International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869, Volume-2, Issue-4, April 2014

Optimization in SQL by Horizontal Aggregation

Krupali R. Dhawale, M.S.Gayathri

***Abstract*—Data mining is similar to discover a new idea. When data are prepared under RDBMs it will be very critical task to make it out. A lot of data mining concepts and algorithms are used to create prepared datasets in tabular format which consist complex queries, joining tables in data mining. Existing SQL aggregations having limited capacity to prepare data sets and it is very time consuming task. They return one column per aggregated group. Hence fundamental methods are used to determine horizontal aggregation to show an outline made by SQL code to return in a horizontal tabular format by using SPJ, CASE and PIVOT methods. This group of new function is called as horizontal aggregation. Aim of this work is to present classification on prepared datasets and further, the decision tree is generated by using C4.5 algorithm to reduce the time constraints. Proposed work might be useful for the programmers to interpret the knowledge in the form of decision tree.**

***Index Terms*— aggregations, SQL, pivoting, data preparations.**

I. INTRODUCTION

Data mining is the analysis step of the ("Knowledge Discovery and Data Mining" process, or KDD), It is the computational process of discover patterns in huge data sets involving a lot of methods at the intersection of machine learning, statistics, and database systems. In general aim of the data mining process is to extract information from a data set and transform it into an understandable structure for advance use. Data mining include major elements are transform, extract and data transformation onto the data warehouse system. It keeps and manage the data in a multidimensional database system. Data may be facts, numeric or text that can be access by computer. Data may be operational e.g. Sales data or nonoperational data such e.g. forecast data and Meta data that is data about data. The method of data preparation consists of three stages – data discovery, data characterization and data set gathering. In data discovery, selected data are made available from their sources and on that basis given data are suitable used for data mining. The data is evaluated for their convenience in data mining, which involves the use of data profile and variable status report. Then a data set is created from the sample through an exact focus based on selected fields or features. The data set improved and enhanced by the data mining tool for data transformations which is used for data preparation. Data preparation ends with a concluded set of reports. The data is evaluate for their convenience in data mining, which involves the use of data profile and variable status report.

Then a data set is created from the sample through an exact focus based on selected fields or features. The data set improved and enhanced by the data mining tool for data transformations which is used for data preparation . Data preparation ends with a concluded set of reports, describing the data and the data sets.

Describing the data and the data sets. There are common terms such as point dimension, statistics literature, observation variable under the data mining. Data mining have a lot of challenges to turn the huge amount of data into knowledge cube for global challenges. Existing SQL Query in DBMS return a data sets but they having limitation to return a data by using complex queries, joining tables and tables.

In this project, three fundamental methods are used to evaluate horizontal aggregation in SQL to prepare data sets. Two common data preparation tasks are explained in this project.

1) Transposition/aggregation and
2) Transforming definite attributes into binary dimensions.

For this purpose include two strategies to evaluate horizontal aggregations using follows strategy.

SPJ strategy
CASE strategy

Selection-Projection –Join (SPJ) this method is fully depends on the relational operators, Pivot method is use to exchange rows and columns, that appears the data transformations useful to create the data into visualization mode and data modeling. Lastly, CASE method is used to construct the SQL CASE programming.

Horizontal aggregations collect extra features of standard SQL aggregations, which return a set of values in a horizontal layout which is standard denormalized tabular layout. This is standard layout used in most data mining algorithms .With the help of three fundamental methods aggregate the data into standard layout. By using three standard methods it gives same result. That prepared datasets is stored in the form of .arff format. On the prepared data sets C4.5 algorithm is applied, further the decision tree is generated in the form of tree structure. IF –Else condition is more comprehensible that interpret easily understand the database.

II. LITERATURE SURVEY

There exist many proposals research that have extended SQL code for the data mining operations. Related works extended on query optimization, comparison between horizontal aggregations with alternative to perform aggregation, pivoting and transposition.

Manuscript received April 19, 2014.
Krupali R. Dhawale, ME student**,** Alard College of Engg. and management, Marunje, Pune, Pune University . Pune, India.
M.S.Gayathri, Associate Professor of computer and science Department of Alard College of Engg. and Management, Pune, India.

Rajesh Reddy Muley, Sravani Achanta and Prof.S.V.Achutha Rao, in [1] are give to support to optimize acquire prepared Data Sets for Data Mining Analysis in SQL and dropping the overload on the databases for recovery of data. In this paper they used CASE method has two possible methods i.e. Vertical view and also the Horizontal View. Durka. C and Kerana Hanirex.D planned that original regular data of pivoting option is built-in by means of Data mining can be achieve with the device SAAS (SQL Server Analysis Services). In [2] knowledge data will be modified based on "Generalized and Suppression Algorithm" and generate the building for the Dataset in Data Mining analysis. Nisha. S and B .Lakshmipathi, [4] optimized Horizontal Aggregation in SQL by Using K-Means Clustering algorithm. The system make use of single parent table and different child tables, operations are then perform on the data overloaded from several tables. PIVOT operator is use to estimate aggregate operations in this paper. "Classifying the large set of data, obtain from the end result of horizontal aggregation in to identical cluster."This task is implemented by K-means algorithm in this paper[4]. Pradeep Kumar and Dr. R. V. Krishna[5] used CASE, SPJ and PIVOT methods to build up a prototype function and the experiential results found. These constructs are able to generate the datasets that can be use for additional data mining operation.

K. Anusha, P. Radhakrishna and P. Sirisha used SPJ Method and correspondence Methods for horizontal aggregation[6]. The CASE method has main contribution in this paper. Since it can be programmed combine GROUP-BY and CASE statements. Proved this three methods construct the identical result. Generate SQL queries with three sets of parameters: grouping columns, sub grouping columns and aggregated column. C. Ordonez proposed two common data preparation tasks are explained [7] those are 1)Transposition/aggregation and 2) Transforming definite attributes into binary dimensions. For this task , C. Ordonez proposed two strategies to estimate horizontal aggregations those are SPJ strategy and CASE strategy. Mr. Ranjith Kumar K and Mrs. Krishna Veni, First, they write to a template to generate SQL code from a data mining tool and make available several unique features and advantages[8]. Such SQL code automate writing SQL queries, optimizing them and testing them for correctness in this paper . Sunil Kumar, N. Surya Prakash Raju used K-means algorithm to prepared datasets for data mining related work and with the help of PIVOT operator horizontal aggregation operation is performed. The existing systems are not defined for the dissimilar fact tables that need better indexing and extraction . To solve this drawback by using Multiple fact tables and k-means algorithm in [9]. Since established query graph models are not enough for model outer join queries with complex predicates. Presents the needed hyper graph abstraction and algorithms for reordering such queries with joins and outer joins are used in [10]. For this purpose, P. Goel, and B.R. Iyer , used Conflict free Assignment algorithm .

Finally, this paper is a significant extension of the preliminary work presented in [3],where horizontal aggregations is carry out by the three fundamental methods are SPJ,CASE and PIVOT methods. In previous work, three fundamental methods provide summarize for SQL code to produce organized datasets in tabular layout for data mining analysis. But, existing SQL query having limited capacity to prepare datasets is very time consuming task. It consist large table. Those prepared datasets tables cannot easily interpret. This paper is helpful to understand the prepared datasets in the form of decision tree structure by using c4.5 algorithm in WEKA.

III. EXISTING SYSTEM

- *Issues in the existing system*

An existing system to prepare a data set for analysis is usually the mainly time consuming task in a data mining project, require a lot of compound SQL queries, combination of tables and aggregating columns. Existing SQL aggregations have boundaries to create data sets because they return one column per aggregated group.

- *Disadvantages of Existing System:*

1) Existing SQL aggregations have boundaries to create data sets.

2) To return single column per aggregated group.

IV. PROPOSED MODEL

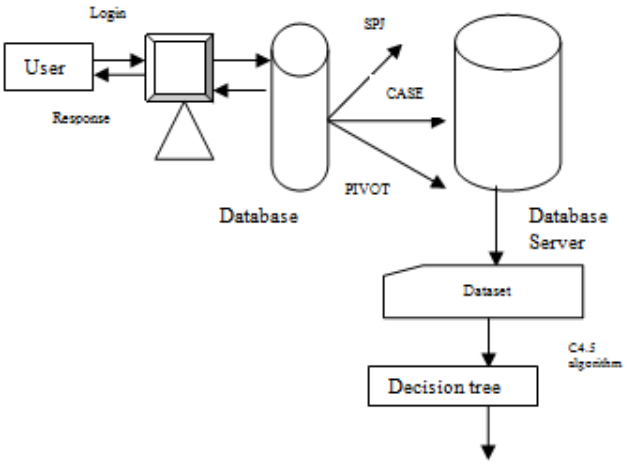
Addressing the constraints of time limitation and complexity with respect to query formation, it is planned to alleviate these hurdles by using three fundamental methods such as SPJ, CASE and PIVOT method for horizontal aggregation in SQL to prepare datasets. The proposed model consists

(1) First, they represent a template to generate SQL code from a data mining tool. Such SQL code automates writing SQL queries, optimizing them, and testing them for correctness. This SQL code reduces manual work in the data preparation phase in a data mining project.

(2) Second, since SQL code is automatically generated it is likely to be more efficient than SQL code written by an end user. For instance, a person who does not know SQL well or someone who is not familiar with the database schema (e.g., a data mining practitioner). Therefore, data sets can be created in less time.

(3) Third, the data set can be created entirely inside the DBMS. In modern database environments, it is common to export denormalized data sets to be further cleaned and transformed outside a DBMS in external tools (e.g., statistical packages).

(4) Further prepared a datasets is classified into decision tree with the help of C4.5 Algorithm.



.Fig.1: System Architecture for Proposed model

V. PROPOSED METHODOLOGY

In this section proposed horizontal aggregations provide process for creating datasets in data mining analysis. The main goal of this paper to define creates an outline to produce SQL code in tabular layout from data mining tools, these SQL code can be created using SQL complex queries, joining tables and used for prepared datasets in SQL for data mining analysis. A

Second goal of this paper to create decision tree by using C4.5 algorithm in WEKA on the prepared datasets. Proposed methodology is explained for four modules given as below.

1) Login Application for User:

Proposed work will be able to upload various details regarding separate username and password. Firstly it will create new login application form, will be registered with various user name and password. It access only authorized person. It provides privacy for prepared datasets within DBMS. If new user want to create account in login application form. It will be created then the dialog box shows the message as login successfully. If username and password are mismatched then dialog box shows incorrect username and password. Datasets are created using SPJ practically. Database created in Microsoft SQL Server 2008. The Microsoft SQL Server 2008 Database Engine is a service for accumulate and giving out data in either a relational (tabular) format or as XML documents. Datasets are preparing from the output of all these three method, so by applying SPJ, PIVOT and CASE methods on given database. Dataset are found in .arff file format.

2) Implementation of aggregated SPJ, CASE and PIVOT methods:

Addressing the problem with the prepared datasets this paper proposes three fundamental methods are SPJ, CASE and PIVOT used for horizontal aggregation in SQL to prepare datasets. The methodology adopted for the proposed plan of implementation, transposition and aggregations by following methods:

- *SPJ method :*

In SPJ Method sub query execute first. After that parent query execute. Select-projection-join (SPJ) method depends on the relational operator. Vertical operations are used in SPJ method. For every column one table is generated in this model. Afterwards, the tables generated are joined in order to obtain final horizontal aggregations

Left Outer Join is use in SPJ method, the left outer join is evaluated in between two tables i.e. right part of table and left part of table. The common fields of both the right and left tables are returned and uncommon fields of left column are also returned.

In a horizontal aggregation having four input parameters to create SQL code:-

(i) The known input table F
(ii) The record of GROUP BY columns L1, …… ,Ln
(iii) The column which to be aggregate into (A) and
(iv) The record of transposing columns R1, … ,Rk.

The actual implementation is based on the details given in data sets.

Proposed syntax is as follows.

SELECT (L1… Lj), H (A BY R1, … ,Rk)
FROM F
GROUP BY (L1… Lj);

- *CASE method:*

This task is based on the CASE construction provided by SQL. It has a lot of built in Boolean expressions. Out of them one of the expressions is returned. Aggregation or Projection is like to this from relational query point of view.

In SQL CASE constructs are available in the SQL CASE programming .It can be done by using many conditions with conjunctions. In this case horizontal aggregations exhibit two strategies.

1) Firstly, the computations of query can be done directly from the given input database table.
2) Secondly, evaluate vertical aggregation and the results are sent to an arbitrary table. This table is used again in horizontal aggregation generation.

- *PIVOT method:*

RDBMS has built in PIVOT operator. This is used for the PIVOT operation proposed in this paper. This construct can provide transpositions. It transposes the fewer of rows into additional new column. Therefore, for evaluating horizontal aggregations pivot operator is used to transfer the data from row into column in it.

3) Implementation of C4.5 algorithm:

Decision tree builds classification models in the form of a tree structure. It splits down in a dataset into smaller and smaller subsets while at that time a related decision tree is incrementally developed. The last outcome is a tree with decision nodes and leaf nodes.

C4.5 algorithm is the latest version of ID3 algorithm. In this module implementation of the C4.5 algorithm will be perform by using the Weka tool. The dataset created by three SPJ, PIVOT and CASE methods, this Prepared dataset is given as an input to C 4.5 algorithm with the help of WEKA tool to generate Decision tree or classification. On that prepared dataset calculate the Entropy and Information gain .Operation are then performed and an appropriate decision rules are produce. Depends on that rule Decision Tree is created. Entropy of each attribute is calculated in every branch. C4.5 algorithm is implemented in WEKA and linked with java file.

4) Decision Tree:

Graphical representation of the output of all the methods implemented before in proposed model. Decision tree is generated on the basis of the prepared datasets.

B) Data Flow Work

Data flow chart fig. 2 shows initially create new login form it will be registered with various user name and password. It access only authorized person. If new account is created then the dialog box shows the message as login successfully.SPJ, PIVOT and CASE method applied on the query. Give the same result by three methods. Prepared datasets store in the form of .arff format. On the prepared datasets c4.5 algorithm is applied and finally the decision tree is generated.

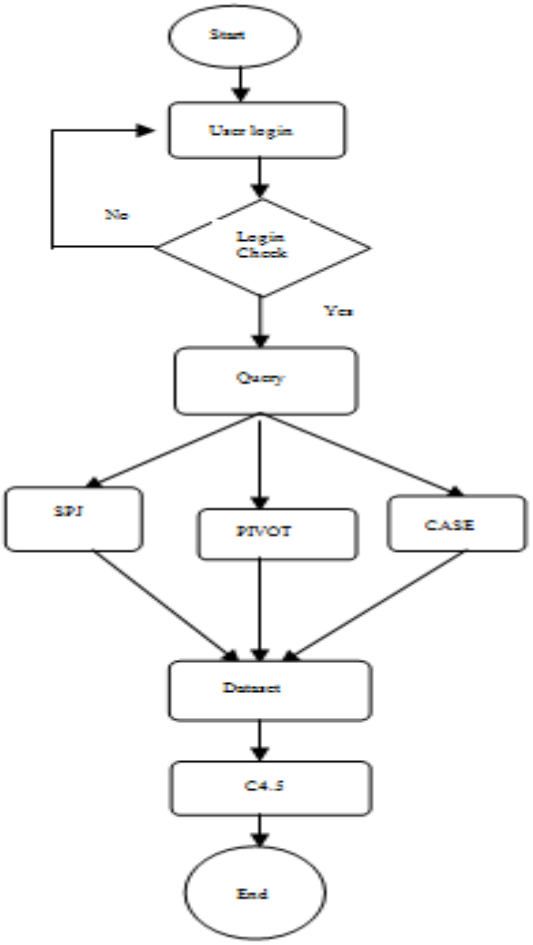


Fig.2: Data Flow Chart

VI. ALGORITHM DESIGN

A) C4.5 Algorithm

C4.5 algorithm constructs classification model in the form of tree structure and its predecessor, that summaries training data in a decision nodes and leaf nodes. Along with final result is a tree with child nodes leaf nodes makes logical rules to satisfy IF-Else condition. Leaf node represents a categorization or decision. The topmost decision node in a tree which corresponds to the top predictor called root node. Decision trees have capacity to handle both definite and numerical data.

C4.5 Algorithm is latest version of ID3 algorithm. C4.5 Algorithm is a well-liked tree based classifier, is used to generate decision tree and from a set of training examples. Nowadays C4.5 is named as J48 classifier in WEKA tool, an open source data mining tool. The function of heuristic used in this classifier is depending on the concept of information entropy.

In general, steps in C4.5 algorithm to build decision tree are:

step1: Choose attribute for root node

Step2: Create branch for each value of that attribute

Step3: Split cases according to branches

Step4: Repeat process for each branch until all cases in the

Branches have the same class .Choosing which attributes to be a root is based on highest gain of each attribute.

To count the information gain, we use formula 1, below :

$$\text{Gain}(S,A)=\text{Entropy}(S)-- \sum_{i=1}^{n} \frac{Si}{S} * Entropy(Si) \quad \ldots\ldots\ldots(1)$$

With:

{S1,... Si,… Sn} = partitions of S according to values of Attribute A

n = number of attributes A

|Si| = number of cases in the partition Si

|S| = total number of cases in S

While entropy is gotten by formula 2 given as below:

$$\text{Entropy}(S) = \sum_{i=1}^{n} - pi * \log_2 \text{pi} \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

With:

S : Case Set

n : number of cases in the partition S

pi : Proportion of Si to S

- *Tool used* **:**

Weka tool was developed at the University of Waikato in New Zealand. It is most popular. Weka is a set of machine learning algorithms for data mining tasks. Weka include tools for classification (e.g. KNN, C4.5 Decision Tree, Neural Networks), data pre-processing (e.g. Data Filters), clustering, association rules, and visualization etc. Input data in the Weka tool is in the form .arff format. This tool is an open source data in Java. Generally, search time is found between ID3 and C4.5 algorithm in sec. which is shown in table I. The result demonstrate optimization approach by horizontal aggregation in SQL to prepare datasets for data mining analysis to aggregate the data in horizontal tabular layout by SPJ,PIVOT and CASE methods. Give the equivalent result by using those three methods . The Database created in Microsoft SQL Server 2008. The Microsoft SQL Server 2008 Database Engine is a service for accumulate and giving out data in either a relational (tabular) format or as XML documents. The technologies used Java Language 1.5 and J creator which is meant for developing the result of horizontal aggregation. The environment used for the development of decision tree by using C4.5 algorithm in Weka tool.

Table I

Search Time Comparison between ID3 and C4.5 Algorithm in Sec.

Algorithm		
Size of Data Set	**ID3**	**C4.5**
14	0 sec	0 sec
24	0 sec	0 sec
35	0.577sec	0.421 sec

VII. CONCLUSION AND FUTURE SCOPE

Existing SQL Query having limitation to prepare datasets. Preparing a data set for analysis is generally the normally time consuming task in a data mining project. Optimizing the workload is challenging problem. For this purpose, the

fundamental methods SPJ, CASE and PIVOT are used to estimate horizontal aggregation in SQL to prepare datasets. It give equivalence results.

It is designed to implement these methods on prepared datasets and further, the decision tree is generated by C4.5 algorithm to reduce the time constraint. With the carrying out of methods of Horizontal Aggregation Datasets are created and these Datasets are used to generate Decision Tree. Decision Tree is generated using C4.5 algorithm in WEKA. This is new technique to reduce the number of optimize decision tree among the prepared datasets. Model built by C4.5 algorithm is require less time than that of ID3. Memory used for storing C4.5 Dataset is comparatively less than ID3.In future, use of C4.5 algorithm will helps to decrease time limit required for building model of a particular dataset and also it require less memory to store its Datasets.

ACKNOWLEDGMENT

I would like to express my sincere thanks to my Guide **M.S.Gayathri, Associate Professor of Alard College of Engg. And Management, Pune** for her consistence support and valuable suggestions.

REFERENCES

[1] Rajesh Reddy Muley, Sravani Achanta and Prof.S.V.Achutha Rao, "Query Optimization Approach in SQL to prepare Data Sets for Data Mining Analysis", *International Journal of Computer Trends and Technology (IJCTT)* – volume 4 Issue 8 , pp 1-5,August 2013.

[2] Durka. C and Kerana Hanirex.D, '' An Efficient Approach for building Dataset in Data Mining", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 3,pp 1-5, March 2013.

[3] Carlos Ordonez and Zhibo Chen," Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis", *IEEE transactions on knowledge and data engineering*, vol. 24, NO. 4, pp 1-14, APRIL 2012.

[4] Nisha. S and B.Lakshmipathi, "Optimization of Horizontal Aggregation in SQL by Using K-Means Clustering", *International Journal of Advanced Research in Computer Science and Software Engineering* , Volume 2, Issue 5, ISSN: 2277 128X, PP. 1-6, May 2012.

[5] Pradeep Kumar and Dr. R. V. Krishna, "Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis" *IOSR Journal of Computer Engineering (IOSRJCE)* ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 5, PP. 36-41, (Nov. - Dec. 2012).

[6] K. Anusha, P. Radhakrishna and P. Sirisha," Horizontal Aggregation using SPJ Method and Equivalence of Methods", *IJCST, Vol. 3, Issue 1, Spl. 5*, pp 1-4, Jan. - March 2012 .

[7] C. Ordonez, "Horizontal Aggregations for Building Tabular Data Sets," *IEEE Trans. Knowledge and Data Eng*," VOL. 24, NO. 4, pp. April 2012.

[8] Mr. Ranjith Kumar K and Mrs. Krishna Veni," Prepare datasets for data mining analysis by using horizontal aggregation in SQL", *Ranjith Kumar K et al, Int.J. Computer Technology & Applications*, Vol 3(6), 1945-1949 IJCTA, pp. 1-5, Nov-Dec 2012.

[9] Sunil Kumar, N. Surya Prakash Raju ,"Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis". *IJCST*, Vol. 3, Issue 3, July – Sept. 2012.

[10] P. Goel, and B.R. Iyer, "Hyper graph Based Reordering of Outer Join Queries with Complex Predicates," *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '95*) , pp. 304-315, 1995.

[11] Swetha .Palabindela and Ch. Rajya Lakshmi, "Custom Aggregations for Generating Datasets for Data mining", *International Journal of Advanced Research in Computer and Communication Engineering*,Vol. 2, Issue 9, PP. 1-4, September 2013.

[12] K. Anusha, P. Radhakrishna and P. Sirisha "Horizontal Aggregation using SPJ Method and Equivalence of Methods", *IJCST* ,Vol. 3, Issue 1, Spl. 5,PP 854-857 Jan. - March 2012.

[13] C. Ordonez, "Data Set Preprocessing and Transformation in a Database System," *Intelligent Data Analysis*, vol. 15, no. 4, pp. 613-631, 2011.

[14] C. Cunningham, G. Graefe, and C.A. Galindo-Legaria, "PIVOT and UNPIVOT: Optimization and Execution Strategies in an RDBM S," *Proc. 13th Int'l Conf. Very Large Data Bases (VLDB '04)*, pp. 998-1009, 2004.

[15] C. Ordonez, "Vertical and Horizontal Percentage Aggregations," *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '04),* pp. 866-871, 2004.

[16] H. Wang, C. Zaniolo, and C.R. Luo, "ATLAS: A Small But Complete SQL Extension for Data Mining and Data Streams," *Proc. 29th Int'l Conf. Very Large Data Bases (VLDB '03)*, pp. 1113- 1116, 2003.

[17] Han and M. Kamber, "Data Mining: Concepts and Techniques, first ed. Morgan Kaufmann", 2001.

[18] Mark A. Hall, "Correlation-based Feature Selection for Machine Learning", *Department of Computer Science Hamilton, NewZealand*, PP.1 -198, April 1999.

[19] S. Sarawagi, S. Thomas, and R. Agrawal, "Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications," *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '98)*, pp. 343-354, 1998.

[20] P. Goel, and B.R. Iyer, "Hyper graph Based Reordering of Outer Join Queries with Complex Predicates," *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '95)* , pp. 304-315, 1995.

Krupali Rupesh Dhawale , received her B.E. degree from M.I.E.T. Gondia, Nagpur University, and Maharashtra state, India in the year 2008. Currently, Pursuing M.E. in Computer Engineering from Alard College of engineering and management, Marunje, Pune, Maharashtra state, India. Her research interests include in Data mining, Mobile computing. Her 2 International and 1 national Publications.

Recent Publication: - IJCEMR, IJSETR, ETIT national conference.

Paper name: - Horizontal Aggregation in SQL to prepare dataset for Data Mining Analysis.

M.S.Gayathri, Associate professor, received the Bachelor's degree in Computer Science & Engg. from P.S.N.A college of Engg & Tech., Madurai Kamaraj University, Tamilnadu state, India and Master degree in computer science and Engg. from Jeusalem College of Engineering, Anna University, Tamilnadu state, India, Her research interests includes network security, biometrics, image processing, data mining, artificial intelligence, compilers and natural language processing.