ISSN: 2250–3676

[IJESAT] INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY Volume-2, Issue-4, 1048 – 1054

PREDICTING THE QUALITY OF OBJECT-ORIENTED MULTIDIMENSIONAL (OOMD) MODEL OF DATA WAREHOUSE USING FUZZY LOGIC TECHNIQUE

Kunwar Babar Ali, Anjana Gosain

University School of Information Technology (USIT), GGSIPU, New Delhi, India, ***kunwarbabarali1@gmail.com***
University School of Information Technology (USIT), GGSIPU, New Delhi, India, ***anjana_gosain@yahoo.com***

Abstract

Data warehouse is a powerful tool which makes decision faster and reliable in organizations where 'information' is the main asset of primary concern. It is very necessary to assure the information quality of the data warehouse. Information quality depends on multidimensional model quality of data warehouse. In the last few years' different authors have suggested several approaches to access the quality of multidimensional models of data warehouse. However empirical validation of these metrics has been made using statistical technique like correlation analysis, univariate and multivariate regression technique etc. But all these technique are not capable to model non linear relationship between the metrics and the quality of multidimensional model. In this paper we firstly proposed a model based on fuzzy logic technique to model nonlinear relationships between the metrics and the quality of object-oriented multidimensional (OOMD) model. Then empirically evaluate the effectiveness of the proposed model, result of fuzzy based model compared with the result of controlled experiment conducted by us and result shows that the proposed fuzzy logic based model is capable to predict the output with considerable accuracy.

Index Terms: *Data warehouse. Information quality, Object-oriented multidimensional model, fuzzy logic, quality metrics*

---------- *** ----------

1. INTRODUCTION

Data warehouse is the backbone of most decision support system; it provides historical information to the decision makers. A be short of quality in the data warehouse can have disastrous consequences from both technical and organizational points of view: loss of clients, important financial losses or discontent amongst employees [2].One way to assure quality of data ware house is to guarantee the quality of the models (conceptual, logical and physical) used in to design of data warehouse. Quality of data warehouse multidimensional model has a great influence on the overall data warehouse quality and hence, in turn on information quality [5][1]. Few researchers have suggested quality factors for multidimensional model for data warehouse like maintainability, simplicity, completeness, consistency, minimality, etc [1][3][6]. Piattini at el [10] defined a set of metrics of object-oriented conceptual multidimensional model for data ware house and they theoretically validate them using Briand framework.

Fuzzy logic offers substantial advantages above other approaches due to its ability to naturally represent qualitative aspect of data and apply flexible inference rules [14]. Model based on fuzzy logic may approximate any non-linear continuous function based on the given data. Due to its natural ability to model imprecise and fuzzy aspect of data and rules, fuzzy logic is an attractive alternative in situation where approximate reasoning is called for. Fuzzy logic technique has been successfully applied in the field of software engineering to predict defect density, software effort estimation, error prone code modules using historical data. Researches also applied fuzzy logic method to predict the quality of different models like UML Class diagram maintainability, Entity relationship (ER) model understanadability.[14][18]. But till date there is no work connected to prediction of quality of data warehouse OOMD model using fuzzy logic technique.

Therefore in this paper we apply fuzzy logic technique to make a prediction model, to predict the understanadability of OOMD model of data warehouse and we also compare the result of proposed model with the results conducted by us manually.

ISSN: 2250–3676

[IJESAT] INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY Volume-2, Issue-4, 1048 – 1054

2. QUALITY PREDICTION USING FUZZY LOGIC BASED MODEL

During the development of a prediction model for systems' quality, we must first identify factors those have major impact on system's quality [14]. Unfortunately, it is not so easy to accurately identify relevant quality factors and also the degree of influence they have on the system's quality [15][16][17]. For that reason, exact and discrete metric data is used.

A fuzzy logic prediction model proposed by us comprise of following steps:

- Build up a set of membership functions for metrics. MF = { m1,m2….,mn}
- Pick a quality factor, QF. In this work, it is understanadability.
- Produce a rule base to convert MF to QF.

Figure 1 shows the basic configuration of a fuzzy logic system with fuzzifier and defuzzifier. This type of fuzzy logic system was first proposed by Mamdani. The main fours components' functions are shown in figure 1 below

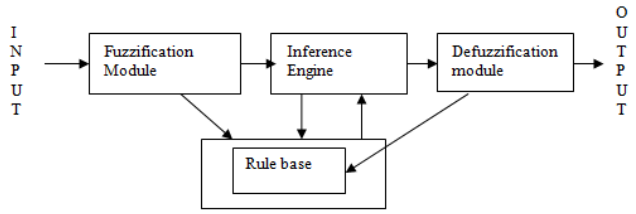

Figure 1: Fuzzy logic system

Fuzzification module: This module does a mapping from input to a fuzzy set.

Fuzzy Rule Base: Fuzzy logic systems use fuzzy IF-THEN rules. In a fuzzy logic system, the collection of fuzzy IF-THEN rules is stored in the fuzzy rule base which is referred to by the inference engine when processing inputs.

Fuzzy Inference Engine: previously all input values have been fuzzified into their respective linguistic values; the inference engine will access the fuzzy rule base of the fuzzy expert system to derive linguistic values for the intermediate as well as the output linguistic variables. The two main steps in the inference process are aggregation and composition. Aggregation is the process of computing for the values of the IF (antecedent) part of the rules while composition is the process of computing for the values of the THEN (consequent) part of the rules.

Defuzzification module: Defuzzifier does a mapping from the fuzzy output to the crisp output.

3 RELEATED WORK

The related work is primarily divided into two parts. The first part deals with the multidimensional modelling and their quality metrics. Second part focus on the work associated with fuzzy logic technique to build a quality prediction models. In past various multidimensional data model have been proposed .Some of them fall into Logical level, some fall into frame model and some fall into conceptual model[02].Object-oriented multidimensional model comes under the conceptual model. Conceptual model provides a set of graphical notations that improve their use and reading. Some conceptual models are The Dimensional-Fact (DF) Model by Golfarelli et al. The Multidimensional ER (M/ER) Model by Sapia et al, The starER Model by Tryfona et al, the Model proposed by Hu¨seman et al., and The Yet Another Multidimensional Model (YAM2) by Abello´ et al.[02] But regrettably none of them has been accepted as a standard for design and maintain high quality data warehouse[02] .Lately a new model is projected that is capable to design efficient data warehouse that is Object-Oriented Multidimensional (OOMD) model [02].N.Prat[02] have proposed metrics for multidimensional schemas analyzability and simplicity Unfortunately none of these metrics proposed have been theoretically as well as empirically validated and therefore, have not proven their practical utility[07] .A proposed metrics has no value if it is not empirically validated then its practical value is zero. Romero et al [15] applied fuzzy logic technique in object – oriented software engineering to build a prediction model. Kwon.Y.R et.al [14] proposed a fuzzy logic based automated prediction system which identifies potential error in early stage of the software life cycle, which help designer to take care of these issues in the starting phase. Mercela Genaro et.al [18] used fuzzy logic technique to predict the maintainability of entity relationship(ER) diagrams. Yaun et al [16] used fuzzy logic technique to predict the no of faults in the system. In this paper we applied fuzzy logic technique to predict the understanadability of OOMDM of data warehouse.

3.1 Metrics for Object -Oriented Multidimensional (OOMD) model

Various authors stated that the complexity of a model may be calculated by the no. and variety of elements and no. and variety of relationship between them. Taking into account this statement we must take three different levels: class, star and diagram. [02] .In this paper we used star level metrics since the star schema is the main issue of data warehouse multidimensional model. Following table shows the metrics [2] for OOMDM of data warehouse.

Metrics	Description
NDC(S)	Number of dimension classes of the star S
NBC(S)	Number of base classes of the star S
NA(S)	Total number of classes of the star S C(S)=NDC(S)+NBC(S)+1
NC(S)	Total number of FA, D and DA attributes of the star S

Table 1: OOMD Model Metrics

To know these metrics, let us take an OOMD model which has all the stereotypes i.e. class stereotypes and attribute stereotypes. Figure 4 have all these things. Table 4 contains the values we have calculated for the metrics given in table 3

Metrics	Calculated values
NDC(S)	3
NBC(S)	5
NA(S)	9
NC(S)	19

Table2: Calculated values for star schema in fig 2

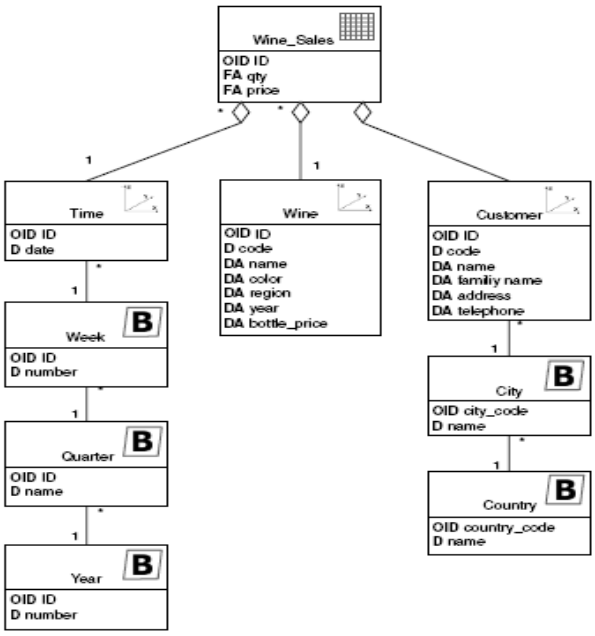

Fig 2.Example of object-oriented multidimensional model

4. EXPERIMENTAL SETUP

In our study, experimental planning includes, subjects considered in this study and data collection etc.

4.1 Data Collection

Eleven object-oriented multidimensional models were collected to carry out this controlled experiment. The domains of these models were general and well known to avoid the problems with domain understanding. The subjects of the experiment are a group of post graduate students at Guru Gobind Singh Indraprastha University, New Delhi. There were twenty participants who participated in the experiment. The students had enough knowledge about Software Engineering, Data modelling and Data Warehouse. Since, the students were post graduate students; few had industrial experience in the field of data warehouse and data bases. The subjects were explained about the experiment. The subjects were supposed to understand and analyse the models and answer few questions. We also explained to them that before studying each schema they had to gloss the starting time (hour, minutes and seconds), then they could look at the design until they were able to answer the given question. Once the answer to the question had been written, they had to annotate the final time (again in hour, minutes and seconds).Example of some questions as follows

Write the starting time (HH:MM:SS)

1). Which classes do you need to use for knowing the color of one wine?

2). Which classes do you need to use for obtaining a list of all the sales of a year?

3) Make the necessary modifications to the model to fit this requirements

4). You need to store information about the month to which a week belongs to

5). You need to store information about the promotions made with the wines

Write the finishing time (HH:MM:SS)

The subjects were divided into two groups of 10 students each, keeping in mind the experience of subjects. Five multidimensional models were given to one group and rest six were given to other group. This was done to reduce the fatigue effect, as all the students in a given group need to complete the experimental task for all the models given to the group. We allowed 2hours time to do the experiment. Table 3 shows the collected data in terms of understanding time.

Subject	Schema no. 1	2	3	4	5	6	7	8	9	10	11
1	255	345	85	255	239	219	169	387	127	68	402
2	359	289	30	420	290	520	348	310	650	78	329
3	180	140	110	140	289	410	249	120	169	29	380
4	200	349	129	210	289	130	360	130	349	87	760
5	130	717	230	120	189	199	430	19	29	70	450
6	190	310	210	250	320	430	190	199	210	87	550
7	30	320	110	210	430	210	310	330	100	130	440
8	340	420	120	110	760	310	840	430	190	98	390
9	320	440	110	210	360	512	320	410	150	110	281
10	420	210	144	430	450	320	110	380	340	30	430
11	120	230	120	160	340	390	210	320	120	76	543
12	180	210	96	160	430	650	320	430	120	120	430
13	190	320	119	140	320	120	190	360	130	144	367
14	430	220	140	190	160	180	420	330	190	60	450
15	330	230	120	230	310	189	314	365	427	11	298
16	711	232	103	544	325	276	387	458	329	90	456
17	331	652	145	231	432	220	560	216	199	27	298
18	320	391	121	238	490	780	490	190	190	110	410
19	330	590	201	220	477	320	381	189	760	101	560
20	310	630	121	421	541	117	377	210	230	45	319

Table 3: Collected Understanding time

Following table shows the descriptive analysis of the collected data

	1	2	3	4	5	6	7	8	9	10	11
Average	302.3	307.1	128.2	244.45	334.05	309.6	348.75	289.15	250.45	83.55	427.5
Minimum	130	140	30	110	160	117	110	19	29	27	81
Maximum	430	717	230	544	760	780	840	430	650	130	760

Table 4: Descriptive statistics

4.2 Experimental design for Model based on Fuzzy Logic

With the help this model we hit upon the effect of OOMD model metrics i.e. NDC, NBC, NC and NA on understanadability of data warehouse.

4.2.1 Fuzzification

We have created following membership functions for the OOMD model metrics

The Number of dimensions classes (NDC) and Number of Base Classes (NBC) is classified in three linguistic variables i.e. low, medium and high as shown in figure 3 & 4 as following

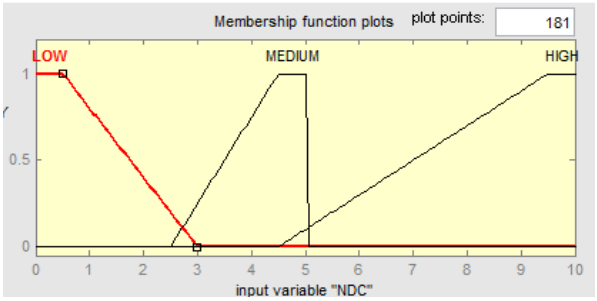

Figure 3: Fuzzification of NDC

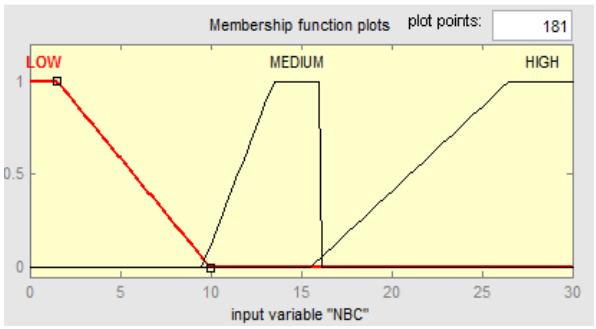

Figure 4: Fuzzification of NBC

The Number of classes (NC) and the Number of attributes (NA) is classified in three linguistic variables i.e. low, medium and high as shown in figure 5 & 6 as below

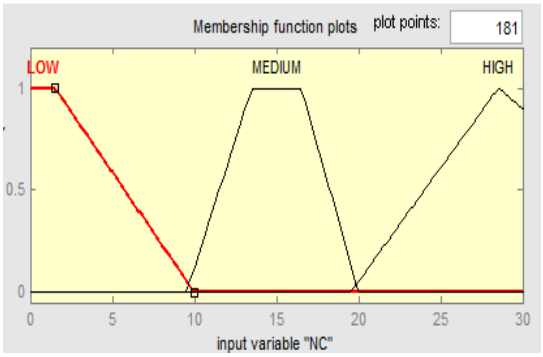

Figure 5: Fuzzification of NC

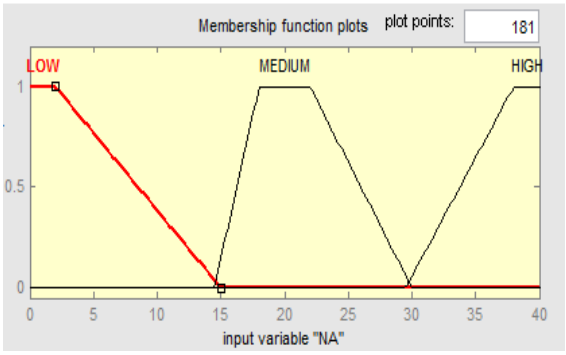

Figure 6: Fuzzification of NA

The output variable i.e. understanadability is classified as very easy, easy, moderate, poor and very poor as shown in following fig.

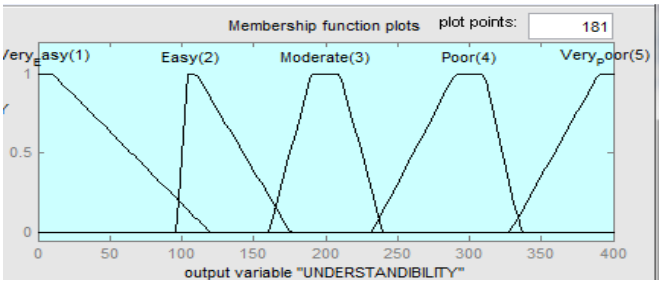

Figure 7: Fuzzification of understanadability

4.2.2 Fuzzy rule base for the model

After the Fuzzification of input data, processing take place in rule base of the fuzzy system. In this step a set of heuristic rules are created which converts the metrics into the quality factors. We have considered all the combinations of inputs during the design of these rules for the system. Total, 81, i.e. 34 (since each of the four metrics have three membership functions) heuristic rules are created for the fuzzy model [14][15][16][17]

The understanadability in case of all the possible combinations is classified with their ratings as discussed above. Following are some heuristic rules we have created for the system

1 If (NDC is low) and (NBC is low) and (NC is low) and (NA is low) then (understanadability is very easy (1)).

2 If (NDC is low) and (NBC is low) and (NC is low) and (NA is medium) then (understanadability is very easy (1)).

.

81 If (NDC is high) and (NBC is high) and (NC is high) and (NA is high) then (Understanadability is very poor (5)).

ISSN: 2250–3676

[IJESAT] INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY Volume-2, Issue-4, 1048 – 1054

4.2.3 Defuzzification: we can perform Defuzzification with the help of Centriod of area (COA), bisector of area, mean of maximum etc. In this paper, we have used COA for Defuzzification

5. RESULTS

Following table (4) shows the predicted output by fuzzy logic model.

Schema No.	NDC	NBC	NC	NA	Predicted Understanding time using fuzzy logic(in seconds	Understanding time (Average) Calculated manually
1	6	16	23	17	295	302
2	5	19	25	32	366	307.1
3	2	05	8	14	56.9	128.2
4	4	17	22	27	285	244.45
5	3	21	25	36	288	334.05
6	5	13	19	34	284	309.6
7	3	6	10	12	136	348.75
8	4	5	10	21	204	289.15
9	3	5	9	19	200	250.45
10	2	4	7	10	51.7	83.55
11	7	24	26	35	371	427.15

Table 5: Comparison of Predicted time

The Output (Understanding time) is predicted as follows: Let us take schema no.1; we have the following inputs to the model. NDC=6, NBC=16, NC= 23 NA =17. When these inputs are fuzzified, we find that NDT=6 belongs to fuzzy set **high**, NBC = 16 belongs to fuzzy set **medium,** NC =23 belongs to fuzzy set **high** and NA =17 belongs to fuzzy set **medium**. With these inputs, rule number **71** is satisfied, which gives understanding time 295(seconds).

These results are observed by using the fuzzy tool box in MATLAB. Using rule viewer, output may observed for the particular set of inputs. Following figs shows the rule viewer of schema no.1.

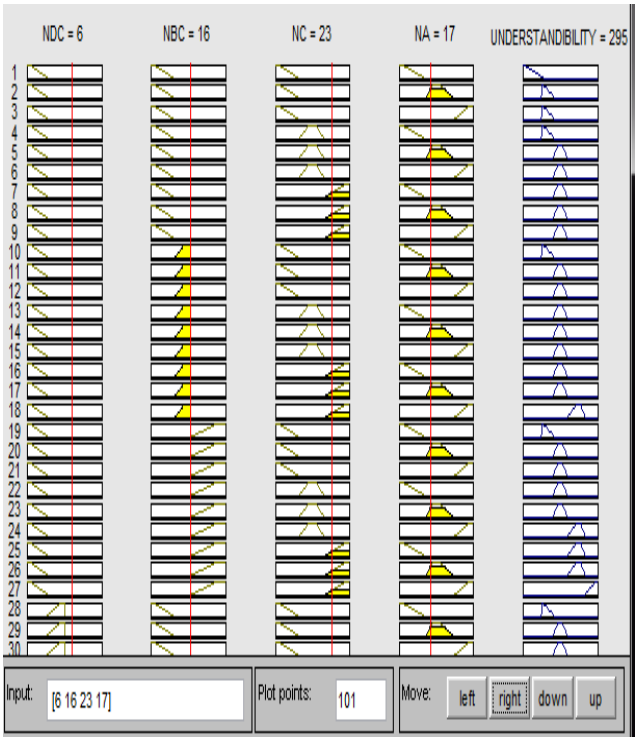

Figure 8: Rule Viewer for schema number 01

6. CONCLUSION AND FUTURE WORK

One way to assure quality of data warehouse is to assurance the quality of the models used to design data warehouse. In this paper, we projected a fuzzy logic based approach to predict understandability of the Object –oriented multidimensional (OOMD) model for data warehouse. This paper presented the effect of structural complexity on the understandability. The planned model is able to successfully measure the understandability based on four metrics i.e. NDC, NBC, NA and NC as inputs. But, usefulness of the proposed model needs to be further validated empirically. The inference rules used in this paper can be improved further. Such improvements would allow this model to perform more correctly.

REFERENCES

[1] Manuel Serrano, Coral Calero, Mario Piattini, "Experimental Validation of Multidimensional Data Models Metrics", Proceedings of the 36th Hawaii International Conference on System Sciences – 2003.

[2] Manuel Serrano , Juan Trujillo, Coral Calero, Mario Piattini , " Metrics for data warehouse conceptual models understanadability", Proceedings of the 36th Hawaii International Conference on System Sciences

[3] Manuel Angel Serrano, Coral Calero, Houari A. Sahraoui, Mario Piattini."Empirical studies to assess
the understanadability of data warehouse schemas using structural metrics", Software Qual J (2008), 6:79–106 DOI 10.1007/s11219-007-9030-7

[4] Daniel L. Moody, Guttorm Sindre, Teqje Brasethvik, "Evaluating the Quality of Information Models:Empirical Testing of a Conceptual Model Quality Framework"**,** 0-7695-1877-X/03 © 2003 IEEE.

[5] Calero C., Piattini M., Carolina Pascual, Serrano, M. A. "Towards Data warehouse quality metrics",3rd International workshop on design and Management of Data warehouses (DMDW 2001), Interlaken,Switzerland

[6] Serrano, M.Calero, C.Piattini, M., "Validating metrics for data warehouses" IEEE Proceedings SOFTWARE, 2002.

[7] Fenton, N., & Fleeter, S. Software metrics: A rigorous approach (2nd Ed.). London: Chapman & Hall. 1997

[8] Serrano, M.Calero, C.,Sahraoui, H.,Piattini, M., Empirical Studies to Assess the Understandability of
Data Warehouse Schemas using Structural Metrics, Software Quality Journal. Springer, 2008. Pages: 79-106

[9] M. Serrano, C. Calero, J. Trujillo, S. Lujan, M. Piattini, Empirical validation of metrics for conceptual models of data warehouse, 16th International Conference on Advanced Information Systems Engineering (CAISE'04), Riga, Latvia, 2004, pp. 506–520.

[10] J. Trujillo, M. Palomar, J. Go´mez, I.-Y. Song, Designing Data Warehouses with OO Conceptual Models. IEEE Computer, Special issue on Data Warehouses 34 (2001) 66–75.

[11] M. Serrano, Definition of a Set of Metrics for Assuring Data Warehouse Quality, University of
Castilla, La Mancha (Spain), 2004.

[12] Jarke, M., LenzerinI, I. M., Vassilou, Y., & Vassiliadis, P. Fundamentals of data warehouses,
Springer, 2000

[13] Serrano, M.Calero, C.Piattini, M., "Validating metrics for data warehouses", IEEE Proceedings SOFTWARE, 2002

[14] So S.S., Cha S.D., Kwon Y.R. Empirical evaluation of a fuzzy logic-based software quality prediction model (2002) Fuzzy Sets and Systems, 127 (2), pp. 199-208.

[15] José A. Cruz-Lemus, Marcela Genero, José A. Olivas, Francisco P. Romero, Mario Piattini: Predicting UML Statechart Diagrams Understanadability Using Fuzzy Logic-Based Techniques. SEKE 2004:238-245
.
[16] Yuan, X.; Khoshgoftaar, T.M.; Allen, E.B.; Ganesan, K.; "An application of fuzzy clustering to software quality prediction," Application-Specific Systems and Software Engineering Technology, 2000. Proceedings. 3rd IEEE Symposium on, vol., no., pp.85-90, 2000.

[17] So S.S., Cha S.D., Kwon Y.R. "Empirical evaluation of a fuzzy logic-based software quality prediction model (2002) Fuzzy Sets and Systems", 127 (2), pp. 199-208.

BIOGRAPHIES

Kunwar Babar Ali pursuing his M.Tech (IT) from university school of information technology (USIT), GGSIPU, New Delhi. Before joining his M.Tech regular course, he worked as a lecturer in the dept. of CSE in more than two engineering colleges located in NCR. He has more than 3 years of teaching experience. He is the member of ICEIT, New Delhi .He has been published more than 5 national and international research paper in reputed journals His research area includes Software engineering, Data Warehouse design metrics and machine learning techniques.

Dr.Anjana Gosain working as an Associate Professor in university school of information technology (USIT), GGSIPU, New Delhi. She is the member many technical bodies in India as well abroad .She has more than 18 years of teaching experience. She has published more than 30 research papers in reputed international journals. Her research area includes software engineering, data warehouse desgn, cloud computing, distributed computing.