - Let u be the view of least cost in S such that $w \leq u$.
 - If $C(v) \leq C(u)$, then $B_w = C(v) - C(u)$. Otherwise $B_w = 0$.
- Define $B(v, S) = \sum_{w \leq v} B_w$.

After selecting a set of view, the benefit derived from materializing that set of views is calculated. The benefit of materializing a certain view 'v' is the improved cost of evaluating views linked to view 'v' and itself. The total benefit $B(v, s)$ is the sum over all views w of the benefit of using v to evaluate w . The greedy algorithm [23] to materialize set of k views is as follows:

```

S = {top view};
For i = 1 to k do begin
  Select that view v not in S such that B(v,S) is maximized;
  S = S union {v};
End;
Resulting S is greedy selection;

```

The Greedy Algorithm (source: Harinarayan et al [23])

Consider the lattice shown in Figure 5.2 consisting of 8 nodes. Each node is associated with its respective space costs. The nodes are labelled a to h.

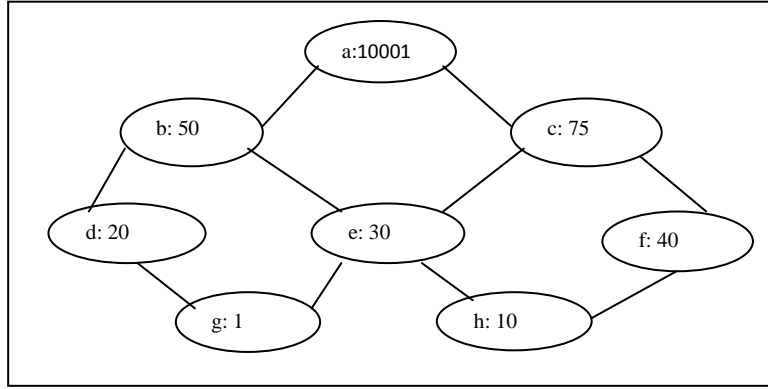


Figure 5.2 Lattice Framework 2 (source: Harinarayan et al [23])

The results of the greedy algorithm when applied to lattice of Figure 5.2 for $k = 3$ excluding top view are shown in Table 5.1.

Table 5.1 Greedy selection

| Nodes | Choice 1 | Choice 2 | Choice 3 |
|-------|------------|------------|--------------|
| b | $50*5=250$ | | |
| c | $25*5=125$ | $25*2=50$ | $25*1=25$ |
| d | $80*2=160$ | $30*2=60$ | $30*2=60$ |
| e | $70*3=210$ | $20*3=60$ | $2*20+10=50$ |
| f | $60*2=120$ | $60+10=70$ | |
| g | $99*1=99$ | $49*1=49$ | $49*1=49$ |
| h | $90*1=90$ | $40*1=40$ | $30*1=30$ |

Pass 1 selects node b (benefit 250) as its benefits are maximum amongst all other nodes. Similarly node f and node d are selected in Pass 2, Pass 3. The greedy algorithm works fine with the lattice specified above.

5.3.2 Need for Refined Greedy Approach

Greedy algorithm proposed by Harinarayan et al [23] does not specify to deal with the situation when one or more nodes with same maximum benefits are encountered in the same pass. The view that is selected in the $(n-1)^{th}$ pass do have an impact on the benefits of the $(n)^{th}$ pass in accordance to greedy approach. A choice is to be made between the two nodes giving same benefits. Consider the example lattice shown in Figure 5.3 below consisting of 8 nodes a to h. Each node is associated with its space costs.

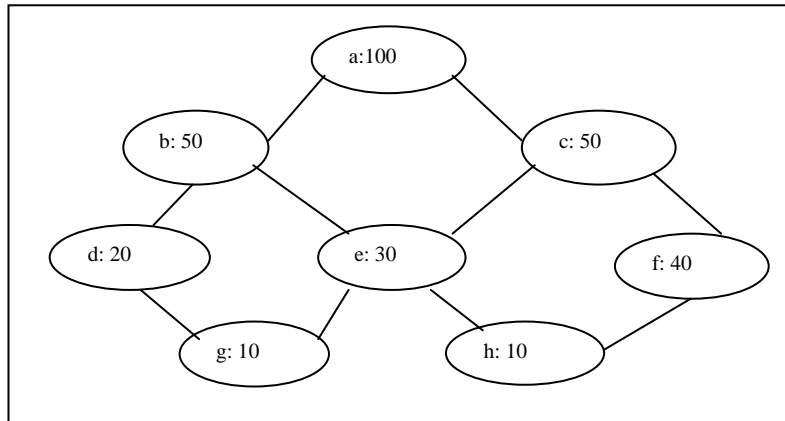


Figure 5.3 Example Lattice Framework

View ‘a’ is trivially assumed to be materialized. In pass1, we calculate the benefits of materializing all other nodes. The benefits are as follows:

Pass 1

| | |
|---|----------|
| B | 50x5=250 |
| C | 50x5=250 |
| D | 80x2=160 |
| E | 70x3=210 |
| F | 60x2=120 |
| G | 90x1=90 |
| H | 90x1=90 |

It can be seen that there is a conflict in pass 1 itself. A choice is to be made, whether to materialize view b or view c. The greedy algorithm is silent on this aspect of the view materialization problem.

5.4 REFINED GREEDY ALGORITHM WITH FORWARD REFERENCING

Picking up the point of greedy algorithm not talking about views having same maximum benefit, a refined greedy algorithm has been proposed that takes into account the concept of greedy algorithm for materialized view selection but does provide a way out when two or more views have the same maximum benefit at a particular pass. The algorithm aims to give a better selection of materialized views in terms of two aspects:

- Space
- Benefit

The Refined greedy algorithm is an improvement over the greedy algorithm for materialized view selection. When confronted with a choice to make, between multiple views having the same maximum benefit, the benefits of the next subsequent pass are calculated, taking each of the multiple views in set of materialized views one by one. The benefits are calculated for all the nodes in the subsequent pass by including each view having same maximum benefits in the set of materialized views one by one. The inclusion of a node in the set of materialized views in $(n-1)^{\text{th}}$ pass affects the benefits at $(n)^{\text{th}}$ pass as per the greedy strategy. A comparison of the maximum benefits of the nodes in the subsequent pass is made and then choice is made for the node of previous pass corresponding to which node in the subsequent pass gives maximum benefit.

If a situation arises where the multiple views in the subsequent pass have same maximum benefits, then we check for space occupied by views in the subsequent pass. That view of the previous pass is selected corresponding to which a view has maximum cost benefit and minimum occupied space in the subsequent pass. The proposed approach considers cost benefits along with space occupied in greedy selection of views. The pseudo code for the proposed refined greedy algorithm is presented as follows:

```

S={Top View} // Selected view

For i=1 to k do begin // k is the no. of passes specified by the user
{
Calculate B(V,S) for view not in S // Benefit of all views V relative to selected view S

Find Bmax (V,S) // Find a view with maximum benefits

Find num(Bmax (V,S)) // Check whether more than one view has same max. benefit

if( num(Bmax (V,S)) == 1 ) // If there exist a single view has max. benefit
{
    Select V Corresponding to Bmax

S = S U V // Set of selected views
}

else // If more than one view has same max. benefit

{
For j=1 to n do begin // n= number of views having same maximum benefit
{
Sj = S U Vj // Include each view one by one in the selected top view

Calculate B(V,Sj) // Find benefit of all views relative to Sj

```

```

T[] = Bmax(V,Sj) // Store the max. benefit in a matrix T[]
V[] = V// Store the view giving max. benefit in matrix V[]
Vj[] = Vj// Store each view Vj in a matrix Vj[]
}

Find max(T[])// Find the max. value of benefit stored in matrix T[]

if(num(max(T[])) == 1)// If there exist a single max. value exist in matrix T[]
{
Vj = Select(Vj[]) corresponding to max(T[])
}
else
{
Vj = Select (Vj[]) such that Space(Vj) = minSpace(Vj[]) corresponding to max(T[])
// Select that view which occupies min. space and gives max. benefit
}
}

if ((i+1)<=k)// Step to select next view by omitting intermediate calculations
{
V = Select (V[]) such that Space(V)=minSpace(V[]) corresponding to selected Vj
// Select a view from matrix V[] that occupies min. space and gives max. benefitcorresponding to
selected view Vj

S=SUVjUV// Set of selected views

i=i+1
} else
S=SUVj // Set of selected views
}

```

Proposed refined greedy algorithm

5.5 EXPERIMENTAL RESULTS

The refined greedy algorithm proposed above was implemented and the results compared with the existing greedy algorithm. The snapshots of the results are shown. The lattice shown in Figure 5.3 was used for experimentation.

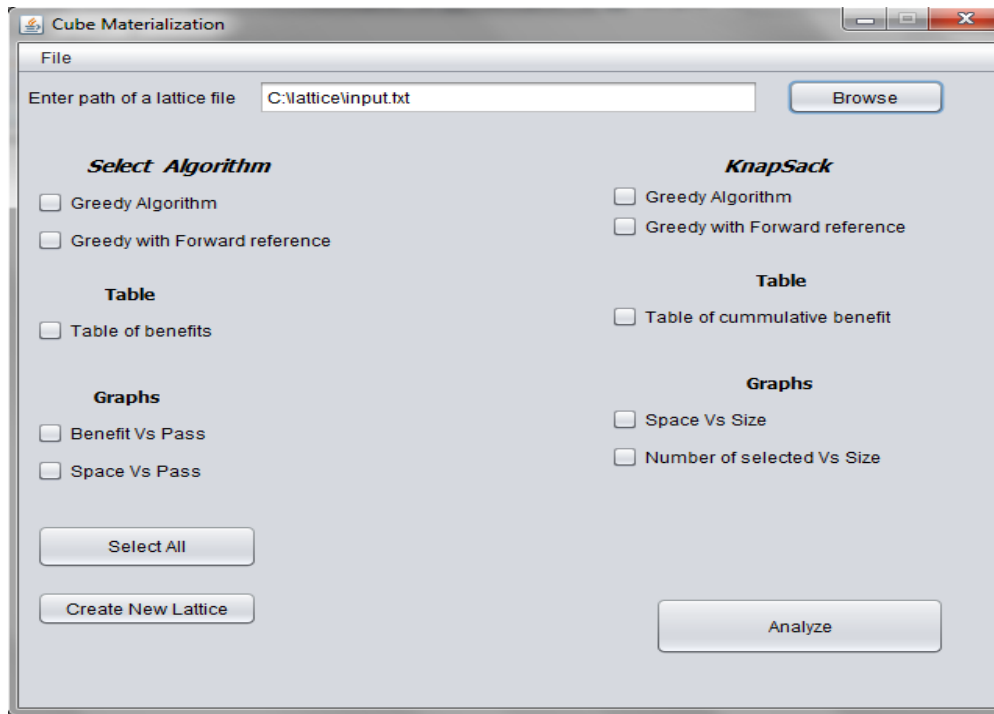


Figure 5.4 Cube Materialization Form

Figure 5.4 provide options to enter the lattice as input and calculate the required results for greedy algorithm and refined greedy algorithm in the form of tables and graphs.

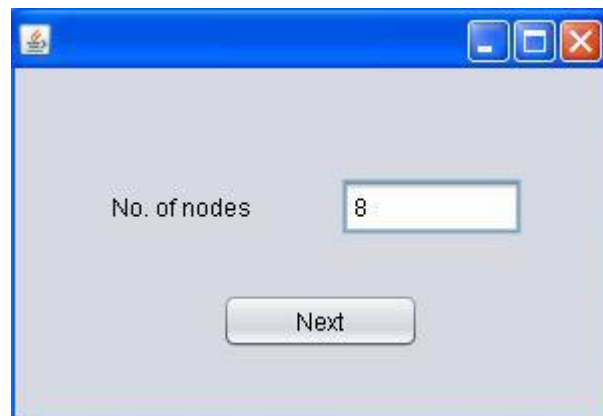


Figure 5.5 Node Entry Form

The icon create new lattice when clicked creates the form shown in Figure 5.5. The number of nodes is entered and next icon clicked to create a new form shown in Figure 5.6.

| SNo. | Node Name | Space | Links |
|------|-----------|-------|-------|
| 1 | a | 100 | b,c |
| 2 | b | 50 | d,e |
| 3 | c | 50 | e,f |
| 4 | d | 20 | g |
| 5 | e | 30 | g,h |
| 6 | f | 40 | h |
| 7 | g | 10 | |
| 8 | h | 10 | |

Save

Figure 5.6 Lattice Entry Form

Figure 5.6 has node field in which all the nodes of lattice are entered. The space field is entered with the space cost associated with each node. The links field stores all the nodes to which a node in node field is directly linked. After entries are made, the save icon is clicked to generate the form shown in Figure 5.7.

Save

Look In: lattice

input.txt

File Name: input1

Files of Type: lattice files (*.txt)

Save Cancel

Figure 5.7 Lattice Storage Form

The file name for the lattice file is entered in File Name field and save icon clicked to save the lattice file in folder lattice. The icon Browse in Figure 5.4 is clicked to

include the created lattice file for processing. Then a Form is generated as shown in Figure 5.8 in which the number of passes for the algorithms is to be entered. The maximum number of passes can be one less than total number of nodes, as the top view is always selected. Then icon OK is clicked.

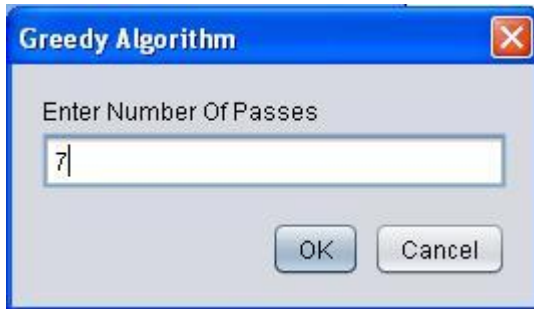


Figure 5.8 Pass Entry Form

As shown in Figure 5.4, all the icons under Select Algorithm column including Tables and Graphs are selected. Analyse icon is then clicked to generate processing results.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------------|------------------|-------------|-------------|-----------|-----------|----------|
| selected | selected | selected | selected | selected | selected | selected |
| b{g,d,h,e} (250) | selected | selected | selected | selected | selected | selected |
| c{g,h,e,f} (250) | c{g,h,e,f} (100) | selected | selected | selected | selected | selected |
| d{g} (160) | d{g} (60) | d{g} (60) | selected | selected | selected | selected |
| e{g,h} (210) | e{g,h} (60) | e{g,h} (60) | e{g,h} (40) | selected | selected | selected |
| f{h} (120) | f{h} (70) | f{h} (20) | f{h} (20) | f{h} (10) | f{h} (10) | selected |
| g{} (90) | g{} (40) | g{} (40) | g{} (10) | g{} (10) | g{} (10) | g{} (10) |
| h{} (90) | h{} (40) | h{} (40) | h{} (40) | h{} (20) | selected | selected |

Figure 5.9 Greedy Algorithm Pass by Pass Output

Figure 5.9 shows pass by pass output following greedy approach. A total of 7 passes are entered. One node is selected in each pass. Node a is already selected prior to pass 1. Each node is associated with the cost benefits and all the nodes covered by it. There are two nodes b,c with same maximum benefit of 250 in pass 1. The greedy algorithm makes a choice of nodes b, c, d, e, h, f and g for each subsequent pass.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------------|------------------|-----------------|-------------|-------------|-----------|-----------|
| selected | selected | selected | selected | selected | selected | selected |
| b{g,d,h,e} (250) | b{g,d,h,e} (100) | b{g,d,h,e} (50) | selected | selected | selected | selected |
| c{g,h,e,f} (250) | selected | selected | selected | selected | selected | selected |
| d{g} (160) | d{g} (110) | selected | selected | selected | selected | selected |
| e{g,h} (210) | e{g,h} (60) | e{g,h} (40) | e{g,h} (40) | e{g,h} (20) | selected | selected |
| f{h} (120) | f{h} (20) | f{h} (20) | f{h} (20) | f{h} (10) | f{h} (10) | f{h} (10) |
| g{} (90) | g{} (40) | g{} (10) | g{} (10) | g{} (10) | g{} (10) | selected |
| h{} (90) | h{} (40) | h{} (40) | h{} (40) | selected | selected | selected |

Figure 5.10 Refined Greedy Algorithm Pass by Pass Output

Figure 5.10 shows pass by pass output following refined greedy approach. The nodes selected by this approach are in order c, d, b, h, e, g and f. The refined greedy algorithm with forward reference is implemented using the following approach:

Take the example of the lattice shown in Figure 5.3 above. The following two arrays are created namely node and source. The node array defines the nodes of lattice with the respective memory occupied by them. The source array defines the source node for materialization of other node covered by a selected node in a given pass. The value in the source node is the value of memory space of a selected node and all the nodes covered by it in a given pass as shown in Array1.

Array 1

| | | | | | | | |
|--------|-------|-------|-------|-------|-------|-------|-------|
| a(100) | b(50) | c(50) | d(20) | e(30) | f(40) | g(10) | h(10) |
| 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

- Cover of b = b,d,e,g,h
Therefore, benefit due to b = $(100-50) \times 5 = 250$
- Cover of c = c,e,f,g,h
Therefore, benefit due to c = $(100-50) \times 5 = 250$
- Cover of d = d,g
Therefore benefit due to d = $(100-20) \times 2 = 160$
- Cover of e = e,g,h
Therefore benefit due to e = $(100-30) \times 3 = 210$
- Cover of f = f,h
Therefore, benefit due to f = $(100-40) \times 2 = 120$

- Cover of g = g
Therefore, benefit due to g = $(100-10) \times 1 = 90$
- Cover of h = h
Therefore, benefit due to h = $(100-10) \times 1 = 90$

As node b and node c are having same benefits, a choice has to be made between them. In the first pass, suppose node b is selected as shown in Array 2.

Array 2

| a(100) | b(50) | c(50) | d(20) | e(30) | f(40) | g(10) | h(10) |
|--------|-------|-------|-------|-------|-------|-------|-------|
| 100 | 50 | 100 | 50 | 50 | 100 | 50 | 50 |

- Cover of c = c,e,f,g,h
Therefore, benefit due to c = $(100-50) + (100-50) = 100$
- Cover of d = d,g
Therefore benefit due to d = $(50-20) + (50-20) = 60$
- Cover of e = e,g,h
Therefore benefit due to e = $(50-30) + (50-30) + (50-30) = 60$
- Cover of f = f,h
Therefore, benefit due to f = $(100-40) + (50-40) = 70$
- Cover of g = g
Therefore, benefit due to g = $(50-10) = 40$
- Cover of h = h
Therefore, benefit due to h = $(50-10) = 40$

The maximum cost benefit of pass 1 is 100. In the first pass, suppose node c is chosen as shown in Array 3.

Array 3

| a(100) | b(50) | c(50) | d(20) | e(30) | f(40) | g(10) | h(10) |
|--------|-------|-------|-------|-------|-------|-------|-------|
| 100 | 100 | 50 | 100 | 50 | 50 | 50 | 50 |

- Cover of b = b,d,e,g,h
Therefore, benefit due to b = $(100-50) + (100-50) = 100$

- Cover of d = d,g
Therefore benefit due to d = $(100-20)+(50-20) = 110$
- Cover of e = e,g,h
Therefore benefit due to e = $(50-30)+(50-30)+(50-30)=60$
- Cover of f = f,h
Therefore, benefit due to f = $(50-40)+(50-40)=20$
- Cover of g = g
Therefore, benefit due to g = $(50-10)=40$
- Cover of h = h
Therefore, benefit due to h = $(50-10)=40$

The maximum cost benefit for pass 1 is 110. On choosing b, a benefit of 100 is obtained in the next pass whereas on choosing c, a benefit of 110 is obtained on the next pass. Therefore, node c is selected for materialization. Select node d in pass 2 and update the node sources from materialized nodes c,d.

Pass 3

| a(100) | b(50) | c(50) | d(20) | e(30) | f(40) | g(10) | h(10) |
|--------|-------|-------|-------|-------|-------|-------|-------|
| 100 | 100 | 50 | 20 | 50 | 50 | 20 | 50 |

- Cover of node b = b,d,e,g,h
Therefore, benefit due to b = $(100-50) = 50$
- Cover of node e = e,g,h
Therefore, benefit due to e = $(50-30) = 20$
- Cover of f = f,h
Therefore, benefit due to f = $(50-40)+(50-40) = 20$
- Cover of g = g
Therefore, benefit due to g = $(20-10) = 10$
- Cover of h = h
Therefore, benefit due to h = $(50-10) = 40$

In the same way, the node selection can be made for next subsequent passes.

5.6 COMPARISON AND ANALYSIS

This section provides the analysis and comparison of results for refined greedy and greedy approach. Lattice shown in Figure 5.3 is taken as input for analysis and comparison of results. The output corresponding to icon Table of Benefits as shown in Figure 5.4 is generated in the form of Table 5.2.

Table 5.2 Comparison of Benefits

| Greedy Algorithm | benefit | Greedy with Forward Reference | benefit |
|------------------|---------|-------------------------------|---------|
| a | 0 | a | 0 |
| b | 250 | c | 250 |
| c | 100 | d | 110 |
| d | 60 | b | 50 |
| e | 40 | h | 40 |
| h | 20 | e | 20 |
| f | 10 | g | 10 |
| g | 10 | f | 10 |

Table 5.2 shows the nodes selected in each pass along with their respective benefits for both greedy algorithm and refined greedy algorithm. The cumulative benefits at the end of 7 passes are same for both the approaches. The cumulative benefits at the end of each pass for refined greedy algorithm are same or better than greedy approach. The same is shown by comparison Table 5.3.

Table 5.3 Comparison of Cumulative Benefits

| Passes | Cumulative Benefits for Greedy Algorithm | Cumulative Benefits for Refined Greedy Algorithm |
|--------|--|--|
| 1 | 250 | 250 |
| 2 | 350 | 360 |
| 3 | 410 | 410 |
| 4 | 450 | 450 |
| 5 | 470 | 470 |
| 6 | 480 | 480 |
| 7 | 480 | 480 |

Table 5.3 shows the cumulative benefits for pass 2 are greater for refined greedy than greedy algorithm. For all other passes cumulative benefits are same for both the approaches. The same results can be visualised with the help of graphs corresponding

to icon Graph in Figure 5.4. The graph corresponding to Figure 5.11 shows Cumulative benefits along Y axis and Number of passes along X axis.

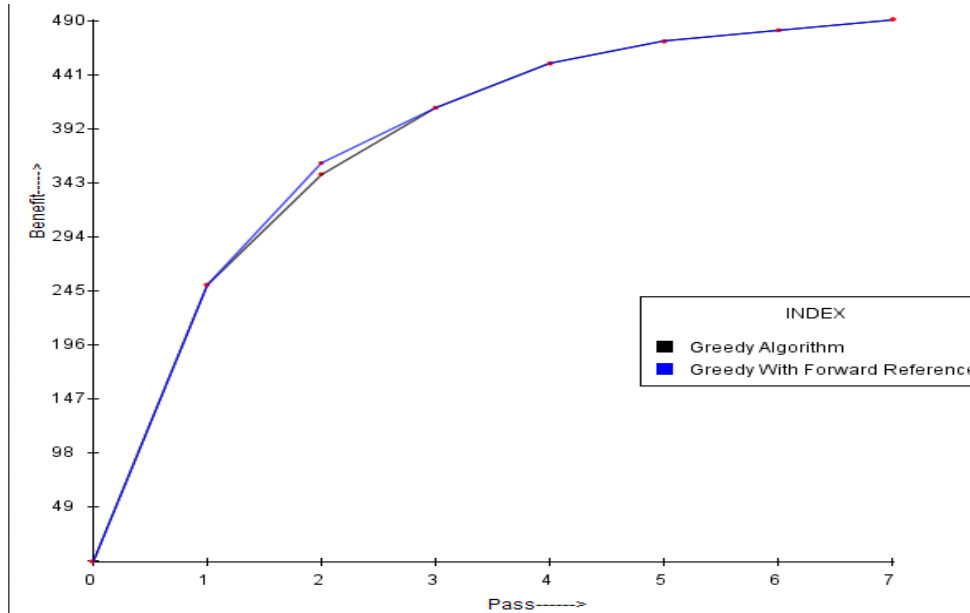


Figure 5.11 Benefits vs. Pass Graph

Another measurement parameter, the Cumulative Space occupied by selected nodes in each pass, is shown by Table 5.4. It can be seen from Table 5.4 that the cumulative space occupied by nodes in each pass is same or lesser for refined greedy algorithm than greedy algorithm. The Cumulative space occupied by nodes in pass 2, 4 and 6 is less for refined greedy approach.

Table 5.4 Comparison of Cumulative Space

| Passes | Cumulative Space for Greedy Algorithm | Cumulative Space for Refined Greedy Algorithm |
|--------|---------------------------------------|---|
| 1 | 150 | 150 |
| 2 | 200 | 170 |
| 3 | 220 | 220 |
| 4 | 250 | 230 |
| 5 | 260 | 260 |
| 6 | 300 | 270 |
| 7 | 310 | 310 |

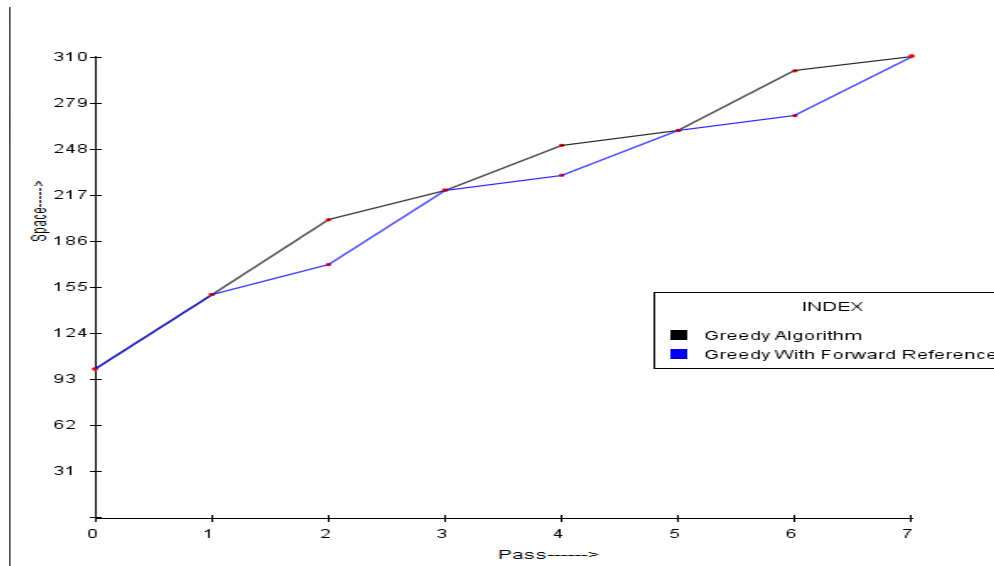


Figure 5.12 Space vs. Pass Graph

The snapshot shown in Figure 5.12 is graphical visualization of Table 5.4. From the above experimental analysis it can be seen that refined greedy approach always perform better than greedy approach when analysed in terms of benefits or space occupied.

The refined greedy and greedy approaches are independent of any space constraints and generate results in accordance to the number of passes specified by the user. For a limited space the optimal selection of materialized views, giving maximum benefits in limited space, is not possible with greedy/refined greedy approach. To achieve optimal view selection with limited space, a knapsack implementation is presented in which the view selection using greedy/refined greedy approach is combined with space constraints. The knapsack approach is given in detail in the next section.

5.7 KNAPSACK IMPLEMENTATION FOR THE RESULTS OF GREEDY AND REFINED GREEDY ALGORITHM

The Greedy approach discussed above is free of any space constraints. The results obtained are in accordance to the number of passes specified by the user. There may arise a situation in which for a limited space optimal materialized view selection is to be made. The greedy/refined greedy approach works only when there are no space constraints. To make materialized view selection with limited space constraints possible, the situation is correlated with knapsack problem. A Knapsack problem

[117] can be stated as: Given a set of items, each with a weight and a value, determine the number of each item to include in a collection so that the total weight is less than or equal to a given limit and the total value is as large as possible.

The set of items in the problem can be correlated with the nodes of lattice, weight with the space occupied by nodes and value with the benefits of nodes as calculated using greedy and refined greedy approach. The knapsack problem in relation to view selection is to select nodes in such a way that for a specified limited space the selected nodes gives maximum space benefits. The icons under Knapsack in Figure 5.4 are selected to generate the form shown in Figure 5.13. The values of space occupied by nodes are taken from lattice [118, 119]. The values of benefits for nodes are as calculated from greedy and refined greedy approach. The user is required to input space constraint and number of iterations for Knapsack implementation. The number of iterations (n) divides the space constraint into (n) equal intervals. The knapsack results are obtained for each interval of the constrained space.

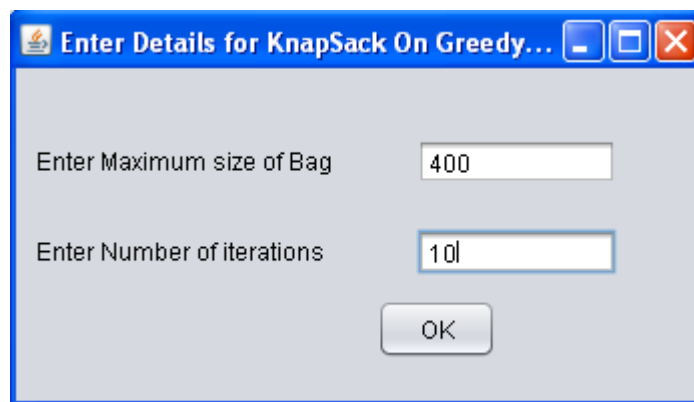
A screenshot of a Windows-style dialog box titled "Enter Details for KnapSack On Greedy...". The dialog box has a blue title bar with standard minimize, maximize, and close buttons. The main area is light gray and contains two text input fields. The first field is labeled "Enter Maximum size of Bag" and contains the number "400". The second field is labeled "Enter Number of iterations" and contains the number "10". Below the input fields is a single "OK" button.

Figure 5.13 Space Constraint Form

By clicking the OK icon, Figure 5.14 is generated, that shows the knapsack selection of items for each interval of constrained space.

| Size Of Bag | Selected nodes | Greedy Algorithm | Size Of Bag | Selected nodes | Greedy with Forward Refe... |
|-------------|----------------|------------------|-------------|----------------|-----------------------------|
| 130 | adh | 80 | 130 | adh | 150 |
| 160 | abh | 270 | 160 | ach | 290 |
| 190 | abdhg | 340 | 190 | acdhg | 410 |
| 220 | abcd | 410 | 220 | acdheg | 430 |
| 250 | abcde | 450 | 250 | acdbhg | 460 |
| 280 | abcdehg | 480 | 280 | acdbheg | 480 |
| 310 | abcdehfg | 490 | 310 | acdbhegf | 490 |
| 340 | abcdehfg | 490 | 340 | acdbhegf | 490 |
| 370 | abcdehfg | 490 | 370 | acdbhegf | 490 |
| 400 | abcdehfg | 490 | 400 | acdbhegf | 490 |

Figure 5.14 Knapsack Benefits for Greedy and Refined Greedy Approach

The size of bag in Figure 5.14 is the constrained space limit. For example enter maximum bag size as 400. The top view a is always selected. The remaining space of 300 is divided into 10 equal parts corresponding to 10 iterations specified. The space of 30 is added to 100 (size of a) to get the starting bag size 130. Increments of 30 are made for subsequent bag size up to limit of 400 as specified. The interpretation of row 1 in Figure 5.14 is that for a limited space of size 130 knapsack algorithm selects 3 nodes a, d, h with total cumulative benefit of 80 as calculated using greedy approach and for a limited space of size 130 knapsack algorithm selects 3 nodes a, d, h with total cumulative benefit of 150 as calculated using refined greedy approach. The fourth row shows that for a constraint space of 220 just 4 items nodes a, b, c, d are selected having cumulative benefits of 410 (knapsack greedy approach) while 6 items a, c, d, h, e, g are selected having cumulative benefits of 430 (knapsack refined greedy approach). The results for knapsack are thus better for refined greedy inputs as compared to greedy inputs.

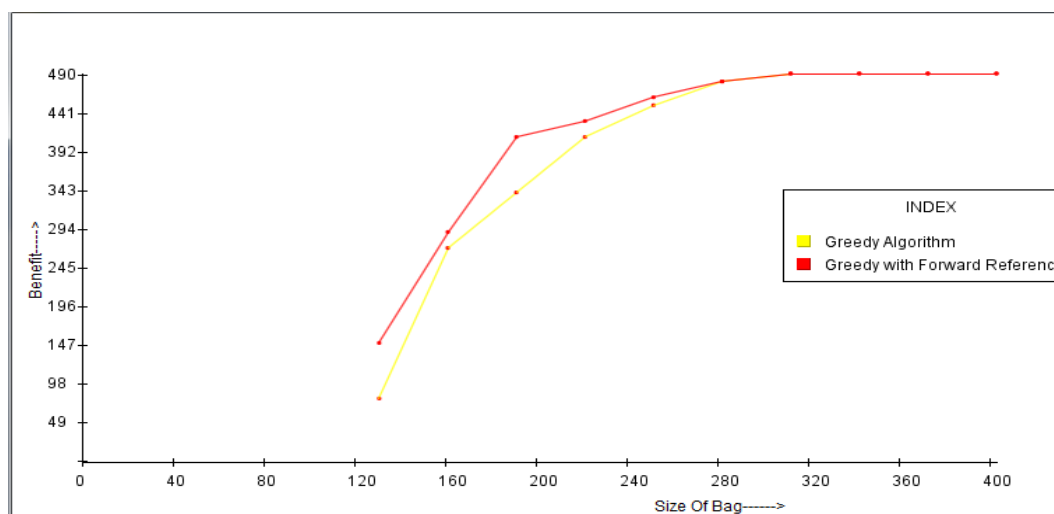


Figure 5.15 Benefits vs. Size Graph

As can be seen from Figure 5.15 the knapsack selection for cumulative benefits of nodes calculated using refined greedy approach is better than knapsack selection for cumulative benefits of nodes calculated using greedy approach up to 5 iterations and thereafter remains same. At the 6th iteration the knapsack selection for both greedy and refined greedy is same meaning that same seven items are selected. The order of selection is a, b, c, d, e, h, g for greedy knapsack selection while the order of selection is a, c, d, b, h, e, g for refined greedy knapsack selection. Only one view f is left. Once all the views are selected, the cumulative benefits remain same thereafter. The maximum difference in benefits is shown in third iteration for a space constraint of 190. The line graph for knapsack refined greedy is always upper than knapsack greedy showing the higher cumulative benefits for knapsack refined greedy selection.

Another analytic perspective showing the number of nodes selected for a given space constraint is shown by Figure 5.16.

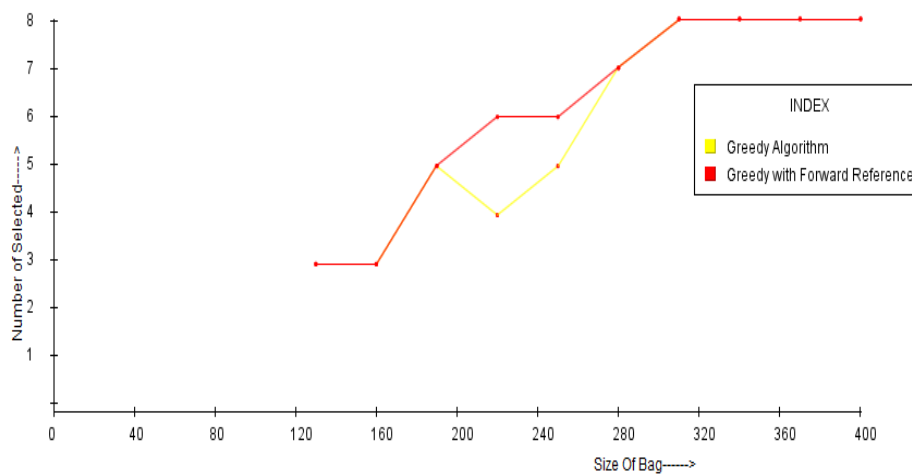


Figure 5.16

Number of nodes selected vs. Size Graph

The selected nodes are same up to third iteration and after fifth iteration. For fourth and fifth iteration knapsack selection refined greedy approach gives better results. The line graphs of Figure 5.16 show the number of nodes selected using knapsack refined greedy approach is same or greater than number of nodes selected using knapsack greedy approach. The above analysis shows that knapsack refined greedy selection gives better results in terms of number of selected nodes and cumulative benefits as

compared to knapsack greedy approach.

5.8 SUMMARY

This chapter presented the research work related to efficient information extraction towards building of an efficient information delivery system. A base approach based on greedy selection of views was taken as reference and some additional enhancements have been made to existing approach. The new proposed approach has been implemented and results calculated in accordance to new proposed approach. Result analysis shows the better selection of views than existing base greedy approach for efficient information extraction towards building of an efficient information delivery system. The proposed approach can be further enhanced by including certain factors such as query pattern frequencies, view pattern frequencies which are directions for future research.

CHAPTER VI

CONCLUSION AND FUTURE RESEARCH SCOPE

6.1 CONCLUSION

The research work carried towards building of an efficient information delivery system has been discussed and presented in previous chapters. Two aspects, one dealing with improving the quality of data warehouse and other dealing with efficient information extraction were identified towards building of an efficient information delivery system. The research work started with the study of existing literature related to both the identified aspects of efficient delivery system. From the study of literature, a classification framework showing various phase of data warehouse development and the techniques used in each of the phases were presented and discussed. The study of literature show that the current research trend focuses on quality evaluation conceptual model and information extraction phase of data warehouse development. So the trend has been given due importance and current research work is oriented toward the quality evaluation and information extraction phase of data warehouse.

During the literature study, it was found that quality evaluation can be carried out at any one conceptual, logical and physical phases of data warehouse design. The present research work has focused on quality issues at conceptual level of design as it is the base level of data warehouse design; the quality of this phase affects the quality of the following phases. The quality at the conceptual level depends on certain factors and measured using certain quality metrics. Various quality metrics for several design configurations, proposed by researchers, were studied. Our current research study is concentrated on object oriented conceptual model metrics, as it incorporates several properties (specialization, generalization, inheritance and polymorphism) that have not been considered by other conceptual design techniques. A new quality metric has been proposed. The metric has been theoretically validated to prove its relevance and utility in the design of conceptual model. The theoretical validation was followed by empirical validation carried out using a controlled experiment in which 22 conceptual models were used and 80 subjects participated for a total of 13 quality metrics.

Various empirical validation techniques like correlation, regression, principal component analysis, case based reasoning were used. The results of empirical validation proved that several metrics including the one proposed had significant effect towards quality evaluation of conceptual data warehouse models.

Further study prompted the thought that if by some mechanism, the quality metrics could be ranked, than it would prove to be a major contribution in design of models at conceptual level. The more important metrics can be stressed upon and taken in consideration during the design of conceptual models. Various ranking methodologies were studied. Three parameters namely understandability, efficiency and effectiveness were identified for ranking of quality metrics. Therefore, a multi-criteria ranking problem has been framed. The opinion of experts was taken to rank the metrics along multiple criteria. As the expert opinion was a result of human thought process based on experience, it was overlapping and ambiguous. To deal with this problem of ambiguity in the opinion of experts, a fuzzy methodology was used for ranking of quality metrics. The results of fuzzy based ranking approach were also compared with actual aggregation ranking based on experts' opinion. The results of fuzzy based ranking approach were better than aggregation ranking approach.

The ranking of quality metrics was further used to develop the fuzzy inference based system so that we predict the understanding time of a conceptual model automatically. For this we prepared a fuzzy rule base based on metric ranking and expert opinions. The final understanding time can be calculated by entering the corresponding values of the quality metrics of the model. The predicted results were compared with the actual results compiled using the data collected from 80 subjects for 22 models in respect of 13 quality metrics in a controlled experiment conducted by us. The results for 19 models out of 22 models were predicted correctly. Thus the fuzzy rule base system can be used for predicting the understanding time of conceptual models without human involvement which needs preparing questionnaire, identifying subjects, conducting survey, collecting and further analysing the data collected.

The research work was carried further in information extraction phase of data warehouse development as one of the identified domain for conducting current research. As we know queries are thrown on big data warehouses for information

extraction used in strategic decision making. The response time to queries is one of the main issues in efficient information extraction from such large databases. In general a powerful query optimization technique selects few and not all views for materialization. Various such query optimization techniques were studied during literature survey including greedy approach. The greedy approach for view selection was taken up as the base approach for its further enhancement. It is enhanced to an approach called refined greedy selection approach which uses forward references to give better selection of views which in turn is responsible for better efficiency as proved by the experimental results. The view selection was further enhanced by including space constraints to the results of greedy and refined greedy approach using knapsack implementation. A comparison of results with the existing knapsack greedy approach show the better materialized view selection in case of proposed scheme.

The objectives of building of efficient information delivery system has been achieved as summarised above. During conducting this research work, we have got certain insights for future extensions in this domain as there is always a scope of improvement and enhancement of any research work performed. The extensible nature of data warehouse domain provides multiple research domains across which research work can be carried on further. Few of the research domains for extension of the current proposed work are following.

6.2 FUTURE SCOPE

The existing work was categorised year wise for better understanding of the developments in the said field. The aim was to give a research summary on data warehouse development approaches. Although this review cannot claim to be exhaustive, it does provide reasonable insight into the subject. Majority of reviewed articles relate to design and evaluation at conceptual level. The conceptual level is given utmost importance due to the fact that it is very costlier to detect and remove errors/bugs in the later logical/physical design phases. The improvement in design and quality at the conceptual level ensures the building of an efficient data warehouse system built up at logical and physical levels. Future research could be aimed on improvement in design and evaluation at logical and actual physical levels as well.

6.2.1 Iterative Approach for Identification of New Metrics

A new quality metric has been proposed for quality evaluation of data warehouse conceptual models. Depending on the application of quality metric in real projects and subsequent performance still new metrics can be evolved and redundant metrics can be discarded following an iterative approach, used for current research work, for data warehouse development.

6.2.2 Wide Sampling for Generalization

Empirical studies were performed to study the effects of quality metrics on understandability of conceptual multidimensional models. Eighty subjects participated in the experiment to perform tasks related to 22 models. The various statistical techniques of correlation, regression, PCA, Nearest Neighbour Analysis and ROC classification were used to analyse the effects of metrics on understandability of models.

The future work can be focused on conducting experiments with more number of models, new questionnaires and subjects for analysing the effects of metrics on understanding time of models. These tasks aim towards generalization of results.

6.2.3 Application to Real Projects

After conducting the theoretical and empirical validation of proposed metric, the proposed metric can be applied to real world projects to judge its performance in real world applications and prove the importance of metric in design of an efficient data warehouse system. Though this future proposal needs full implementation of the information delivery system in which not only conceptual design is there but also other design phases as well.

6.2.4 Extending Domain of Experts, Criteria and Ranking Methodologies

A fuzzy based ranking methodology was proposed to rank quality metrics of conceptual data warehouse models along criteria of understandability, efficiency and effectiveness. The opinion of experts was taken in terms of fuzzy linguistic variables to assign weights to criteria and ratings to metrics. A criteria matrix of ratings and

rankings was formed for each of the metrics. The permanent of criteria matrix was calculated to rank the metrics. A comparison was also made with other methodology to validate the results of calculation. The results of fuzzy based approach were more reliable, accurate as compared to expert opinion approach due to the consideration of ambiguity, imprecision prevalent in human thought process and consideration of all interdependencies of attributes by the use of permanent function. The proposed work can be further extended by discovery of more criteria, metrics for quality evaluation and then applying the proposed fuzzy methodology. Also more number of experts from diverse domains and having wide experience can be consulted for generalization and validation of results which could not be done in this work due to constraints of limited time and resources. Further, a broad comparison can be made with other ranking methodologies to measure the accuracy of results obtained using the proposed fuzzy based approach.

6.2.5 Factor Expansion for Query Optimization

A refined greedy approach for materialized view selection to enhance query optimization has been proposed and implemented. The refined greedy approach makes use of forward referencing to select views of lattice and can be used at the backend of decision support systems to provide near to optimal results. The second approach for materialized view selection is implemented using knapsack refined greedy approach, where a limited space is available for storage of materialized views. A comparison of results with the existing knapsack greedy approach shows the better materialized view selection. The future work in the field of view selection and query optimization can be carried out along various domains. Query evaluation is a complex task and to make its processing faster, along with materializing certain views more factors like average frequency of access of a view, average query response time of a view in addition to the cost benefits can be used for calculation of view benefits. This will go a long way in getting more realistic results.

REFERENCES

- [1] W. Inmon, *Building the Data Warehouse*. John Wiley, ISBN: 0-471-08130-2, 2010.
- [2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, ISBN: 155860-489-8, 2012.
- [3] T. Connolly and C. Begg, *Database Systems*. Addison Wesley, ISBN: 0-201-70857-4, 2012.
- [4] P. Pooniah, P., *Data Warehousing Fundamentals*. Wiley Publication, ISBN: 8126509198, 2010.
- [5] E. Franconi and U. Sattler, "A data warehouse conceptual data model for multidimensional aggregation," in *Proc. of the Workshop on Design and Management of Data Warehouses (DMDW-99)*, 1999.
- [6] N. Gamal, G. Galal-Edeen and E. Bastawissy, "Towards a generic conceptual model for data warehouses," in *Proceedings of 5th international business information management association (IBIMA)*, Egypt, 2005..
- [7] M. Golfarelli, D. Maio and S. Rizzi, "The dimensional fact model: a conceptual model for data warehouses," in *IJCIS*, Vol. 7, Issue 2, pp. 215-247, 1998.
- [8] A. Kamble, "A Conceptual Model for Multidimensional Data," in *Fifth Asia-Pacific Conference on Conceptual Modelling (APCCM 2008)*, Wollongong, NSW, Australia, 2008.
- [9] M. Golfarelli, S. Stefano Rizzi, S. and E. Turricchia, "Modern Software Engineering Methodologies Meet Data Warehouse Design: 4WD," in *Data Warehousing and Knowledge Discovery*, LNCS 6862, pp. 66-79, 2011.
- [10] D. Mishra, A. Yazici and B. Basaran, "A case study of data models in data warehousing," in *ICADIWT*, Vol. 9, pp. 314-319, 2008.
- [11] M. Serrano, C. Calero, H. Sahraoui and M. Piattini, M., "Empirical studies to assess the understandability of data warehouse schemas using structural metrics," in *Software Quality Journal*, Vol. 16, Issue. 01, pp. 79-106, 2008.

- [12] L. Bradji and M. Boufaida, M., “Knowledge Based Data Cleaning for Data Warehouse Quality,” in *Digital Information Processing and Communications in Computer and Information Science*, Vol. 189, pp. 373-384, 2011.
- [13] Y. Singh, A. Kaur and R. Malhotra, “Empirical validation of object oriented metrics for predicting fault proneness models,” in *Software Quality Journal*, Vol. 18, Issue 1, pp. 3-35, 2010.
- [14] C. Calero, M. Piattini, C. Pascual and M. Serrano, “Towards Data warehouse Quality Metrics,” in *International Workshop on Design and Management of Data Warehouses (DMDW’01)*, 2001.
- [15] I. Caballero, A. Vizcaino and M. Piattini, M., “Optimal data quality in project management for global software developments,” in *4th international conference on COINFO*, IEEE, pp. 210-219, 2009.
- [16] M. Serrano, J. Trujillo, C. Calero and M. Piattini, M., “Metrics for data warehouse conceptual models understandability,” in *Information and Software Technology*, Vol. 49, Issue 08, pp. 851–879, 2007.
- [17] G. Poels and G. Dedene, G., “DISTANCE: A Framework Software Measure Construction,” in *Research Report DTEW9937*, Dept. Applied Economics Katholieke Universiteit Leuven, Belgium, 1999.
- [18] L. Briand, S. Morasca and V. Basili, V., “Property based software engineering measurement,” *IEEE transaction on Software Engineering*, Vol. 22 No.1, pp. 68-86, 1996.
- [19] H. Zuse, H., “A framework of software measurement,” in *Walter de Gruyter*, Berlin, 1998.
- [20] R. Garg, K. Sharma, C.K. Nagpal, R. Garg, R. Kumar and Sandhya, “Ranking of software engineering metrics by fuzzy based matrix methodology,” in *Software testing, verification and reliability*, Wiley, Vol. 23, pp. 149-168, 2013.
- [21] L. Zadeh, “Fuzzy sets,” in *Information and control*, Vol.8, pp. 338-353, 1965.
- [22] B. Ali and A. Gosain, “Predicting the quality of OOMD model of data warehouse using fuzzy logic technique,” in *IJESAT*, Vol. 2, Issue 4, pp. 1048-54, 2012.

- [23] V. Harinarayan, A. Rajaraman and J. Ulmann, "Implementing Data Cubes efficiently," in *SIGMOD Conference*, pp. 205-16, 1996.
- [24] E. Annoni, F. Ravat, O. Testeand G. Zurfluh, G., "Towards multidimensional requirement design," in *Proceedings of the 8th international conference on Data Warehousing and Knowledge Discovery*, Springer-Verlag , pp. 47-56., 2006.
- [25] R. Hofman, "Behavioral economies in software quality engineering," in *Journal of empirical software engineering*, Vol.16, Issue 02, pp. 278-293, 2011.
- [26] A. EvenandG. Shankarnarayanan, "Utility-driven configuration of data quality in data repositories," in *Int. J. of Information Quality*, Vol. 1, No. 1, pp. 22-40, 2007.
- [27] C. Blanco, J. Trujillo, E. Fernandez-Medina and M. Piattini, M, "Implementing multidimensional security in OLAP tools," in *3rd international conference on ARES*, IEEE, pp. 1248-1253, 2008.
- [28] C. Blanco, D. Guzman, J. Trujillo, E. Fernandez-Medina and M. Piattini, "Applying an MDA based approach to consider security rules in development of secure data warehouse," in *International conference on ARES*, IEEE, pp. 528-533, 2009a.
- [29] C. Blanco, D. Guzman, J. Trujillo, E. Fernandez-Medina and M. Piattini, "Including security rules support in MDA approach for secure data warehouse," in *International conference on ARES*, IEEE, pp. 516-521, 2009b.
- [30] A. Caro, C. Calero, E. Mendes and M. Piattini, M, "A probabilistic approach to web portal data quality evaluation," in *6th international conference on QUATIC*, IEEE, pp. 143-153, 2007.
- [31] K. Pabreja and K. Datta, "A data warehousing and data mining approach for analysis and forecast of cloudburst events using OLAP-based data hypercube," in *International Journal of Data Analysis Techniques and Strategies*, Vol. 4, No. 1, pp. 57-82, 2012.
- [32] M. Pighin and L. Ieronutti, "A statistical and syntactical approach to data warehouse design quality," in *Int. J. of Information Quality*, Vol. 1, No. 04, pp. 368-391, 2007.

- [33] M. Haigh, "Software Quality, non-functional software requirements and IT-business alignment," in *Software Quality Journal*, Vol. 18, Issue. 03, pp. 361-385, 2010.
- [34] R. Villarroel, E. Soler, J. Trujillo, E. Medina and M. Piattini, M, "Representing levels of abstraction to facilitate the secure multidimensional modeling," in *1st international conference on ARES*, IEEE, pp. 678-684, 2006.
- [35] A. Rodriguez, E. Medina and M. Piattini, "Security requirement with a UML 2.0 profile," in *1st international conference on ARES*, IEEE, pp. 1-8, 2006.
- [36] E. Solar, V. Stefanov, J. Mazon, J. Trujillo, E. Medina E. and M. Piattini, "Towards comprehensive requirement analysis for data warehouses: considering security requirements," in *3rd international conference on ARES*, IEEE, pp. 104-111, 2008.
- [37] E. Verbo, I. Caballero, R. Perez, C. Calero and M. Piattini, "An approach based on i^* for security requirement analysis in data warehouses," in *LAT*, IEEE, Vol. 6, Issue 3, pp. 282-289, 2007.
- [38] E. Duggan and C. Thachenkary, "Integrating nominal group technique for improved software requirement determination," in *Information and management journal*, Vol. 41, pp. 399-411, 2004.
- [39] L. Munoz, J. Mazon and J. Trujillo, "A family of experiments to validate measures for UML activity diagrams of ETL processes in Data Warehouse," in *Information and Software Technology*, Vol. 52, Issue 11, pp. 1188-1203, 2010.
- [40] J. Norberto, J. Lechtenborger and J. Trujillo, "A survey on summarizability issues in multidimensional modeling," in *Data & Knowledge Engineering*, Vol. 68, Issue 12, pp. 1452-1469, 2009.
- [41] F. Tria, E. Lefons, E. and F. Tangora, "Hybrid Methodology for Data Warehouse Conceptual Design by UML Schemas," in *Information and software technology*, Vol. 54, Issue 04, pp. 360-379, 2012.
- [42] M. Rifaie, K. Kianmehr, R. Alhajj and M. Ridley, "Data modelling for effective data warehouse architecture and design," in *Int. J. of Information and Decision Sciences*, Vol. 1, No. 3, pp. 282 – 300, 2009.

- [43] M. Golfarelli and S. Rizzi, "Data warehouse testing: A prototype-based methodology," in *Information and Software Technology*, Vol. 53, Issue 11, pp. 1183–1198, 2011.
- [44] L. Munoz, J. Mazon and J. Trujillo, "Measures for ETL processes models in data warehouses," in *Proceeding of the first international workshop on Model driven service engineering and data quality and security*, ACM, pp. 33-36, 2009.
- [45] C. Zhang, X. Wang and Z. Peng, "Extracting dimensions for OLAP on multidimensional text databases," in *Proceedings of the 2011 international conference on Web information systems and mining, Volume Part II*, Springer-Verlag, pp. 19-26, 2011.
- [46] M. Hendawi and S. Sappagh, "EMD: entity mapping diagram for automated extraction, transformation, and loading processes in data warehousing," in *Int. J. of Intelligent Information and Database Systems*, Vol. 6, No. 3, pp. 255 – 272, 2012.
- [47] A. Cuzzocrea, "Improving range-sum query evaluation on data cubes via polynomial approximation," in *Data & Knowledge Engineering*, Vol. 56, Issue 2, pp. 85-121, 2006.
- [48] A. Simitsis and P. Vassiliadis, "A method for the mapping of conceptual designs to logical blueprints for ETL processes," in *Decision Support System*, Vol. 45, Issue 1, pp. 22-40, 2008.
- [49] E. Medina, J. Trujillo, R. Villarroel and M. Piattini, "Developing secure data warehouses with a UML extension," in *Information Systems*, Vol. 32, Issue 6, pp. 826-856, 2007.
- [50] M. Genero, G. Poels and M. Piattini, "Defining and validating metrics for assessing the understandability of entity–relationship diagrams," in *Data & Knowledge Engineering*, Vol. 64, Issue 03, pp. 534–557, 2008.
- [51] R. Villarroel, J. Trujillo, E. Medina and M. Piattini, "A UML profile for designing secure data warehouses," in *LAT, IEEE*, Vol. 3, Issue 1, pp. 40-48, 2005.
- [52] D. Schuff, K. Corral and O. Turetken, "Comparing the understandability of alternative data warehouse schemas: An empirical study," in *Decision Support Systems*, Vol. 52, Issue 01, pp. 9–20, 2011.

- [53] K. Ramamurthy, A. Sen and A. Sinha, "An empirical investigation of the key determinants of data warehouse adoption," in *Decision Support Systems*, Vol. 44, Issue 4, pp. 817-841, 2008.
- [54] C. Batini, C. Cappiello, C. Francalanci and A. Maurino, "Methodologies for data quality assessment and improvement," in *Computing Surveys*, Vol. 41, Issue 3, pp. 1-52, 2009.
- [55] D. Moody, "Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions," in *Data & Knowledge Engineering*, Vol. 55, Issue 3, pp. 243-276, 2005.
- [56] S. Khurram and G. Mustafa, "Virtual data warehouse: implementation and experimental comparison," in *Int. J. of Management and Decision Making*, Vol. 11, No.1, pp. 69 – 88, 2010.
- [57] S. Kpodjedo, F. Ricca, F. Galinier, Y. Gueheneuc and G. Antoniol, "Design evolution metrics for defect prediction in object oriented systems," in *Empirical Software Engineering*, Vol. 16, Issue 01, pp. 141-175, 2011.
- [58] E. Verbo, I. Caballero, R. Perez, C. Calero and M. Piattini, "A Methodology based on ISO/IEC 15939 to elaborate data quality measurement plans," in *LAT, IEEE*, Vol. 7, Issue 3, pp. 361-368, 2009.
- [59] S. Mojaveri, E. Mirzaeian, Z. Bornae and S. Ayat, "New approach in data stream association rule mining based on graph structure," in *Proceedings of the 10th industrial conference on Advances in data mining: applications and theoretical aspects*, Springer Verlag, pp. 1-12, 2010.
- [60] N. Rahman and J. Harding, "Textual data mining for industrial knowledge management and text classification: A business oriented approach," in *Expert Systems with Applications: an International Journal*, Vol. 39, Issue 5, pp. 4729-4739, 2012.
- [61] G. Bhamra, A. Verma and R. Patel, "Agent enriched distributed association rules mining: a review," in *Proceedings of the 7th international conference on Agents and Data Mining Interaction*, Springer Verlag, pp. 30-45, 2011.
- [62] D. Bobby and J. Lee, "A framework for discovering relevant patterns using aggregation and intelligent data mining agents in telematics

- systems,” in *Telematics and Informatics*, Vol. 26, Issue 4, pp. 343-352, 2009.
- [63] S. Nedjar, A. Casali, R. Cicchetti and L. Lakhal, “Reduced representations of Emerging Cubes for OLAP database mining,” in *Int. J. of Business Intelligence and Data Mining*, Vol. 4, No. 3, pp. 267 – 300, 2009.
- [64] E. Ngai and F. Wat, “A literature review and classification of Electronic commerce research,” in *Information & Management*, Vol. 39, Issue 05, pp. 415-429, 2002.
- [65] E. Ngai and A. Gunasekaran, “A review for mobile commerce research and applications,” in *Decision support systems*, Vol. 43, Issue 01, pp. 3-15, 2007.
- [66] D. Mellado, C. Blanco, L. Sanchez and E. Medina, “A systematic review of security requirements engineering,” in *Computer Standards & Interfaces*, Vol. 32, Issue 4, pp. 153-165, 2010.
- [67] A. Bara, V. Diaconita, I. Lungu and M. Velicanu, “Improving performance in integrated DSS with object oriented modeling,” in *WSEAS Transactions on Computers*, World Scientific and Engineering Academy and Society (WSEAS), Vol. 8, Issue 4, pp. 599-609, 2009.
- [68] M. Genero, E. Manso, A. Visaggio, G. Canfora and M. Piattini, “Building measure based prediction models for UML class diagram maintainability,” in *Journal of Empirical Software Engineering*, Vol. 12, pp. 517-549, 2007.
- [69] M. Haider and T. Kumar, “Materialized views selection using size and query frequency,” in *Int. J. of Value Chain Management*, Vol. 5, No. 2, pp. 95-105, 2011.
- [70] R. Rjagan, E. Chang and T. Dillon, “Conceptual design of XML FACT repository for dispersed XML document warehouses and XML marts,” in *5th international conference on CIT*, IEEE, pp. 141-149, 2005.
- [71] L. Ribeiro, R. Goldschmidt and M. Cavalcanti, “Complementing data in the ETL process,” in *Proceedings of the 13th international conference on Data warehousing and knowledge discovery*, Springer Verlag, pp. 112-123, 2011.

- [72] A. Simitsis, D. Skoutas and M. Castellanos, “Natural language reporting for ETL processes,” in *Proceeding of the ACM 11th international workshop on Data warehousing and OLAP*, ACM, pp. 65-72, 2008.
- [73] A. Gosain, S. Sabharwal and S. Nagpal, “Assessment of quality of data warehouse multidimensional model,” in *Int. J. of Information Quality*, Vol. 2, No. 4, pp. 344 – 358, 2011.
- [74] A. Gosain, S. Sabharwal and S. Nagpal, “Predicting quality of data warehouse using fuzzy logic,” in *Int. J. of Business and Systems Research*, Vol. 6, No. 3, pp. 255 – 268, 2012.
- [75] H. Kefi and N. Koppel, “Measuring data warehousing success: an empirical investigation applying the DeLone and McLean model,” in *International Journal of Data Analysis Techniques and Strategies*, Vol. 3, Issue 2, pp. 178-201, 2011.
- [76] A. Smith, “Quality assurance practices for competitive data warehouse management systems,” in *Int. J. of Business Information Systems*, Vol. 7, No. 4, pp. 440 – 457, 2011.
- [77] N. Aggarwal, A. Kumar, H. Khatter and V. Aggarwal, “Analysis of the effect of DM techniques on database,” in *Advances in Engineering Software*, Vol. 47, Issue 01, pp. 164-169, 2012.
- [78] B. Thuraisingham, M. Kantarcioglu and S. Iyer, “Extended RBAC-based design and implementation for a secure data warehouse,” in *Int. J. of Business Intelligence and Data Mining*, Vol. 2, No. 4, pp. 367 – 382, 2007.
- [79] V. Basili, L. Briand and W. Melo, “A validation of object-oriented design metrics as quality indicators,” in *IEEE transactions software engineering*, Vol. 22 No. 10, pp. 751-61, 1996.
- [80] M. Serrano, C. Calero and M. Piattini, “Experimental Validation of Multidimensional Data Model Metrics,” in *36th Hawaii International conference on System Sciences*, 2003.
- [81] F. Shull, J.C. Carver, S. Vegas and N. Juristo, “The Role of Replications in Empirical Software Engineering,” in *Journal of Empirical Software Engineering*, Springer, Vol. 13, pp. 211-218, 2008.

- [82] A. Lucia, Carmine, Gravino, Oliveto and G. Tortora, "An experimental comparison of ER and UML class diagrams for data modeling," in *Journal of Empirical Software Engg.*, Springer, Vol. 15, pp. 455-492, 2010.
- [83] G. Johnson and X. Yu, "Objective software quality assessment," in *Proceedings of nuclear science symposium*, Seattle, USA, pp. 1691-1698, 1999.
- [84] T. Dyba, "An instrument for measuring the key factors of success in software process improvement," in *Empirical software engineering*, Vol. 5, pp. 357-390. DOI: 10.1023/A:1009800404137, 2000.
- [85] L. Briand, B. Freimut and F. Vollei., "Assessing the cost effectiveness of inspections by combining project data and expert opinion," in *Proceedings of 11th international symposium on software reliability engineering*, San Jose, USA, Vol. 23, pp. 246-258, 2000.
- [86] X. Zhang and H. Pham, "An analysis of factors affecting software reliability," in *The journal of systems and software*, Vol. 50, pp. 43-56. DOI: 10.1016/S0164-1212(99)00075-8, 2000.
- [87] L. Ming and S. Carol, "A ranking of software engineering measures based on expert opinion," in *IEEE transactions on software engineering*, Vol. 29, No. 9, pp. 811-824, 2003.
- [88] H. Lau, C. Wong, P. Lau, K. Pun, K. Chin and B. Jiang, "A fuzzy multi criteria decision support procedure for information delivery in extended enterprise networks," in *Engineering applications of artificial intelligence*, Vol.16, pp. 1-9. DOI: 10.1016/S0952-1976(03)00020-4, 2003.
- [89] J. Wang and Y. Lin, "A fuzzy multi criteria group decision making approach to select configuration items for software development," in *Fuzzy sets and systems*, Vol. 134, pp. 343-363. DOI: 10.1016/S0165-0114(02)00283-X, 2003.
- [90] J. Cochran and H. Chen, "Fuzzy multi criteria selection of object oriented simulation software for production system analysis," in *Computers and operations research*, Vol. 32, No. 1, pp. 153-168, 2005.

- [91] R. Garg, V. Gupta and V. Agrawal, "Quality evaluation of thermal power plants by graph theoretical methodology," in *Int. J. of power and energy systems*, Vol. 27, No. 1, pp. 42-48, 2007.
- [92] K. Khatatneh and T. Mustafa, "Software reliability modeling using soft computing technique," in *European J. of scientific research*, Vol. 26, No. 1, pp. 154-160, 2009.
- [93] G. Bailador and G. Trivino, "Pattern recognition using temporal fuzzy automata," in *Fuzzy sets and systems*, Vol. 161, No. 1, pp. 37-55, 2010.
- [94] Rafik A. Aliev and Oleg H. Huseynov, "Fuzzy Geometry-Based Decision Making with Unprecisiated Visual Information," in *Int. J. Info. Tech. Dec. Mak.*, Vol. 13, 1051. DOI: 10.1142/S0219622014500709, 2014.
- [95] Xia Meimei and Xu Zeshui, "A Novel Method for Fuzzy Multi-Criteria Decision Making," in *Int. J. Info. Tech. Dec. Mak.*, Vol.13, 497. DOI: 10.1142/S0219622014500205, 2014.
- [96] A. Kaufmann and M. Gupta, "Fuzzy mathematical models in engineering and management science," in *Elsevier science publisher*, Netherlands, 1998.
- [97] J. Buckley and S. Chanas, "A fast method of ranking alternative using fuzzy numbers," in *Fuzzy sets and systems*, Vol. 30, pp. 337-338, 1989.
- [98] M. Marcus and H. Minc. Permanents. *American mathematics*, Vol. 72, pp.571-591, 1965.
- [99] A. Aldea, R. Alcantara and S. Skrzypczak, "Managing information to support the decision making process," in *JIKM World scientific publishers*, Vol. 11, No. 3, 2012.
- [100] Z. Bellahsene, M. Cart and N. Kadi, "A cooperative approach to view selection and placement in P2P systems," in *OTM*, pp. 515-22, 2010.
- [101] C. Dhote and M. Ali, "Materialized view selection in Data Warehousing: A Survey," in *Journal of Applied Sciences*, pp. 401-14, 2009.
- [102] A. Halevy, "Answering queries using views: A survey," in *VLDB Journal*, Vol. 10, Issue 4, pp.270-94, 2001.

- [103] A. Shukla, P. Deshpande and J. Naughton, "Materialized view selection for multi-cube data models," in *LNCS 1777*, Springer Verlag, pp. 269-84, 2000.
- [104] J. Nilsson. *Problem solving methods in artificial intelligence*. New York: McGraw–Hill publishing company Ltd, 1971.
- [105] K. Ross, D. Srivastava and S. Sudharshan, "Materialized view maintenance and integrity constraint checking: Trading space for time," in *SIGMOD*, pp. 447-58, 1996.
- [106] I. Miami and Z. Bellahsene, "A survey of view selection methods," in *SIGMOD Record*, Vol. 41, Issue 1, pp. 20-29, 2012.
- [107] S. Lujan–Mora, "Extending UML for Multidimensional Modeling UML '02'," in *Proceedings of the 5th International Conference on the Unified Modeling Language*, London, UK, 2002.
- [108] M. Piattini, M. Genero, M. and L. Jimenez, "Metrics Based Approach for Predicting Conceptual Data Model Maintainability," in *International Journal of Software Engineering and Knowledge Engineering*, Vol. 11, Issue 6, pp. 703-729, 2001.
- [109] www.inmoncif.com
- [110] A. Kaufmann and M. Gupta, "Fuzzy mathematical models in engineering and management science," in *Elsevier science publisher*, Netherlands, 1998.
- [111] C. Huang and T. Lin, "Software reliability analysis by considering fault dependency and debugging time lag," in *IEEE Trans. Reliability*, Vol. 55, No. 3, pp. 436-50, 2006.
- [112] C. Smidts, R. Stoddard and M. Stutzke, "Software reliability models: an approach to early reliability prediction," in *IEEE Trans. Reliability*, Vol. 47, No. 3, pp. 268-78, 1998.
- [113] C. Huang and S. Kuo, "Analysis of incorporating logistic testing effort function into software reliability modelling," in *IEEE Trans. Reliability*, Vol. 51, No. 3, pp. 261-70, 2002.
- [114] A. Sukert, "Empirical validation of three software errors predictions models," in *IEEE Trans. Reliability*, pp. 199-205, 1979.

- [115] K. Pillai and V. Nair, "A model for software development effort and cost estimation," in *IEEE Trans. Software Engg.*, Vol. 23, No. 8, pp. 485-97, 1997.
- [116] M. Lyu and A. Nikora, "Applying software reliability models more effectively," *IEEE Softw.*, pp. 43-52, 1992.
- [117] Corman. *Introduction to Algorithms*. PHI publication; 2008.
- [118] W. Pedryez, "Knowledge management and semantic modeling: a role of information granularity," in *IJSEKE World scientific publishers*, Vol. 23, Issue 1, pp. 5-11, 2013.
- [119] C. Lofi, "Analogy queries in information systems- a new challenge," in *JIKM World scientific publishers*, Vol. 12, Issue 3, 2013.

BRIEF PROFILE OF THE RESEARCH SCHOLAR



Naveen Dahiya received his B.E. (first division with honours) in Computer Science and Engineering from Maharshi Dayanand University, Haryana, India in 2003 and M. Tech. (first division with honours) in Computer Engineering from Maharshi Dayanand University, Haryana, India in 2005 and is currently pursuing Ph.D in Computer Engineering (Faculty of Engg. & Technology) from Y.M.C.A. University of Science and Technology, Faridabad, Haryana, India. He is currently working as an Assistant Professor and Head in Computer Science and Engineering Department at Maharaja Surajmal Institute of Technology, C-4, Janak Puri, New Delhi, India. He has ten years of experience in teaching as an Assistant Professor. He has qualified GATE & NET examinations. His Research interests include database systems, data warehouse and data mining.

LIST OF PUBLICATIONS

List of Published Papers

| S.No. | Title of Paper | Name of Journal/Conference where published | No. | Vol. & Issue | Year | Page |
|-------|---|---|-----|--------------|------|-----------|
| 1. | Modeling data warehouse using quality metrics: The need of software process. (Paper indexed in DOAJ, Google Scholar) | IJCA special issue on Confluence 2012- The next generation information technology summit, CONFLUENCE(1) | | | 2012 | 28-31 |
| 2. | Effective data warehouse for information delivery: a literature survey and classification. (Paper indexed in Scopus, ACM) | International journal of networking and virtual organizations, Inderscience. | 3 | 12 | 2013 | 217 - 237 |
| 3. | Enhancing consistency of conceptual data warehouse design, Vol. 2, No. 1, 2015, pg. 11-24. (Paper indexed in Google Scholar, Gale) | International journal of computational systems engineering, Inderscience | 1 | 2 | 2015 | 11-24 |
| 4. | An empirical experimentation towards predicting understandability of conceptual schemas using quality metric. (Paper indexed in Google Scholar, Gale) | International journal of big data intelligence, Inderscience | 1 | 2 | 2015 | 09-22 |
| 5. | Applications of data mining in software development life cycle: A literature survey and classification (Book chapter), | Book entitled: Data mining and analysis in engineering field, IGI Global, DOI: 10.4018/978-1-4666-6086-1.ch004. | | | 2014 | 67-69 |
| 6. | An Experiment towards metrics validation for data warehouse conceptual models. | 5th International Conference (IEEE), Confluence 2014: The Next Generation Information Technology Summit on the theme: Cloud Security and Big Data | | | 2014 | 116 - 123 |
| 7. | A Conceptual Framework for Effective Data Warehouse Design. | International conference FOBE, IMT Ghaziabad | | | 2012 | 1-14 |

List of Accepted Papers

| S.No. | Title of Paper | Name of Journal/Conference where published | Present Status | Year |
|--------------|--|--|-----------------------|-------------|
| 1. | A fuzzy based matrix methodology for evaluation and ranking of data warehouse conceptual model metrics.(Paper indexed in SCI Expanded , Scopus) | International Arab Journal of Information Technology | Accepted | 2015 |

List of Communicated Papers

| S.No. | Title of Paper | Name of Journal/Conference where published | Present Status | Year |
|--------------|---|---|-----------------------|-------------|
| 1. | Efficient Materialized View Selection for Multi-Dimensional Data Cube Models. | IJIRR, IGI Global | Communicated | 2015 |