A Survey on Clustering Methods in Data Mining

Bhagyashree Pathak
Computer Science & Engineering,
Mody University of Science & Technology,
Lakshmangarh, Sikar, Rajasthan, India

Niranjan Lal
Computer Science & Engineering,
Mody University of Science & Technology
Lakshmangarh, Sikar, Rajasthan, India

ABSTRACT

With increasing amount of data in information industry, there may be an immense amount from claiming information accessible in the majority of the business data. This information will be of no utilization until it is changed over under suitable data. It found important to examine this immense amount of data and withdrawal meaningful knowledge from it. Data mining is an important methodology in withdrawal of meaningful knowledge from large cluster of data. Clustering is included in the tasks of data mining. Clustering is one of the task in which making a group of physical objects into classes of similar objects. In this review paper, we give a study of various clustering methods in data mining for information retrieval and other purposes. We will describe fundamental study of clustering and will analyze each methodology by doing comparative study in table format and examine the clustering algorithms for heterogeneous data.

General Terms

Clustering Techniques, k-means, k-medoids

Keywords

Data mining, Clustering, Clustering methods.

1. INTRODUCTION

Data mining is an operation on immense amount of data to unhide the patterns and extract knowledgeable information from wide amount of data sets. Mining reveals a minor set of huge pieces from lots of raw material. So running with both ―records and ―mining changed into a well known opportunity. This announcement having incredibly wonderful significance or about like data mining are getting to know extraction, information mining from facts, KDD method and instance examination. What's more, alongside these traces for discovering likenesses in statistics and accumulating uncertain records into gatherings, grouping of information is exceptionally advantageous system. Grouping venture walls information set into different lessons with the end intention that the homogeneous inside a class is bigger than that amongst classes [2].

The concept which is simple in nature and is very close to the normal people approach is data clustering, or grouping; at whatever point we are offered with a lot of information, we generally like to abridge this large data into a small sets or categories, to carry out the further study of that data. More often than not, information gathered, seem to have some inherent properties that group themselves to natural groupings, in numerous issues. Yet, to find these groups or attempt to classify the data is not a direct job for individuals till the information holds low dimensionality (two or three dimensions at maximum.) Because of this alone, soft computing methods have been proposed to conquer such issue. Such techniques that are utilized to beat the downside of the low dimensionality issue are called as the ― Data Clustering Strategies.

The developed method that we ought to take after for mining the information from gigantic databases is information accumulation to begin with, after that information cleaning is required and once the information is cleaned after that lone we ought to decipher the outcome. Another explanation behind clustering is to find importance information in information. Actualized, a Case Based Reasoning (CBR) framework based on a Growing Cell Structure (GCS) show. Information base can store information which is ordered or arranged by cases, is known as the Case Base. Cases in every gathering are relegated to some groups. Utilizing a Growing Cell Structure (GCS) information can be included or evacuated based the arrangement which is utilized. At the point when an inquiry is offered to the model, the framework will get the most relevant cases from the case construct situated in light of how inquiries and cases are near each other [5, 14].

The rest of the sections are described as follows. Section 3 introductions for basic concepts of data mining and its tasks, section 4 includes basic concepts of clustering and examples in this field. Section 5 presents research reviewed in types of clustering methods in Data mining, Section 6 has basics of clustering of heterogeneous data, and Section 7 consists of Analysis of clustering methods. Finally, Section 8 concludes this paper.

2. RELATED WORK

As determined some time recently, clustering of data is concerned with the isolation of data set into various sets to such a degree, to the point that things inside the gathering are more relative while objects of various groups are less similar. This proves that the data set to be divided into natural grouping to some extent or when the data is homogeneously dispersed method of discovering clusters of data will not succeed or will form the falsely introduced partitions. Problem that can arise is overlapping of the clusters, which can decrease the efficiency of the clustering technique and this diminish in efficiency of results in clusters overlaps. Today, it can be found that many research works have done on clustering techniques in data mining space. [1-8] offered a board view for the different clustering techniques and comparison between them considering the advantages and disadvantages of the same. Different soft computing methodologies like Fuzzy sets are generally good to handle the controversies regarding understandability of patterns, mixed media information; noisy data can provide approximate solutions faster [3]. Information excess problem can be resolve by biomedical text retrieval, which will facilitate the researchers, by hunting the analytical relationships among different pieces of extracted data, and after that we can relate the data mining algorithm to increase the effectiveness of the information retrieval system or the process[4,15]. The various algorithms are compared on the root of factors like as size of datasets, elements of cluster, variety of dataset and variety of software used. Performance, quality and accuracy of the algorithms are discussed with respect to all these four factors [5]. Reference [1] has given a common

schematic of the architecture of participation anticipating system in presidential election by using KNN, Classification Tree and Naïve Bayes and tools orange based on crisp and by anticipating the political behavior of people in elections can determine the future prospect of each country domestic and foreign policies and characterize domestic and international relationships. Overall four opinions are considered and specified in point and those are knowledge resource; knowledge types and/or knowledge datasets; data mining tasks; and data mining tasks and applications used in data running and beside with this terms data mining functionalities are surveyed [2]. Graham Cormode has introduce their idea on the key concepts in data mining along with that he has discussed the things like as clustering problems and some of the other clustering techniques. As tremendous amount of work is being done on the clustering methods in data mining space , Among those fuzzy c-means is nothing but the famous strategy of clustering and the reference[1] suggests that the fuzzy c-means (FCM) algorithm is important tool for clustering real dimensional data, yet we cannot apply it to the incomplete data. In these clustering techniques, a soft computing technique play a extremely significant role and it is because outstanding to the reality that soft computing strategies allows the pattern to belong to more than 1 cluster. Reference [13] put forward the significance of soft computing together with fuzzy logic (FL), artificial neural networks (ANNs), genetic algorithms (GAs), and (RS), rough sets highlighted. A study of the prose on ―soft web mining, (which is accessible), is offered with the commercially accessible systems. The word data clustering is ultimately refers to the text clustering and the reference [14] tells that the conceptual text clustering extends to web ID, containing diverse markup language structure linked with the documents. Based on the markup languages such as presentations, procedural and descriptive markup and the web document's text clustering is done efficiently using the concept-based mining model.

3. DATA MINING

With the enormous measure of information put away in records, databases, and different warehouses, it is progressively vital, if a bit much, to grow intense means for examination and perhaps defining of such information and for the extraction of fascinating patterns that could help in decision-making.. Data mining is one of the processes in entire knowledge discovery phenomena of information retrieval. Data Mining, also well liked is known as Knowledge Discovery in Databases (KDD) refers to the significant extraction of unexpressed, previously unknown and prospectively meaningful information from huge data in databases. Both data mining and knowledge discovery in databases (or KDD) are often regarded as equivalent words, data mining is core of the knowledge discovery process. The following Fig.1 shows knowledge discovery iterative process including working of data mining. The said functionalities are measured to perceive the type of patterns to be found in data mining tasks, Data Mining tasks can be categorized into two categories.

1). Descriptive Task: These tasks present the general properties of data stored in database. The descriptive tasks are used to find out patterns in data i.e. cluster, correlation, patterns and inconsistency etc. and 2). Predictive Tasks: Predictive tasks of data mining anticipate the evaluation of one property on the bases of estimations of different properties, which is identified as target or reliant variable and the attributes used for building the prediction are known as independent variables. The important Data mining techniques are prediction, classification, clustering Mining Frequent patterns, Associations and

correlations and outliner analysis. We are working one of its functionality i.e. clustering.
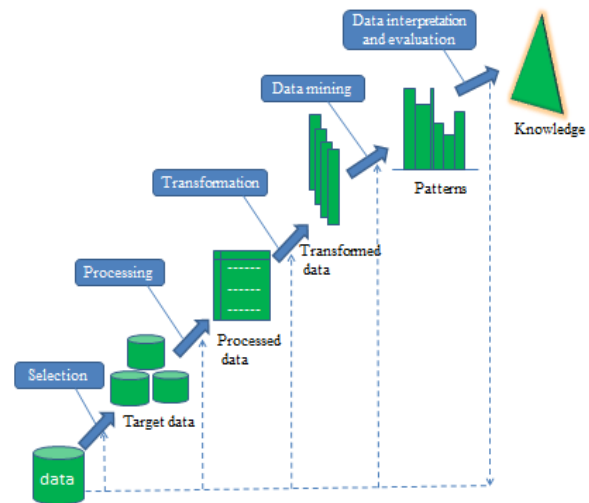

Fig. 1 Data mining: the key factor in Knowledge discovery process

4. BASIC CONCEPT OF CLUSTERING

Cluster is a collection of objects that belongs to the same class, where the objects with similarity put (or related) to one another within the same group and dissimilar (or unrelated) to the objects in other groups.

Cluster analysis is operation on large collection of data which discovers likenesses between information as per the attributes discover in the information and gathering comparative information objects into clusters. An efficient clustering method will deliver high quality clusters. We classified two main concepts on similarity that is high intra-class similarity have cohesive within clusters and low inter-class similarity have distinctive between clusters. The nature of a grouping result relies on upon both the similitude measure utilized by the strategy and its usage. The nature of a grouping strategy is likewise measured by its capacity to find a few or the majority of the inaccessible patterns. In information retrieval (IR), cluster analysis has been used to create groups of documents with the goal of improving the coherence and potency of retrieval, or to determine the semantic of the literature of a field. The terms in a document collection can also be clustered to show their relationships

There are two primary types of cluster analysis methods are the non-hierarchical(partitioning method, density based method, grid based method, soft computing and model based method), which splits a data set of X items into Y clusters, and the hierarchical, which produce tree structured data set in which pairs of objects or clusters are progressively connected. By Comparing nonhierarchical methods with hierarchical methods it is analyze that non hierarchical methods are analytical in nature and require less calculation. Clustering is for unsupervised learning (some clustering models are for both). The aim of clustering is descriptive (mostly applied on descriptive applications).
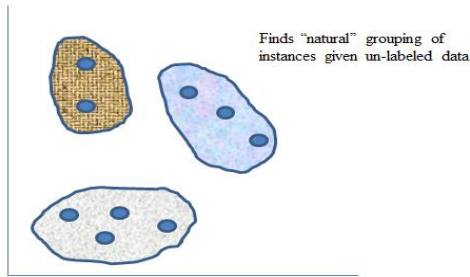
Fig.2 A view of clusters in graph form

5. OVERVIEW OF TYPES OF CLUSTERING METHODS

There are diversities in the kinds of clustering techniques are reviewed in following section.

a). Hierarchical methods

b). Partitioning methods

c). Density based clustering

d). Grid based clustering

e). Model based clustering

f). Soft computing clustering

5.1 Hierarchical methods

These strategies develop the groups by periodically dividing the samples in either a downward or upward form. This strategy can be subdivided as following:

Agglomerative hierarchical clustering: All items at first have a place with one bunch. Then these bunches are progressively combined until the needed cluster structure is achieved.

Divisive hierarchical clustering: initially All objects are put together in one bunch. Then the bunch is splits into sub-clusters, which are progressively divided into their own sub-clusters. This operation proceeds until the required group structure is accomplished. The consequence of the hierarchical techniques is in form of cluster dendrograms, the representing of objects in form of nested grouping and similarity levels at which groupings change. A clustering of the information objects is obtained by cutting the tree at the needed similarity level. The hierarchical clustering techniques could be further separated by way that the closeness measure is computed.

- Single-link clustering (likewise called the connectedness, the base strategy or the nearest neighbour technique) — strategies that consider the separation between two clusters to be equivalent to the shortest separation from any individual from one group to any individual from the other group. In the event that the information comprise of similitude, the closeness between a couples of clusters is thought to be equivalent to the best comparability from any individual from one group to any individual from the other group.

- Complete-link clustering (also named as the diameter, the maximum method or the furthest neighbour method) - techniques that consider the separation between two groups to be equivalent to the longest separation from any individual from one cluster to any individual from the other cluster.

- Average-link clustering (also called minimum variance method) – strategies that consider the separation between two bunches to be equivalent to the average separation from any individual from one group to any individual from the other group.

5.2 Partitioning methods

Partitioning methods relocate instances by moving them from one group to another, starting from an initial partitioning. Such techniques commonly require that the quantity of groups will be pre-set by the client. To accomplish global optimality in partition based clustering, a comprehensive identification procedure of every conceivable segment is required, because this is not feasible, certain greedy heuristics are used in the pattern of iterative optimization. Namely, a relocation method recursively moves points between the x clusters. The following subsections present various types of partitioning methods.

5.2.1 Error Minimization Methods

The algorithms, who work on this principle can actively work with remote and compact clusters, are the most intuitive and frequently used methods. The essential thought is to discover a bunching structure that minimizes specific error criteria which measures the "separation" of every sample to its delegate esteem. The most renowned criterion is the Sum of Squared Error (SSE), which measures the total squared Euclidian distance of samples to their characteristic values. SSE may be globally optimized by exhaustively enumerating all partitions, which is extremely time-consuming, or by offering an estimated result (not essentially foremost to a global minimum) using heuristics. The modest and most regularly used algorithm, employing a squared error criterion is the K-means algorithm.

5.2.2 Graph Theoretic Clustering

Graph theoretic strategies are those methods that deliver clusters via histogram. The sides of the histogram connect the samples denoted as nodes. A well-known graph-theoretic algorithm is work on the Minimal Spanning Tree—MST Inconsistent edges are edges whose weight (in the case of clustering-length) is significantly greater than the average of nearby edge lengths. Another graph-theoretic approach constructs graphs based on limited neighbourhood sets. There is also a relation between hierarchical methods and graph theoretic clustering.

- Single-link clusters are subgroups of the MST of the data cases. Each sub-graph is a linked element, namely a set of instances in which each instance is linked to minimum one of the other element set, so that the set is maximal as per this property. These subgroups are framed according to some similarity threshold.

- Complete-link clusters are maximal complete sub-graphs, framed using a similarity threshold. A maximal complete sub-graph is a sub-graph so that each node is connected to every other node in the sub-graph and the set is maximal according to this property.

5.3 Density Based Methods

Density-based methods accept the points that have a place with every group are drawn from a particular probability distribution. The general allotment of the data is accepted as a mixture of several distributions. The main point of these strategies is to recognize the clusters and their distribution parameters. These techniques are intended for finding groups

of self-assertive shape which are not really raised. The thought is to keep developing the given bunch the length of the thickness (number of articles or information points) in the area surpasses some limit. In particular, the area of a given sweep needs to contain no less than a base number of items. At the point when every group is portrayed by neighbourhood mode or maxima of the thickness work, these techniques are called mode-seeking.

5.4 Grid-based Method

These techniques divide the space into a countable number of cell that form a grid structure on which all of the operations for clustering are performed.

5.5 Model Clustering Methods

These strategies endeavour to advance the fit between the given information and some numerical models. Dissimilar to customary clustering, which recognizes gatherings of articles; model-based clustering methods additionally discover trademark portrayals for every gathering, where every gathering speaks to an idea or class. The most every now and again utilized enlistment techniques are choice trees and neural systems.

5.5.1 Decision tree

In decision trees, the data will be represented in the form of hierarchical tree, where every leaf refers to an opinion and holds a probabilistic explanation of that concept. Several algorithms produce classification trees for representing the unlabelled data.

5.5.2 Neural Network

This type of algorithm represents each cluster by a neuron or "prototype". The input data is also represented by neurons, which are linked to the prototype neurons. Each such connection has a Weight, which is learned adaptively during learning.

5.6 Soft computing Clustering

This fragment promotes the essential utility of other soft-computing methods in clustering tasks.

5.6.1 Fuzzy Clustering

In this clustering, each pattern is associated with every cluster using some sort of membership function, namely, every cluster called as fuzzy set of all the patterns. Larger membership values indicate higher confidence in the assignment of the pattern to the cluster. A hard clustering can be acquired from a fuzzy partition by using a threshold of the membership value. The most prominent fuzzy clustering algorithm is the fuzzy c-means (FCM) algorithm. Because it is better one than hard K-means algorithm at avoiding (local minima).

5.6.2. Evolutionary technique

Evolutionary techniques are stochastic broadly useful methods for solving optimization problems. The concept is to use evolutionary administrators and a populace of clustering structures to converge into a globally optimal clustering. Candidate clustering are encoded as chromosomes. The most generally used evolutionary operators are: selection, recombination, and mutation. The most frequently used

evolutionary technique in clustering problems is genetic algorithms (GAs). A fitness value is involved with each clusters structure. A higher fitness value indicates a better cluster structure.

5.6.3 Stimulated annealing for clustering

Another general-purpose stochastic search technique that can be used for clustering is simulated annealing, which is a sequential stochastic search technique designed to avoid local optima. This is accomplished by accepting with some probability a new solution for the next iteration of lower quality. The probability of acceptance is administered by a critical parameter called the temperature, which is typically specified in terms of a starting and final temperature value. SA is statistically insured to find the global optimal solution.

6. CLUSTERING METHODS FOR HETEROGENEOUS DATA

Now-a-days, we all are working with large data which is heterogeneous in nature say HTML, XML, web pages, audio, video etc. Heterogeneous data includes three main classifications that are structured data, semi-structured data and unstructured data. There are some different types of algorithms for clustering of distributed heterogeneous data so far it is difficult but possible. some important algorithms are reviewed that are K-means clustering algorithm, hybrid clustering algorithms(K-medoids algorithm) and C-means clustering algorithm(mainly used in soft computing clustering applications).K-means algorithm is also called centroid-based clustering and also popular in machine learning. The *K-medoids* algorithm is a clustering algorithm which is also belongs to *K-means* algorithm. Both the K-means and K-medoids algorithm are separating the dataset into sets and both endeavor to minimize the separation between points named to be in a cluster and a point assigned as the center of that group. Rather than the K-means algorithm, K-medoids picks data points as points and works with self-assertive measurements of separations between information points. Usual *K-means* cannot cluster heterogeneous data with numerical and non-numerical features appropriately because the Euclidean distance between non-numerical values cannot be calculated. To crack this problem, numerous centroid-based clustering algorithms have been projected, such as *k*-prototypes, enhanced *k*-prototypes and KL-FCM-GM, but these still have a few drawbacks. Clustering of semi structured data sets can be completed using algorithm known as UTF-K-means so for reliable results integrated UFT-*K-means* is a most favorable choice for clustering heterogeneous data with numerical and non-numerical features.

7. INTERPRETATIONS ON CLUSTERING METHODS

In the investigation of clustering methods, we consider distinctive strategies for clustering; below table shows relative investigation of clustering methods with their advantages and disadvantages.

Table 1: Comparison between clustering methods

S.no.	Methods	Sub-methods	Advantages	Disadvantages
1.	Hierarchical clustering method	✓ Single-link clustering ✓ Complete-link clustering ✓ Average-link clustering	Versatility- The single-link methods, e.g., maintain good performance on groups of data containing non-identical clusters. Well separated, Chain-like and concentric clusters. Multiple partitions - hierarchical methods generate not only one partition, but multiple nested partitions, which permit different users to opt different partitions, as per the preferred similarity level.	- Inability to scale well- The time complexity of hierarchical algorithms is at least O(m2) (where m is the total number of instances), which is Non-linear with the number of objects. - Hierarchical methods can never undo what was done formerly. Namely, There is no back-tracking capability. - Hierarchical algorithm is additionally portrayed by huge I/O Costs.
2.	Partitioning clustering method	✓ Error Minimization Algorithms (K-means algorithm) ✓ Graph-Theoretic Clustering	- This algorithms have linear complexity - ease of analysis, - straightforwardness in implementation, - rapidity of convergence and - flexibility to sparse data - K-means algorithm uses the selection of the initial partition.	- Its processing is added costly - It requires more estimation Effort. - K-means cannot represent density-based clusters - Graph theoretic methods include methods that carry clusters via charts only.
3.	Density based clustering method	✓ Expectation maximization algorithm	- Discovers clusters of random shapes and is efficient for large structural Databases. - This algorithms explore for clusters by searching the neighbourhood of each object in the database and ensure if it holds more than the least number of objects. - Density-based clustering can also utilize non parametric methods, such as searching for cases with huge counts in multi- dimensional circular graphs.	- The region around the given radius of a given cluster has to contain at least a minimum number of points.
4.	Grid based clustering method	_____	- The major advantage of this method is less processing time. - It is dependent only on the number of cells in each part in the quantized space. - The algorithm's complexity is for low size of data.	- Algorithms Complexity exponentially grows with the elements. - Data have order dependency f - well-built dependency due to clusters initialization - reliance on parameters (threshold used as initialization)
5.	Model clustering method	✓ Decision tree ✓ Neural network	-Find the best fit of data for a given model. -It is taking outlier or noise into account and compute. -yields robust clustering methods	- Mostly work with mathematical models - This algorithms are not appropriate for the clustering of large database data

| 6. | Soft computing clustering | ✓ Fuzzy clustering c-means algorithm
✓ Evolutionary Approaches for Clustering genetic algorithms (GAs)
✓ Simulated annealing for clustering | - These are Cost Function Optimization Algorithms | - have high computational costs that limit their application in data mining.

- Usage of combined strategies (e.g., making of initial presume for k-means) has been attempted.

- Not produce good results on definite problems like pattern clustering. |

8. CONCLUSION

We have investigated, the five popular data clustering algorithms in data mining are specifically hierarchical clustering, partitioning clustering, density based clustering(expectation maximization algorithm),grid based clustering, model based clustering and the soft computing methods. Soft computing techniques contain genetic algorithms, simulated annealing, fuzzy clustering; neural networks are recently useful to resolve the issues in data mining from large databases. They attempt to provide us the low cost results, and hence the procedure will speed -up. Data mining is the extremely high-quality area for learning the clustering algorithms. From the paper it is observed that the hierarchical algorithms can achieve better results than the partitioning algorithms.

9. ACKNOWLEDGMENT

10. REFERENCES

[1] Chuan Shi, Member, IEEE, Yitong Li, Jiawei Zhang, Yizhou Sun, Member, IEEE, and Philip S. Yu, Fellow, IEEE "A Survey of Heterogeneous Information Network Analysis" , IEEE transactions on knowledge and data engineering.

[2] Yizhou Sun, Jiawei Han, Xifeng Yan and Philip S. Yu,"Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach" 38th International Conference, August 27th - 31st 2012, Vol. 5, No. 12

[3] Min Wei, Tommy W. S. Chow and Rosa H. M. Chan, "Clustering Heterogeneous Data with K-means by Mutual Information-Based Unsupervised Feature Transformation" Entropy 2015, 17, 1535-1548.

[4] Y. Sun and J. Han, "Mining heterogeneous information networks: a structural analysis approach," SIGKDD Explorations, vol. 14, no. 2, pp. 20–28, 2012.

[5] Y. Sun and J. Han, "Mining Heterogeneous Information Networks: Principles and Methodologies". Morgan & Claypool Publishers, 2012.

[6] Y. Sun and J. Han, "Meta-path-based search and mining in heterogeneous information networks," Tsinghua Science and Technology, vol. 18, no. 4, pp. 329–338, 2013.

[7] Amin Babazadeh Sangar, Seyyed Reza Khaze and Laya Ebrahimi, "participation anticipating in elections Using data mining methods", International Journal on Cybernetics & Informatics, Vol.2, No.2, April 2013.

[8] Tipawan Silwattananusarn and Dr. Kulthida Tuamsuk, ―Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012‖, International Journal of Data Mining & Knowledge Management Process, Vol.2, No.5, 2012.

[9] Sushmita Mitra, Sankar Pal, and Pabitra Mitra, "Data Mining in Soft Computing Framework :A Survey", IEEE transactions on neural networks, vol. 13, no. 1, 2002.

[10] Sumit Vashishta, Dr. Yogendra Kumar Jain, "Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm", International Journal of Advanced Computer Science and Applications, Vol. 2, No. 4, 2011.

[11] Osama abuabbas, "Comparision between data clustering algorithms", International arab journal of information technology, vol. 5, No. 3, july 2008.

[12] Richard J. Hathaway, and James C. Bezdek, " Fuzzy c-Means Clustering of Incomplete Data", IEEE transactions on systems, man, and cybernetics—part b: cybernetics, vol. 31, no. 5, 2001.

[13] Sankar K. Pal, Varun Talwar, and Pabitra Mitra, "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions", IEEE transactions on neural networks, vol. 13, no. 5,2002.

[14] Niranjan Lal, Monika Kalra "Data mining of heterogeneous data with research challenges" IEEE, Colossal Data Analysis and Networking (CDAN), Symposium on 18-19 March 2016.

[15] Niranjan Lal , Samimul Qamar , Savita Shiwani "Search Ranking for Heterogeneous Data over Dataspace" Indian Journal of Science and Technology, Vol 9(36) , September 2016.

[16] Bhagyashree Pathak, Niranjan Lal "mining of unstructured data with clustering approach", international journal of engineering reasearch and management Technology, volume 3, issue-6, november-2016.