International Journal of Computer Science and Informatics

Volume 2 | Issue 4 Article 13

April 2013

Clustering Algorithms For High Dimensional Data – A Survey Of Issues And Existing Approaches

B.Hari Babu
Department of CSE, NOVA College of Engg.& Tech., Jangareddygudem, AP, India,
hari.sreenu4u@gmail.com

N.Subash Chandra
Department of CSE, Holy Mary Inst. of Tech. & Science, Bogaram (v), Keesara (M), R.R Dist.,
subhashchandra_n@yahoo.co.in

T. Venu Gopal
Department of CSE, JNTUH College of Engineering, Jagithyal,Karimnagar , AP, INDIA,
t_vgopal@rediffmail.com

Recommended Citation

Babu, B.Hari; Chandra, N.Subash; and Gopal, T. Venu (2013) "Clustering Algorithms For High Dimensional Data – A Survey Of Issues And Existing Approaches," *International Journal of Computer Science and Informatics*: Vol. 2 : Iss. 4 , Article 13.
DOI: 10.47893/IJCSI.2013.1108
Available at: https://www.interscience.in/ijcsi/vol2/iss4/13

Clustering Algorithms For High Dimensional Data – A Survey Of Issues And Existing Approaches

B.Hari Babu[1], N.Subash Chandra[2] & T. Venu Gopal[3]

[1]Department of CSE, NOVA College of Engg.& Tech., Jangareddygudem, AP, India
[2]Department of CSE, Holy Mary Inst. of Tech. & Science, Bogaram (v), Keesara (M), R.R Dist.
[3]Department of CSE, JNTUH College of Engineering, Jagithyal,Karimnagar , AP, INDIA
E-mail : hari.sreenu4u@gmail.com, subhashchandra_n@yahoo.co.in, t_vgopal@rediffmail.com

Abstract - Clustering is the most prominent data mining technique used for grouping the data into clusters based on distance measures. With the advent growth of high dimensional data such as microarray gene expression data, and grouping high dimensional data into clusters will encounter the similarity between the objects in the full dimensional space is often invalid because it contains different types of data. The process of grouping into high dimensional data into clusters is not accurate and perhaps not up to the level of expectation when the dimension of the dataset is high. It is now focusing tremendous attention towards research and development. The performance issues of the data clustering in high dimensional data it is necessary to study issues like dimensionality reduction, redundancy elimination, subspace clustering, co-clustering and data labeling for clusters are to analyzed and improved. In this paper, we presented a brief comparison of the existing algorithms that were mainly focusing at clustering on high dimensional data.

Keywords:- *High dimensional data, Dimensionality Reduction, Redundancy Reduction, Subspace Clustering, Co-Clustering.*

I. INTRODUCTION

In recent years, the dramatic rise in the use of the web and the improvement in communications in general have transformed our society into one that strongly depends on information. The huge and amount of data that is generated by this communication process contains important information that accumulates daily in databases and is not easy to extract. The field of data mining developed as a means of extracting information and knowledge from databases to discover patterns or concepts that are not evident. So, it is esteemed that there is a mounting need for a more sophisticated automated system of partitioning the datasets into groups, or clusters [1]. Clustering is defined as the process of finding a structure where the data objects are grouped into clusters which are similar behavior". For example, as digital libraries and the World Wide Web are growing exponentially, the ability to find useful information progressively depends on the indexing infrastructure or search engine. Clustering techniques can be used to discover natural groups in data sets and to identify a structure that might reside there, without having any specific background knowledge as characteristics of the data [1]. Clustering has been used in a variety of areas, including computer vision, VLSI design, psychology, data mining, bioinformatics, statistics, pattern recognition, machine learning and information retrieval.

Clustering can be considered as the most important unsupervised learning problem. Clustering deals with finding a primitive structure in a collection of unlabeled data. A cluster a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters [2]. The objective of the clustering technique is to determine the intrinsic grouping in a set of unlabeled data. The similarity between data objects can be measured with the imposed distance values. Specifying the distance measures for the high dimensional data is becoming very trivial because it holds different data values in their corresponding attributes. Following is the analysis of different distance measures used for measuring similarity between data objects in clustering.

Distance Measure: Most of the clustering techniques relay on distance measure as an important step for selecting data objects, which will determine the similarity between two elements [2]. The cluster shape

or density will be influenced by the similarity between the data objects [3], as some elements may be close to one another according to one distance and farther away according to another. In general there are two types of distance measures. 1) Symmetric measure and 2) Asymmetric measure. The common distance measures used in the clustering process [3][4] are *i)* The Euclidean distance or Squared Euclidean distance, *ii*)The Manhattan distance , *iii)* The Maximum Likelihood Distance, iv)The Mahalanobis distance,*(v)* The Hamming distances ,(vi) The angle between two vectors used as a distance measure when clustering high dimensional data.

II. ANALYSIS OF HIGH DIMENSIONAL DATA FOR CLUSTERING

The rapid growth in various new application domains, like bioinformatics and e-commerce, reflects the need for analyzing high dimensional data. Many organizations have massive amounts of data containing valuable information for running and building a decision making system. To do this, it makes study and to analyze high dimensional and large amount data for effective decision making. Generally, in a gene expression microarray data set, there could be tens or hundreds of dimensions, each of which corresponds to an experimental condition. Researchers and practitioners are very eager in analyzing these data sets. However, before analyzing the data mining models, the researcher will analyze the challenges of attribute selection, the curse of dimensionality, redundancy reduction, data labeling and the specification of similarity in high dimensional space for analyzing high dimensional data set.

In data mining, the objects can have hundreds of attributes or dimensions. Clustering in such high dimensional data spaces presents a tremendous difficulty, much more so than in predictive learning. In clustering, however, high dimensionality presents two problems.

1) The clustering tendency will lose when the dataset contains irrelevant attributes [5]. Searching for clusters is a hopeless enterprise where there are no relevant attributes for finding clusters. Attribute selection is the best approach to address the problem of selecting irrelevant attributes.

2) Dimensionality curse is another problem in high dimensional data. As the number of attributes or dimensions increases in a dataset, the distance measures will become increasingly meaningless [6] [7]. The resultant clusters with high dimensions; they are equidistant from each other.

A. Dimensionality Reduction

Dimensionality curse is a loose way of speaking about lack of data separation in high dimensional space [7], [6], and [8]. The complexity of many existing data mining algorithms is exponential with respect to the number of dimensions [7]. With increasing dimensionality, these algorithms soon become computationally intractable and therefore inapplicable in many real applications.

In general, there are two approaches that are used for dimensionality reduction. One is attribute Transformation and another one is attribute Decomposition. Attribute Transformations are simple function of existent attributes. For example, the sales profiles and OLAP-type data, rollup as sums or averages over time intervals can be used. In multivariate attribute selection can be carried out by using Principle Component Analysis (PCA) [9]. Attribute Decomposition is a process of dividing data into subsets. Using some similarity measures, so that the high dimensional computation over smaller data sets will happen [10]. Dimensions stay the same, but the costs are reduced. This approach targets the situation of high dimensions, large data. It was proven that, for any point in a high dimensional space, the expected gap between the Euclidean distance to the closest neighbor and that to the farthest point shrink as the dimensionality grows [8]. This phenomenon may render many data mining tasks ineffective and fragile because the model becomes vulnerable to the presence of noise.

An Adaptive dimension reduction for clustering, a new semi-supervised clustering framework based on feature projection and fuzzy clustering is proposed for clustering high dimensional data [11]. In this proposed model, the standard practice of reporting the results directly obtained in the reduced-dimension subspace is not accurate enough.

B. Sub Space Clustering

Subspace clustering is the task of detecting clusters in all subspaces [12]. A data object may be a member of multiple clusters, each one existing in a different subspace. In general the Subspace clustering techniques involve two kinds of approaches. One is projection pursuit clustering assumes that the class centers are located on same unknown subspaces. The other hand, principal component clustering assumes that each class is located on an unknown specific subspace.

The problem with subspace clustering is given by the fact that there are 2d different subspaces of a space with d dimensions. If the subspaces are axis-parallel, a finite number of subspaces are possible. Hence, subspace clustering algorithm utilizes a kind of

heuristics to remain computationally feasible, at the risk of producing inferior results. Most of the traditional algorithms for subspace clustering are CLIQUE [13] and SUBCLU [14]. An innovative algorithm, called "Non Redundant Subspace Cluster mining" NORSC, to efficiently discover a succinct collection of subspace clusters while also maintaining the required degree of data coverage [15]. But this proposed model limited to discover the solution for information overlapping in subspace clustering.

C. Co-Clustering

Simultaneous clustering of data points and their attributes is Co-clustering [16]. Grouping attribute in conjunction with clustering of data points themselves is called co-clustering. Co-clustering will improve clustering of points based on their attributes. It tries to cluster attributes based on their points. Grouping rows and columns into point-by-attribute data representation is the concept of co-clustering.

Co-clustering of data points and attributes is proposed by [16]. Co-clustering is also known as simultaneous clustering, conjugate clustering, distributional clustering, bi-dimensional clustering, block clustering, and information bottleneck method. Co-clustering on categorical data [17] is prominent area for research now a day. In the Co-Occurrence of Categorical Data the similar way of building groups of items was presented. A new mechanism named Maximal Resemblance Data Labeling (MARDL) [18]. However this proposed model delivering quality is relatively proportional to quantity of the dataset.

III. CLUSTERING ALGORITHMS FOR HIGH DIMENSIONAL DATA

The main aspiration of clustering is to find high quality clusters within reasonable amount of time. Clustering in data mining is the process of discovering groups. Each group is a dataset such that the similarity among the data inside the group is maximized and the similarity in outside group is minimized. The discovered clusters are then used to explain the characteristics of the data distribution. To day there is tremendous necessity in clustering the high dimensional data. For example, many business applications, clustering can be used to describe different customer groups and allows offering customized solutions. Clustering can be used to predict customer buying patterns based on their profiles to which cluster they belong. In the following section 4 presents various types of clustering algorithms used for two dimensional data space and section 5 represents types clustering algorithms high dimensional data space.

IV. TYPES OF CLUSTERING ALGORITHMS FOR TWO- DIMENSIONAL DATA SPACE

From the broad variety of evaluations in the data base community and applications, the clustering algorithms for two dimensional data space are specified below. The general categories of the clustering algorithms are listed below.

A. K-means clustering algorithm

The traditional clustering algorithm is the k-means algorithm [2]. In k-means it assigns each point to the cluster which is nearer to the center called centroid. The center is the average of all the points in the cluster that means the coordinates are the simple arithmetic mean for each dimension separately over all the points in the cluster.

Simplicity and speed is the main advantage of this algorithm. It also allowed running on large datasets. The disadvantage is that at each run it does not produces the same result, since the resulting clusters depend on randomly initialized assignments. The problem by seeking to choose better starting clusters is addressed by k-means++ [19]. The intra-cluster variance is minimized, but it does not sure about minimizing global variance. Another disadvantage is the requirement of mean to be definable for the concept which the case is not always. For such datasets the variant of k-mean is k-medoid is appropriate. A different criterion for which points are best assigned to which centre are k-medians is clustering.

B. Hierarchical Clustering Algorithms

Hierarchical clustering builds cluster hierarchy or it's a tree of clusters. It finds successive clusters using previously established clusters. These algorithms can be agglomerative or divisive [20] [2]. Agglomerative hierarchical clustering is a bottom-up clustering. It begins with each element as a separate cluster and merges them into successively larger clusters. Divisive hierarchical clustering is top-down clustering. It's clustering starts with everybody in one cluster and ends up with everyone in individual clusters. Divisive algorithms begin with a set and keep on dividing it into successively smaller clusters. CURE and CHAMELEON are better known hierarchical clustering algorithms [20].

C. Partitioning Relocation Clustering algorithms

Partitioning algorithms divides the data into several data sets. It uses a greedy heuristic as in the form of iterative optimization. Different relocation schemes can be iteratively used to reassign points between k clusters. It typically determines all clusters at once. The hierarchical clustering can be used as divisive algorithms. The common variants of this of algorithms

are probabilistic clustering [21] such as EM framework, MCLUST, k-means and k-medoid methods [20] where a more detailed explanation can be found.

D. Density-Based Clustering Algorithms

The data space is divided into a set of its connected components. The basic idea for partitioning into sets requires a concept of density. A cluster defined as a dense component, where it can grow in any direction that density leads. These are devised to discover arbitrary-shaped clusters. In this approach, a cluster is considered as a density region in which the data objects exceeds a threshold. Since it requires space as metric for clustering Density based algorithms are also called as Spatial Data Clustering. There are two approaches. The first approach is a density to a training data point like DBSCAN [22] and OPTICS [23]. The second approach is a density to a data point in the attribute space uses a density function like DENCLUE [2].

E. Conceptual Clustering Method

A machine learning approach for unsupervised classification is Conceptual clustering [24]. It generates a concept description which distinguishes it from ordinary data clustering. Conceptual clustering methods are capable of generating hierarchical category schemes. Hence it is closely related to decision tree learning, and mixture model learning. COBWEB [25] is the most common Concept clustering technique, which uses the incremental concept clustering for generating object description and producing observations yields good predictions. It is also called as formal concept analysis. It is a technique for generating clusters for objects and attributes [26]. Concept analysis uses a bipartite graph representing the relation between the set of objects and attributes. A method for generating overlapping clusters rather than partitions in discussed in [27].

F. Quality Threshold Clustering Algorithm

Quality Threshold clustering algorithm is an alternative method of partitioning data, particularly invented for gene clustering [28]. It requires excessive computing power than k-means, and does not require the number of clusters in advance. But it always returns the same result as it runs for several times. In QT-Clust algorithm [28], the distance between a point and a group of points is computed using complete linkage, which is the maximum distance from the data point in a group to any member of the group.

G. Spectral clustering

Spectral clustering method used for reducing the dimensionality in a given data space. In a given data set A, the similarity matrix can be defined as S where Sij represents a similarity matrix between data points. Spectral clustering techniques will use the similarity matrix of the data to reduce dimensionality to fewer dimensions suitable for clustering. One such spectral clustering technique is the Normalized Cuts algorithm [29]. This technique is commonly used for image segmentation. This algorithm partitions data points into two sets S1 and S2. These S1 and S2 are selected based on the Eigenvector values corresponding to the second-smallest Eigen value of the Laplacian matrix. This partitioning may be done by taking the median m of the data points in value, and placing all points in S1 which is greater than m, and the rest in S2. The algorithm along with hierarchical clustering can be used to repeatedly partition the subsets in this style.

V. TYPES OF CLUSTERING ALGORITHMS FOR HIGH- DIMENSIONAL DATA SPACE

In this section, we describe some of the clustering algorithms for High Dimensional data space. These are specific and need more attention because of high dimensionality. To day, most of the research work is carrying under this. Due to high dimensionality it is becoming tedious and needs more generalized techniques to cluster various dimensions of the data. Due its dimensionality, there is a need for dimensionality reduction and redundancy reduction at the time of clustering. This section discusses the main subspace clustering and projected clustering strategies and summarizes the major subspace clustering and projected algorithms.

A. Subspace Clustering

Subspace clustering methods will search for clusters in a particular projection of the data [12]. These methods can ignore irrelevant attributes and also problem is known as Correlation clustering. Two-way clustering, or Co-Clustering or Biclustering are known as the special case of axis-parallel subspaces. In these methods the objects are clustered simultaneously as the feature matrix consisting of data objects as they are span in rows and. As in general subspace methods they usually do not work with arbitrary feature combinations. But this special case it deserves attention due to its applications in bioinformatics. The classification of subspace clustering is shown in the following figure.1

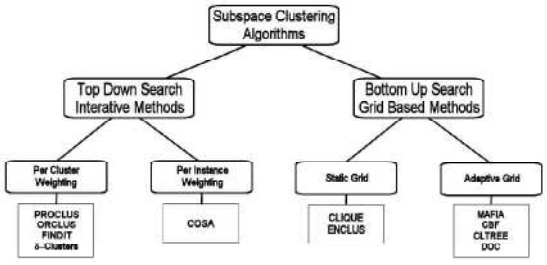

Fig. 1 : Hierarchy of Subspace Clustering Methods

The above figure will represents the subspace clustering. These are distinguished based on their search strategy. It uses both top-down and bottom-up iterative process for clustering the subspace data points [12]. The following section will better describe existing techniques along with the projected clustering approach.

Subspace clustering is the method of detecting clusters in all subspaces in a dimensional space. There are multiple clusters consists of a point as a member, and each cluster exists in a different subspace. The term can be used synonymous with general clustering in high-dimensional data. In a two-dimensional space the number of clusters can be identified. The one-dimensional subspaces, the clusters in subspace {X} and in subspace {Y} can be found. There are different 2D subspaces each with a space of D dimension is the major problem of subspace clustering. If the subspaces are axis-parallel, a finite number of subspaces are possible. If any subspace in a cluster can be found in a full space that contain the cluster. This approach taken by most of the traditional algorithms is Clique [13] And Subclu [14], Enclus [30], Mafia [31].

CLIQUE-Clustering in QUEst [13], is the fundamental algorithm used for numerical attributes for subspace clustering. It starts with a unit elementary rectangular cell in a subspace. If the densities exceeds the given threshold value, those cell are will be retained. It applies a bottom-up approach for finding such units. First, it divides units into 1-dimensional equal units with equal-width bin intervals as grid. Threshold and bin intervals are the inputs for this algorithm. It uses Apriori-Reasoning method as the step recursively from q-1-dimensional units to q-dimensional units using self-join of q-1. The total subspaces are sorted based on their coverage. The subspaces which are less covered are pruned. Based on MDL principle a cut point is selected and a cluster is defined as a set of connected dense units. A DNF expression that is associated with a finite set of maximal segments called regions is represented whose union is equal to a cluster.

A. Projected Clustering

Projected clustering tries to assign each point to a unique cluster, but the clusters may exist in different subspaces. The general approach uses a special distance function along with a regular clustering algorithm. PROCLUS -Projected Clustering, [32], is associates with a subset of a low-dimensional subspace S such that the projection of S into the subspace is a tight cluster. The pair (subset, Subspace) will represent a projected cluster. The number of clusters k and average subspace dimension *n* will be specified by the user as inputs. It finds k-medoid in iterative manner and each medoid is associated with its subspace. A sample of data is used along with greedy hill-climbing approach and the Manhattan distance divides the subspace dimension. An additional data passes follow after the iterative stage is finished to refine clusters with subspaces associated with the medoids. ORCLUS-ORiented projected CLUSter generation [33] is an extended algorithm of earlier proposed PROCLUS. It uses projected clustering on non-axes parallel subspaces of high dimensional space.

B. Hybrid Clustering algorithm

Sometimes it is observed that not all algorithms try to find a unique cluster for each point nor all clusters in all subspaces may have a result in between. It is because of having a number of possibly overlapping points. The exhaustive sets of clusters are found necessarily. FIRES [34], can be used as a basic approach a subspace clustering algorithm. It uses a heuristic aggressive method to produce all subspace clusters.

C. Correlation Clustering

Correlation Clustering is associated with feature vector of correlations among attributes in a high dimensional space. These are assumed to persistent to guide the clustering process. These correlations may found in different clusters with different values, and cannot be reduces to traditional uncorrelated clustering. Correlations among attributes or subset of attributes results different spatial shapes of clusters. Hence, the local patterns are used to define their similarity between cluster objects. The Correlation clustering can be considered as Biclustering as both are related very closely. In the biclustering, it will identify the groups of objects correlation in some of their attributes. The correlation is typical for the individual clusters.

VI. CONCLUSION

The purpose of this article is to present a comprehensive classification of different clustering techniques for high dimensional data. Clustering high dimensional data sets is a ubiquitous task. The incosent growth in the fields of communication and technology, there is tremendous growth in high dimensional data spaces. It study focuses on issues and major drawbacks of existing algorithms. As the number of dimensions increase, many clustering techniques begin to suffer from the curse of dimensionality, de-grading the quality of the results. In high dimensions, data becomes very sparse and distance measures become increasingly meaningless. This problem has been studied extensively and there are various solutions, each appropriate for different types of high dimensional data and data mining procedures.

There are many potential applications like bioinformatics, text mining with high dimensional data where subspace clustering, projected clustering

approaches could help to uncover patterns missed by current clustering approaches. As with any clustering techniques, finding meaningful and useful results depends on the selection of the appropriate clustering technique. In order to do this, one must understand the dataset in a domain specific context in order to be able to best evaluate the results from various approaches. From the above discussion it is observed that the current techniques will suffers with many problems. To improve the performance of the data clustering in high dimensional data, it is necessary to perform research in the areas like dimensionality reduction, redundancy reduction in clusters and data labeling.

REFERENCES

[1] Jacob Kogan, "Introduction to Clustering Large and High-Dimensional Data" , University of Maryland, Baltimore County.

[2] Jiawei Han and Micheline Kamber," Data Mining: Concepts and Techniques",2006

[3] Holmes Finch ." comparison of distance measures in cluster analysis", in the proceedings of journal of data science, 85-100, 2005

[4] Paul E Green and Vithala R Rao, " A Note on Proximity Measures and Cluster Analysis", in Journal Of Marketing Research, 359-64, 1969.

[5] Mark A. Hall, Geo_rey Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining", IEEE Transactions on Knowledge and data engineering, VOL. 15, NO. 3, MAY/JUNE 2003.

[6] Beyer, k., Goldstein, j., Ramakrishnan, r., and Shaft, U. " When is nearest neighbor meaningful?", In Proceedings of the 7th ICDT, Jerusalem, Israel., 1999.

[7] Aggarwal, C.C., Hinneburg, A., and Keim, D.A. "On the surprising behavior of distance metrics in high dimensional space", . IBM Research report, RC 21739, 2000.

[8] Beyer K and Ramakrishnan. "Bottom-up computation of sparse and iceberg cubes". In Proceeding of the ACM-SIGMOD International Conference on Management of Data, Philadelphia, pp 359–370, 1999.

[9] Mardia.K, Kent, J and Bibby.J."Multivariate Analysis". Academic Press, 1980.

[10] McCullum. A., Nigam, K., and Ungar, L.H." Efficient clustering of high dimensional data sets with application to reference matching". In proceedings of the 6th ACM SIGKDD, 167-178, Boston., MA, 2000.

[11] Chris Ding Xiaofeng He, "Adaptive dimension reduction for clustering high dimensional data ", in the Proceedings of IEEE International Conference on Data Mining, Washington DC, USA, 2002.

[12] Lance Parsons, Ehtesham Haque and Huan Liu ," Subspace Clustering for High Dimensional Data: A Review", in the proceedings of SIGKDD Explorations, Volume 6, Issue 1, pages 90-105.

[13] Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P, "Automatic subspace clustering of high dimensional data for data mining applications", In Proceedings of the ACM SIGMOD Conference, Seattle, WA. 1998, 94-105,

[14] Kailing, Karin; Kriegel, Hans-Peter; Kröger, Peer ,"Density-Connected Subspace Clustering for High-Dimensional Data", in the Proceedings of the 5th SIAM International Conference on Data Mining : 2004. 246–257.

[15] Yi-Hong Chu, Ying-Ju Chen ,"Reducing Redundancy in Subspace Clustering", in proceedings of IEEE Transaction on Knowldege and Data Engineering, VOL. 21, NO. 10, OCTOBER 2009.

[16] Hartigan, J, "Clustering Algorithms", John Wiley & Sons, NY, 1975

[17] Nishisato, S. "Analysis of Categorical Data: Dual Scaling and Its Applications". University of Toronto. 1980.

[18] Hung-Leng Chen; Kun-Ta Chuang; Ming-Syan Chen "On Data Labeling for Clustering Categorical Data", IEEE Trans. Knowl. Data Eng., 2008:

[19] David Arthur and Sergei Vassilvitskii , " K-means++: the advantage of Careful Seeding", in Proceeding SODA eighteenth annual ACM-SIAM symposium on Discrete algorithms, USA 2007.

[20] Pavel berkhin.," Survey Of Clustering Data Mining Techniques"., 2005.

[21] Pasi Fränti, Olli Virmajoki and Ville Hautamäki, "Probabilistic Clustering by Random Swap Algorithm", in the proceedings of IEEE conference 2008

[22] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu," A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", In the proceedings of 2nd

International Conference on Knowledge Discovery and Data Mining, 1996.

[23] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander,"OPTICS: Ordering Points To Identify the Clustering Structure", in Proceedings of ACM SIGMOD Intl.. Conf. on Managt of Data, 1999.

[24] Andreas Hotho, Gerd Stumme ," Conceptual Clustering for Text Clusters", Publiction.

[25] Fisher, D.H.," Knowledge Acquisition via Incremental Conceptual Clustering, Machine Learning", 1987

[26] Ganter, Bernhard; Stumme, Gerd; Wille, Rudolf, " Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence", Springer-Verlag, ISBN 3-540-27891-5, 2005

[27] Jardine, N. and R. Sibson. "The construction of hierarchic and non-hierarchic classifications". Comp. J. 11 (2): 177-184.1968.

[28] 28Heyer, Kruglyak, Yooseph, "QT-Clust: Exploring Expression Data: Identification and Analysis of Coexpressed Genes ", in the proceedings of Genome Research, 1999

[29] Jianbo Shi and Jitendra Malik ," Normalized Cuts and Image Segmentation", in the proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI) 2000

[30] C.-H. Cheng, A. W. Fu, and Y. Zhang," Entropy-based subspace clustering for mining numerical data". In Proceedings of the5th ACM SIGKDD internl conf. on Knowledge discovery and data mining, 1999.

[31] S. Goil, H. Nagesh, and A. Choudhary. ,"Mafia: Efficient and scalable subspace clustering for very large data sets". Technical Report CPDC-TR-9906-010, Northwestern University, 2145 Sheridan Road, Evanston IL 60208, June 1999.

[32] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park.,"Fast algorithms for projected clustering". In Proceedings of the ACM SIGMOD international conf.. on Management of data, 61-72. , 1999.

[33] C. C. Aggarwal and P. S. Yu. Finding generalized projected clusters in high dimensional spaces. In Proceedings of the ACM SIGMOD international conference on Management of data, pages 70-81., 2000.

[34] AchtertE, BöhmC, Kriegel H-P, Kröger P ,"Online hierarchical clustering in a data warehouse environment", In: Proceedings of the 5th international conference on data mining (ICDM), Houston,, 2005

◈◈◈