ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 3, Issue 1 (Jan. - Mar. 2015)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2015)

CoDe Model for Exploiting Visualization Report from Data Warehouse

[I]T.P.Latchoumi, [II]P.Neeranjana, [III]D.Devi, [IV]E.Suganya
[I]Assistant Professor (CSE), CCET, Pondicherry, India
[II,III,IV]Student (CSE), CCET, Pondicherry, India

Abstract
CoDe model helps in organizing the visualization through the visual language CoDe [14] which represents relationships between information graphically. The visualization of different data in any reports should create a user-friendly manner for efficiency and manage complexity. A complete analysis has been made to every individual report using graphical representation. The standard graphs such as bar charts, pie diagrams, and scatter plot charts can be used for visualizing the data. The use of different types of graphs allows the user to view the specific information in more detailed manner. This model only allows for visualizing the data but in this paper additional details are also visualized. This paper presents a more improved model of CoDe using clustering technique. A cluster is hence a collected works of items which are "alike" between them and are "unlike" to the items belong to previous clusters. An important factor of a clustering algorithm is the distances assess between objects. The proposed method defines the stages of CoDe modelling, distributed clustering, subset selection algorithm, providing data into table and finally visualizing the data. The final visualization adds some effects of information visualization to exploit the paradigm with improved additional information.

Keywords
Information Visualization; CoDe Modelling; Distributed Clustering; Feature Selection; Datasets.

I. Introduction

Statistics presentation can be provided in effective styles, elegant and eloquent manner. There is a wide range of predictable ways to visualize data – tables, histograms, pie charts and bar graphs are being used every day and on every possible occasion. However, to deliver a message to the readers efficiently, occasionally you necessitate additional than just a simple pie chart of your results. In fact, there are much better, superficial, creative and absolutely captivating ways to visualize data. Many of them might become omnipresent in the next few years.

A major goal of statistics visualization is to communicate information clearly and efficiently to users via visualization particular, namely scatter tables and bar charts. Effectual image helps users in analysing and reasoning about data and evidence. Graphical Representations makes it easy to understand and interpret data at a glance.

Visual representation also helps to do contrasts among many things. Moreover visualization makes data easy to recall. This makes complex data more manageable, reasonable and functional. Users may have particular analytical tasks, such as making evaluations or understanding connection, and the design principle of the graphic involves the chore. Table data are usually done where people will make look at an exact assess of an element, whereas graphical representation of a variety of classifications are worn to illustrate patterns or relationships in the data for one or more variables. A Graphic language named CoDe [14] (Complexity Design) designates visual representations which are based on the concept of macroscope. This graphical language is an agenda for a procedure intended to support visual representation of composite information. CoDe is projected as an interactive tool that supports the designer of an imagining project in describing the complex data which is the basis for most decision processes. The idea of introducing CoDe is introduced by de Rosnay with the concept of macroscope is that a complex system can be better understood but it is portrayed like complete, relatively than as a work of split metaphors. For instance, microscope helps in observing tremendously minor occurrences, and a telescope helps in observing enormously enormous items, so a macroscope is required to compact with tremendously composite methods.

This paper presents a technique to group the objects which are similar between them and are dissimilar to the objects be appropriate to others clusters. Clustering [18] can be considered the most important unsupervised learning problem and deals with finding a structure in a collection of unlabelled data. This technique could be "the process of organizing objects into groups whose members are similar in some way. This technique has been used to cluster words into groups based either on the distance between the items they contain.

In Section II, we depict the background which is the basis for the techniques used. In Section III, we point up the related works based on the proposed methods In Section IV, we theoretically define the approach of clustering and CoDe model we describe the proposed methods which involve algorithm and final visualization designs and in Section V, we illustrate final observations and further works.

II. Backgrounds

We give a brief background of data mining and data warehouse inspired by that concepts, as well as related work on CoDe models and attempts to mitigate visualization.

A. Data Mining

Data mining is a process of discovering patterns in large data sets involving methods at the intersection of database methods and other schemes. The physical mining of patterns from data has occurred for centuries. Untimely methods of identifying patterns in data comprise Bayes' theorem (1700s) and regression analysis (1800s). The complete goal of the data mining process is to extract information from a data set and transform it into a comprehensible structure. This involves database and data management aspects, visualization, and online updating.

The definite data mining task is the spontaneous analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records cluster analysis, unusual records and association rule mining. The concrete data mining task is the usual or semi-automatic analysis of huge quantities

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

of data to take out earlier unknown appealing patterns such as groups of data records, strange records and dependencies. The data mining stage may recognize numerous groups in the data, which can then be used to attain more precise prophecy results by a decision support system.

B. Data Warehouse

Data warehouse is a process of integrated data management and retrieval. A data warehouse refers to a database that is maintained individually from an organization's functioning database. Data warehousing is arelatively new term although the concept itself has been about for few years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Present an effortless and concise view around particular subject issues by apart from data that are not useful in the decision support process. The time prospect for the data warehouse is considerably longer than that of operational systems.

In addition, data warehouse is a semantically reliable data store that serves as an objective functioning of a decision support data model and stores the information on which an enterprise needs to make intentional decisions. A data warehouse is based on a multidimensional data model which views data in the form of a data cube.

III. Related Work

Augmentation in the area of data analysis has been unquestionably rapid. Numerous techniques and tools have been projected in the literature to relate and display information extracted from a data warehouse.Ciuccarelli et al. [14] present graphic Language named CoDe (Complexity Design)involves relating graphical representations which provide information according to the basic idea of visualization. CoDe is a frame aimed to support visual representation of complex information.

Ma et al. [12, 13] illustrate the design and the implementation of a meteorological data warehouse. This approach uses Microsoft SQL Server. In exacting, the proposed system creates meteorological data warehousing processes based on SQL Server Analysis Services and uses SQL Server Reporting Services to design, execute, and manage multidimensional data reports.

Hsu and Li [11] concern a clustering analysis on OLAP reports to verify the group information between dissimilar OLAP reports. Su and Su [9] depict a procedural structure to create a report system based on a three-layer calculating architecture which implements a metadata mapping, an ETL module and a DW. In particular, the future system uses numerous unusual data sources and performs statistical estimate to combine them. Moreover, it allows essential and extending different details models that are in tabular form or graphically represented.

Anfurrutia et al. [8] present a product-line approach to database reporting based on preventability and relationship among reports. In particular, this approach exploits a aspect model that provides an conceptual and short language rules to convey commonality and variability in a database reporting. Data reports are provided textually. Another technique is based on reports represented as animated transitions.

In meticulous, Heer and Robertson [10] inspect the efficiency of animated transitions between common statistical data graphics such as Bar-charts, Pie, and Scatter-plots. Disparate our approach they do not generate a single visualization and do not show relationships between data; data used to visualize the initial graph are the same used for the conclusion one.

Papageorgiou [7] investigation work tries to check the most recent applications and trends on fuzzy cognitive maps (FCMs) at the last ten years. FCMs use cyclic directed graphs, for knowledge representation and reasoning. In the past decade, FCMs have gained considerable research interest and are widely used to analyse causal systems such as system control, decision making, management, risk analysis, text categorization, prediction etc. Their dynamic characteristics and learning methodologies make them essential for modelling, analysis, prediction and decision making tasks as they improve the performance of these systems.

Mansmann and Scholl [16] establish a visual structure called Enhanced Decomposition Tree in which each level of the tree structure is created in a disaggregation step along a chosen dimension, the nodes contain the corresponding sub aggregates arranged in a graph and the arcs are labeled with their respective values. There are various layouts offered to support various activities of analysis. Data cubes are queried using a visual browser based upon a scheme which has dimensions hierarchies of their granularity levels, thus providing an effi create hierarchical views. Data cubes can also multiple according to shared dimensions. The main area of the visual browser is exploited to show the results of user interaction in a selected visual format.

Iakovidis and Papageorgiou [6] describe the medical decision making techniques which can be regarded as a process, combining both analytical cognition and intuition. It involves reasoning within complex causal models of multiple concepts. Focusing towards a model called medical decision making, in which it propose a novel approach based on cognitive maps and intuitionist fuzzy logic. The new model, called intuitionist fuzzy cognitive map (iFCM), extends the existing fuzzy cognitive map (FCM) by considering the expert's hesitancy in the determination of the causal relations between the concepts of a domain.

Qinbao Song, Jingjie Ni, and Guangtao propose a feature selection method which identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. Based on these criteria, a fast clustering-based feature selection algorithm (FAST) [18] is proposed. Features in different clusters are relatively independent and the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features.

From the survey of the related works we come to know about few techniques and design methodologies as follows

- CoDe[14] supports visualization of complex information
- Clustering techniques are used need more efficiency
- Data can be viewed in terms of animated transitions
- The performance of system can be improved using Fuzzy Cognitive Maps
- The dimensions of data can be analyzed using hierarchical manner

IV. Proposed Method

The proposed approach will be presented in the support of CoDe [14] model in which the methodologies are discussed based on different techniques involves clustering and others as mentioned previously. The aim is to improve user satisfaction by recurring data that have a higher efficiency and understanding of complex data to be accepted by the user.

A. Theoretical Vision of Code

CoDe [14] supports the deeds of graphic representation of complex systems, by a environment that simplifies both the process of components, and the process of interconnection, which are the basis for many decision activities. This provides visualization to be composed relating one or more different types of graphs. The graphical language visualizes representation of data and possesses stability between data. In demand to define a graphic visualization in CoDe, we measured a starting set of graph types representing data such as Histogram, Pie charts. The OLAP allows data analysing process and decision support system. This operation permits dat in a multidimensional manner. This OLAP definition extracts data from the data warehouse and split into hierarchies. The data are reported in a multidimensional cube.

Table 1. Shares report

SHARES		
COMPANIES	PRICE	CHANGE
Reliance	907.00	26.75
Coal India	378.00	10.60
ICICI bank	354.35	9.15
Cipla	341.90	6.30

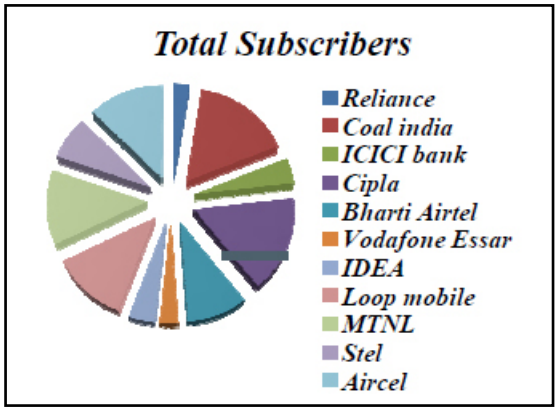

Fig.1: Report of total subscribers

The production of a report from a data mart maps the cube dimensions on a structure collected by resultant to the components in the report and one or more data. The resulting report is extracted by applying a combination of selection and or aggregation slicing, dicin0g, pivoting, rolling, drilling dimensional operators (i.e., OLAP operations that allow multidimensional data analysis) [14]. We define operation pattern the combination of OLAP operations to be achieved. Process patterns are described taking into account only metadata of the data mart. The definite implementation of action patterns to extract data is performed during the OLAP Operation phase at the end of the design process.

Table 1.describes the shares from different companies are invested in stock market and gain profitable returns. This table shows the view of data items stored in the data warehouse. The surveys of the total subscribers are visualized in the Fig. 1. That shows the information in graphical representation in the form of pie charts with difference between each data item.

B. Clustering Tactic

Cluster analysis was originated by Driver and Kroeber in anthropology in 1932 and famously used by Cattell. Cluster analysis involves different types of techniques which can be achieved by various algorithms. These algorithms differ significantly in their notion which constitutes cluster and find them efficiently. Clusters include collection of objects with small distances among the cluster members and statistical distributions.

Cluster analysis depends on the individual dataset and use of the reports. There is no objectively perfect clustering step by step procedure, although as it was illustrious, cluster analysis is in the zenith of the beholder. The clustering model most closely related to statistics is based on distribution technique. This technique is based on same distribution of object. This methodology usually be able explain the data better and more complex model. The distribution clustering pays extreme burden on the user for real data from many data sets. There have been several suggestions for a measure of similarity between one or more clusters.

This measure can be used to compare how well different data clustering algorithms perform on a set of data. The types of evaluation methods measure how close the clustering is to be predetermined.

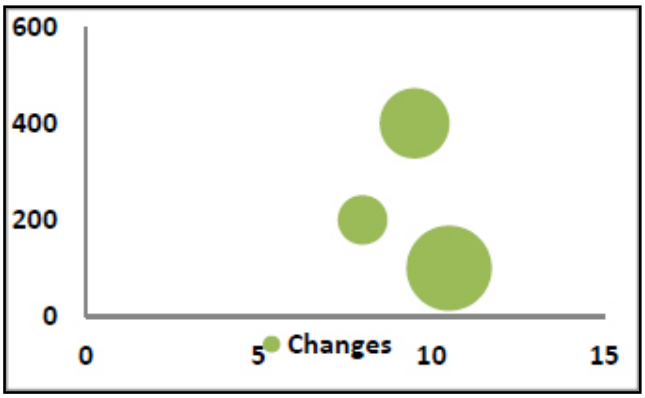

Fig. 2:: Clustering of changes in shares

Data objects are grouped according to reasonable relations or end user inclination. For example, data can be mined to identify market segments or end user resemblance. Cluster analysis itself is not one particular procedure, however the common chore to be worked out. It can be performed by different algorithms that differ significantly in their notion of what constitutes a cluster and how to capably find them. Accepted design of come together consists of collections with small distances among the group components, intense spots of the statistics gap, period or exacting arithmetical divisions.

Clustering is the task of categorizing the vertices of the graph into clusters taking into consideration the edge structure of the graph in such a way that there should be many edges within each cluster and relatively few between the clusters. Graph clustering in the sense of combining the vertices of a given input graph into clusters should not be confused with the clustering of sets of graphs based on structural similarity. Fig.2 displays an example of clustering techniques with the help of distance measures between data. Cluster analysis or clustering is the charge of consortium of set of objects in such a way that objects in the same group are more similar to each other than to those in other groups (clusters). It is a foremost task of tentative data mining, and a common technique for statistical data examination, used in numerous fields.

C. Distributed Clustering

The Distributional clustering has been used to cluster words into groups based either on their participation in particular

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

grammatical relations with other words or on the distribution of class labels associated with each word. As distributional clustering of expressions result in suboptimal word clusters and high computational cost, planned a newfangled algorithm for word clustering and applied it to text classification proposed to cluster features using a special metric of distance, and followed by makes use of the of the ensuing cluster hierarchy to choose the most relevant attributes. In case of hierarchical type of clustering it can be described by clusters that form a tree. This type of clustering includes two categories such as agglomerative and divisive. But statistical data can be clustered based on distributional clustering algorithm so that user can understand even the complex data. The process can be functioned with a simple step by step procedure and grouped into single data based similar type of data elements. The most important part of clustering lies on the distance between the data elements or the key elements present in the data. The algorithm gives brief outlines of the algorithms proposed by Baker and McCallum in 1998. For simplicity we will refer to the algorithm of Distributional Clustering. This algorithm uses an alternate strategy. It uses the whole terms but maintains only n word clusters at any instant. Combining of those two clusters fallout in n−1 clusters after which a single cluster is created to get back n clusters.

Algorithm: Distributed Clustering

1. Arrange the complete terms by common Information with the class variable.
2. Initialize n single cluster with the top n words. Evaluate the inter-cluster distances between each two of a kind of clusters.
3. Loop until all words have been set into one of the clusters:

i. Combine the two clusters which are most related resultant n−1 clusters.
ii. Include a new single cluster consisting of the next word from the sorted list of words.
iii. Update the inter-cluster distances

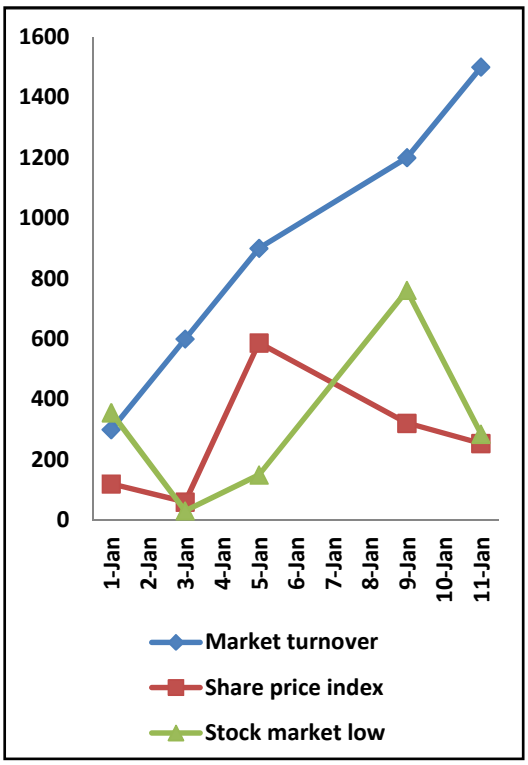

Fig.3: Visualization of clustered terms

After the process of clustering the data is visualized this is the ultimate aim of this paper that helps in understanding of the data with ease. Table.2. Shows the data stored in the form of table which is the view of data in the data warehouse and fig.3.Shows the clustered data in a visualized manner.

Table.2 Clustered data

YEAR	MARKET TURNOVER	SHARE PRICE INDEX	STOCK MARKET LOW
JAN-2001	300	120	356
JAN-2003	600	60	31
JAN-2005	900	587	150
JAN-2009	1200	321	761
JAN-2011	1500	254	285

D. Subset Selections

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. It can be evaluated from both the efficiency and effectiveness observation. Whereas the effectiveness distress the time required to find a subset of features, the effectiveness is related to the worth of the sub division of features. Based on the criteria, a fast clusteringbased feature selection algorithm (FAST) [18] is proposed and experimentally evaluated by two steps. The first step features are divided into clusters by using graph-theoretic clustering methods. The second step involves the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features Characteristics in unlike clusters are comparatively autonomous; the clustering-based strategy of FAST has a high portability of producing a subset of useful and independent features. A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, alongside with an estimation compute which scores the altered feature subsets. The most effortless algorithm defines to check each one probable subset of feature finding the one which minimises the error rate.

E. Visualization Performances

Visualization is an explore area that focus towards assist users in accepting, and investigating data all the way through progressive, iterative visual exploration. With the explosion in big data analytics, Visualization has become widely used in a variety of data analysis applications. Visualization intend extremely depends on the core data. Unusual types of data have unusual distinctiveness and patterns to visualize. A graph is a prevailing notion of data that consist of fundamentals and relations between elements. Now-a-days text documents are widely available in digital format, and have received more and more focus as getting higher visualization topic.

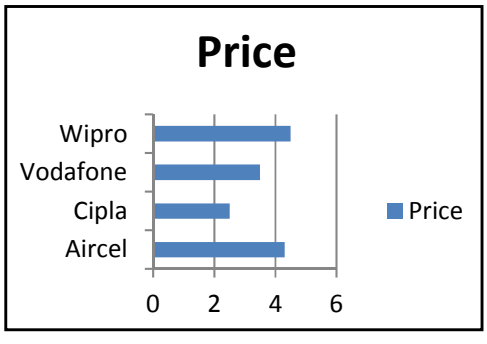

Fig.4: Visualization of price

In this section, we review and categorize recent visualization techniques based on their target Visualization is concerned with exploiting the visual perception in order to convey meaningful patterns and trends hidden in datasets. As data has gradually become more complex, new techniques are created that are more sophisticated. The visualization phase displays the visualization of extracted reports and their relationships according to the CoDe model. Fig.4 shows the visualization of price invested by different companies in the form of bar chart that visually describes the leading invested company. During this phase the designer selects the type of standard graph to draw the reports and places them in specific locations of the drawing area. Furthermore, extra data labels or ocular symbols can improve visualization details.

V. Conclusion and Future Works

In this system, we have presented a clustering-based feature subset selection algorithm for high dimensional data. The projected graphical representation technique based on the CoDe model gives as a replacement for a system to deal with the difficulty of information to be depicted. It is appeal that a suitable edge must be sited on the quantity of data and information that come out in any image, in order not to surpass the capability of the viewer. The visualization intend based on CoDe representation let the user to precisely choose information to be viewed in accordance with their consequence and interrelations. The most important aim of CoDe language is to support the independent thinking of visualization plan of composite data by providing a tool to deal with a reserved idea of information items to be depicted and associated. Specifically, we shall consider the visualization of statistical data, could also be helpful to revise an image clustering techniques and animations in the field of data mining.

References

[1] *Battiti R., Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537- 550, 1994.*

[2] *Bell D.A. and Wang, H., A formalism for relevance and its application in feature subset selection, Machine Learning, 41(2), pp 175-195, 2000.*

[3] *Cardie, C., Using decision trees to improve case-based learning, In Proceedings of Tenth International Conference on Machine Learning, pp 25-32, 1993.*

[4] *Chanda P., Cho Y., Zhang A. and Ramanathan M, Mining of Attribute Interactions Using Information Theoretic Metrics, In Proceedings of IEEE international Conference on Data Mining Workshops, pp 350-355, 2009.*

[5] *Chikhi S. and Benhammada S., ReliefMSS: a variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min. 4(3/4), pp 375-390, 2009.*

[6] *D. Iakovidis and E. Papageorgiou, "Intuitionistic Fuzzy Cognitive Maps for Medical Decision Making," IEEE Trans. Information Technology in Biomedicine, vol. 15, no. 1, pp. 100-107, Jan. 2011.*

[7] *E. Papageorgiou, "Review Study on Fuzzy Cognitive Maps and Their Applications during the Last Decade," Proc. IEEE Int'l Conf. Fuzzy Systems (FUZZ), pp. 828-835, 2011.*

[8] *F. Anfurrutia, O. Diaz, and S. Trujillo, "A Product-Line Approach to Database Reporting," IEEE Latin Am. Trans. (Revista IEEE Am. Latina), vol. 4, no. 2, pp. 70-76, Apr. 2006.*

[9] *H. Su and J. Su, "A Study and Practice of Report System Techniques Based on Three-Layer Calculating Architecture," Proc.Second Int'l Workshop Education Technology and CS (ETCS), pp. 654-657, 2010.*

[10] *J. Heer and G. Robertson, "Animated Transitions in Statistical Data Graphics," IEEE Trans. Visualization and Computer Graphics, vol. 13, no. 6, pp. 1240-1247, Nov. 2007.*

[11] *K. Hsu and M.-Z. Li, "Applying Clustering Analysis on Grouping Similar Olap Reports," Proc. Second Int'l Conf. Computer Eng. And Applications (ICCEA), pp. 417-423, 2010.*

[12] *N. Ma, M. Yuan, Y. Bao, Z. Jin, and H. Zhou, "The Design of Meteorological Data Warehouse and Multidimensional Data Report," Proc. Second Int'l Conf. Information Technology and CS (ITCS), pp. 280-283, 2010.*

[13] *N. Ma, Y. Zhai, Y. Bao, and H. Zhou, "Design of Meteorological Information Display System Based on Data Warehouse," Proc. Int'lConf. Management and Service Science (MASS), pp. 1-4, 2010.*

[14] *P. Ciuccarelli, M.I. Sessa, and M. Tucci, "Code: A Graphic Language for Complex System Visualization," Proc. Italian Assoc. for Information Systems (ItAIS), 2010.*

[15] *P. Hanrahan, "Vizql: a Language for Query, Analysis and Visualization," Proc. ACM SIGMOD Int'l Conf. Management of Data(SIGMOD), pp. 721-721, 2006.*

[16] *S. Mansmann and M.H. Scholl, "Exploring OLAP Aggregates with Hierarchical Visualization Techniques," Proc. ACM Symp. Applied computing (SAC), pp. 1067-1073, 2007.*

[17] *T. Wojciechowski, B. Sakowicz, D. Makowski, and A. Napieralski, "Transaction System with Reporting Capability in a Web-Based Data Warehouse Application Developed in Oracle Application Express," Proc. 10th Int'l Conf. CAD Systems in Microelectronics (CADSM), pp. 273-276, 2009.*

[18] *Qinbao Song, Jingjie Ni, and Guangtao Wang "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data" IEEE transactions on knowledge and data engineering, vol. 25, no. 1, january 2013.*

Author Profiles

T.P.Latchoumi received the Master degree in Computer Science in Pondicherry University, Pondicherry. Currently pursuing PhD degree in Computer Science from Sathyabama University. So far published 15 National and International Conference and journals. Her research interests include data warehouse, data mining and applications.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

P.Neeranjana pursuing Bachelor of Technology in Christ College of Engineering and Technology, Pondicherry.

D.Devi pursuing Bachelor of Technology in Christ College of Engineering and Technology, Pondicherry.

E.Suganya pursuing Bachelor of Technology in Christ College of Engineering and Technology, Pondicherry.