www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 6 June, 2014 Page No. 6505-6508

Implementation of Unsupervised Clustering Using Geometric Approach

Ms. Ayushi Laud[1] and Prof. Ritesh Shah[2]

[1]Ms. Ayushi Laud PG Scholar, Sanghvi Institute of Management & Science,
Indore (M.P), India
aayushi.lad06@gmail.com,

[2]Prof. Ritesh Shah Assistant Professor, Sanghvi Institute of Management & Science,
Indore (M.P), India
ritesh.shah@sims-indore.com

Abstract

Various medical sciences, educational, scientific research works are depends upon data clustering. Clustering is an application of data mining and knowledge processing which used find the data pattern using visual pattern analysis. Using this technique, data based on their attribute distance is mapped over space to find the pattern of data available. Clustering is type of unsupervised learning technique where no class level is available to utilize as feedback parameter for error correction. In this paper a study based on cluster analysis is provided. This includes study of k-mean clustering scheme, and outlier detection algorithm technique and performance enhancement using outlier technique, after concluding them we propose a geometric distance based clustering scheme. This technique is implemented using MATLAB simulator for experiment and academic research purpose, the implemented technique is works over numerical and the estimated results are provided in the further sections.

Keywords—Data mining, clustering, outlier detection, algorithm design, performance analysis.

1. INTRODUCTION

Machine learning is domain of engineering where algorithms and computer based programs are learnt from data and past experience and provides decisions and analysis based on their experience and learned knowledge. Machine learning introduces stepwise process to learn and utilize the knowledge to solve the real world problems in effective and efficient manner. This process includes data pre-processing, training of algorithm and implementation using any application sometimes the last step is also known as testing of data model.

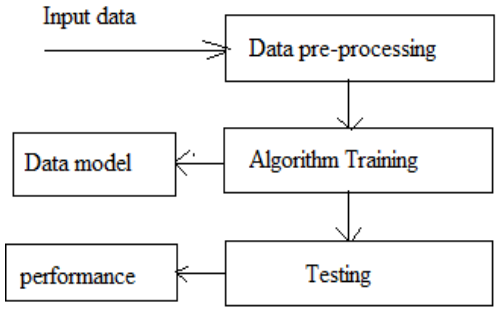

Figure 1.1 shows the machine learning

1. **Data pre-processing:** Data in real world is never found in a structured manner, which is hidden between unstructured data or between uneven data sources. In this phase data is separated from its original format and organized in a desired format to read and utilized for future use.
2. **Algorithm training:** Using the previous phase data machine prepare a data structure or data model by which decisions are made for providing solutions.
3. **Testing:** performance of the system can be measured using this phase of system, here manually or automatically real time or sample data is produces to test the intelligence of trained algorithm.

The proposed study is cluster analysis of data, clustering is process of data mining for finding groups of objects such that the objects in a group will be similar or related to one another and different from or unrelated to the objects in other groups. [2] An example of clustering is given using figure 1.2 where three different clusters are given.

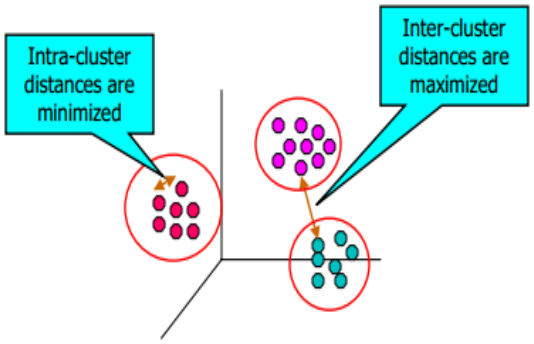

Figure 1.2 shows the clusters of data

This section provides an overview of domain of study in next section of the documents provides the background and literature study.

2. BACKGROUND

The aim and objective is to be accomplished a new clustering scheme during the study. Proposed study includes the following work.

1. **Study of clustering algorithm:** In this phase of study K-mean clustering algorithm is utilized and analyzed; this study includes the working of K-mean algorithm and performance under different dataset.
2. **Study of performance enhancement algorithm:** Here the detailed study is performed for finding the approaches and techniques by which clustering techniques can be improved for that purpose outlier detection technique is analyzed and studied.
3. **Design and propose a new algorithm:** A new algorithm which is based on geometric analysis is proposed and designed and implemented.
4. **Implement proposed algorithm:** Implement all the algorithms using MATLAB simulation tool.
5. **Performance evaluation:** In this step performance of all the implemented algorithms are compared, in order of justification of the proposed algorithm.

Clustering algorithms are categorized according to their application that is our main motivation of the proposed study. Therefore, clustering techniques can be divided into a number of techniques according to analysis and calculations involved to form cluster organizations. [5]

1. **Hierarchical clustering:** This cluster scheme is also known as connectivity based clustering, the basic working of this kind of cluster approach is more related to nearby objects than to objects farther away. These algorithms do not provide a single partitioning of data set, but in its place provide a general hierarchy of clusters that merge with each other at certain distances.
2. **Centroid based clustering:** In centroid-based clustering, clusters are characterized by a central vector which is also called centroid, which may not automatically be a member of the data set. This vector can be calculated using given input data. K-mean is the part of this clustering scheme.
3. **Distribution based clustering:** The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. A nice property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution.
4. **Density based clustering:** In density-based clustering, [3] clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points.

An suitable example of the clustering scheme is provided in order to implementation of real time dataset a research work found in **[Tuhin 2012]** Detection and segmentation of brain tumours in Magnetic Resonance Images is an important but very time-consuming task that required to be completed manually by experts. Due to the occurrence of the high degree of intensity and textural similarity between normal areas and tumour areas in Brain MRI images Automation of this process is a challenging task. This paper proposes a fully automated two step segmentation process of MRI images. Firstly, the scull is stripped from the MRI images by generating a scull mask from the original MRI image. Finally, an advanced K – means algorithm improvised by two level granularity oriented grid based localization process based on standard local deviation is used to segment the image into grey matter, white matter and tumour region. Finally, the length and breadth of the tumour is assessed.

And to improve clustering **[K. Mumtaz et al 2010]** provides a way for mining knowledge from large amounts of spatial data is known as spatial data mining. Huge amounts of spatial data have been collected in various applications from geo-spatial data to bio-medical and similar applications. So, it far from human's ability to analyse these data. Clustering has been accepted as a mining method for knowledge discovery in spatial data. The data can be clustered in different manners depending on the clustering algorithm and other factors. In this paper, a novel density based k-means clustering algorithm has been proposed to overcome the drawbacks of DBSCAN and k-means clustering scheme. The result will be an improved version of k-means clustering algorithm. The given algorithm performs better than DBSCAN while handling clusters of circularly distributed data points and slightly overlapped clusters.

This section of paper provides the literature study and background of the clustering schemes, in next section of the paper includes the study of algorithms.

3. PROPOSED SYSTEM

The K-Means clustering algorithm is a partition-based cluster analysis method [1]. According to the algorithm we firstly select k objects as initial cluster centres, then calculate the distance between each object and each cluster centre and assign it to the nearest cluster, update mean of all clusters, repeat this process until the criterion function converged. Square error criterion for clustering

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| x_{ij} - m_i \right\|^2 \quad (1)$$

x_{ij} Is the sample j of i-class, $m_i$is the centre of i-class, n_i i is the number of samples of i-class. K-means clustering algorithm is simply described as

Input: N objects to be cluster (xj, xz … xn), the number of clusters k;
Output: k clusters and aggregate dissimilarity between each object and its nearest cluster centre is the smallest;
Process: 1. Arbitrarily select k objects as initial cluster centres $(m_1, m_2, \dots, m_k)$; 2. Calculate the distance between each object Xi and each cluster centre, then assign each object to the nearest cluster, formula for calculating distance as: $$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d} (x_i - m_{j1})^2}, i = 1 \dots N, j = 1 \dots k$$ $d(x_i, m_i)$is the distance between data i and cluster j. 3. Calculate mean of attributes in each cluster as the new cluster centres, $$m_i = \frac{1}{N} \sum_{j-1}^{n_i} x_{ij}, i = 1,2, \dots, K$$ N_i is the number of samples of current cluster i; 4. Repeat 2) 3) until the criterion function E converged, return$(m_1, m_2, \dots, m_k)$ Algorithm terminates.

The K-Means algorithm is very sensitive to the initial cluster centres so that the clustering results will be very different result from different initial cluster centres. If the data points exist in isolation, in other words, a small amount of data points are far from data-intensive areas, the calculation of the mean point would be affected, and the new cluster centre may deviate from the true data-intensive area which eventually lead to a clustering output result of large deviation. Therefore we firstly remove isolation point in Data set before data clustering. The abnormal degree of each object in database is measured by local outlier factor LOF. LOF first generates k-Neighbourhood and k-nearest neighbour distance of all objects, and then calculates the distance between each object and the objects which are in its k-Neighbourhood; at last LOF identifies local outlier according to the local outlier factor of each object [1]. The procedure of outlier detection is briefly described as follows:

1. Calculate k-nearest neighbourhood distance named distance $(P, i)(i \in N_k(P))$ of each object p, distance (p, i) is defined as the direct connection distance between object p and i,

$$distance = \sqrt{(x^1 - y^1)^2 + (x^2 - y^2)^2 + \cdots + (x^n - y^n)^2} \quad (2)$$

Where n is dimension of dataset

2. Calculate the density of each object p. The density of object p which reflects the distribution of the data near is defined by the reciprocal of k-nearest neighbour mean. It is described as follows:

$$lrd(p) = \frac{1}{\frac{1}{K}\sum_{i=1}^{K} distance(p, i)} \quad (3)$$

3. Calculate the local outlier factor of p.

$$lof(p) = \frac{\sum_{i=1}^{k} \frac{lrd(i)}{lrd(p)}}{k} \quad (4)$$

lrd (i) is the local density of k-nearest neighbour of p, lrd (p) is the local density of p. lof (p) reflects the extent of p as a outlier. The value of outlier factor is about I in the data set of density distribution. As the local density of the outliers in the data set is much less than the local density of its neighbourhood, outlier factor by which outlier can be distinguished is larger than others.

4. If lof (p) is much larger than 1, P is an isolated point, delete the object p, return the first step until the number of data sets remains unchanged. The new data sets generated is the data set to be clustered.

We first make a pre-treatment with the data to be clustered to remove the outliers using outlier detection method based on LOF, so that the outliers cannot participate in the calculation of the initial cluster centres, and excluded the interference of outliers in the search for the next cluster centre point. We secondly apply fast global K-Means clustering algorithm on the new data set which is generated previously. Fast global K-Means clustering algorithm is an improved global K-Means clustering algorithm by AristidisLikas. The improved clustering algorithm is described as follows:

Input: data set M $\{x_1, x_2, \ldots, x_m\}$ to be clustered, the number of clusters k;
Output: k clusters and aggregate of dissimilarity between all objects and their nearest cluster centres is the smallest.
Process: 1. Traversal on data set M, calculate lof (p), $$lof(p) = \frac{\sum_{i=1}^{k} \frac{lrd(i)}{lrd(p)}}{k}$$ if lof (p) is much larger than 1,remove the isolated point p, otherwise leave p; finally get the new data set N; 2. Calculate the mean of data set N as the first cluster centre. $$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$ 3. Find the next cluster centre. Calculate the distance between the remaining points and the cluster centre. $$b_n = \sum_{j=1}^{N} \max\left(d_{k-1}^{j} - \|x_n - x_j\|^2, 0\right)$$ make the sample point X_n whose $b_n$is the largest as the next initial cluster center, calculate the distance between each object X_n and the center of each cluster, and assign it to the nearest cluster 4. Calculate the mean of objects in each cluster as a new cluster centre. 5. Repeat 3) 4) until the criterion function E converged, return$(m_1, m_2, \ldots, m_k)$. Algorithm terminates

Clustering is an interesting domain of separating points geometrically to classify the objects in a given solution space. The proposed study is based on previously made effort and improvements over traditional K-mean clustering scheme. The selected algorithm K-mean has some deficiencies.

1. **Handling Empty Clusters:** One of the problems with the basic K-means algorithm given earlier is that empty clusters can be obtained if no points are allocated to a cluster during the assignment step. If this situation occurs then a strategy is needed to choose a replacement centroid, since otherwise, the squared error will be larger than necessary.
2. **Outliers:** When outliers are present, the resulting cluster centroids (prototypes) may not be as representative as they otherwise would be and thus, the SSE will be higher as well.
3. **Reducing the SSE with Post processing:** In k-means to get better clustering we have to reduce the SSE that is most difficult task. There are various types of clustering methods available which reduces the SSE.

The above given problems are the most common problem with k-mean clustering schemes, to overcome these problems methods are suggested in [1]. To obtain high efficient results here solution steps are provided in further sections.

4. PROPOSED ALGORITHM

The proposed system includes implementation of three different algorithms first traditional K-mean clustering algorithm, second dense cluster formation method using outlier detection method and third proposed algorithm. As we discussed previously k-mean clustering scheme are type of unsupervised learning techniques of clustering, thus the system basically works in the below given phases.

Figure 4.1 shows the clustering

In pre-processing step data is accepted in and filtered using different technique, the proposed system is developed

through machine learning dataset available in ARFF format. In training process algorithm works in the steps involved in algorithm and produces the data clusters in 2 d space.

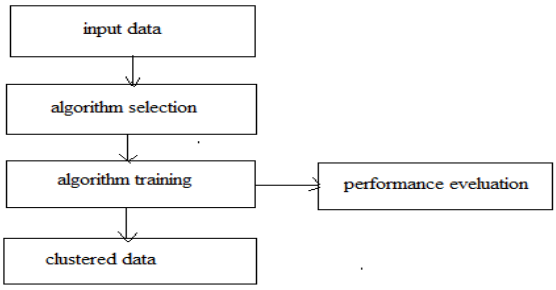

Figure 4.2 shows the system architecture

After clustering the performance of the system is evaluated based on the clustered classified data. Here cross validation technique is implemented which returns the bit error after comparing the real classified classes and algorithms classifications.

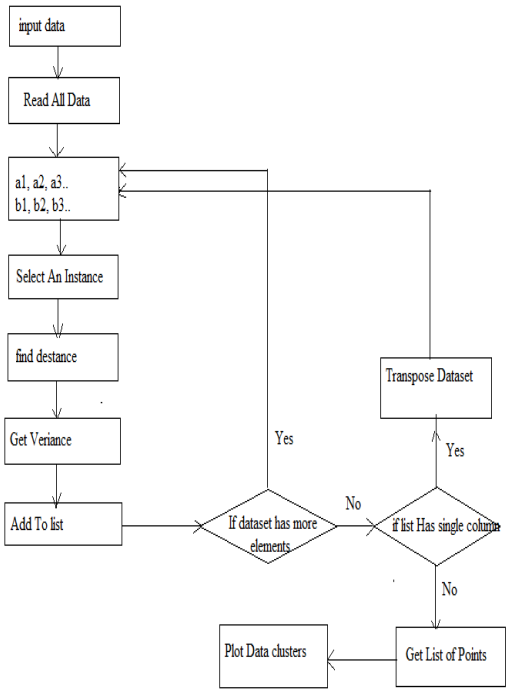

Figure 4.3 show algorithms flow diagram

Input: Dataset with n number of data samples (instances)
Output: similar classification over 2D hyper plane
Process: 1. Read all data sequentially 2. Select an instance X_i 3. For each X_i in data set D_n a. Find $d(p-q) = \sqrt{\sum_{i=0}^{n}(Q_i - P_i)}$ b. Create List Dlist $d(p,q)$ 4. End for 5. A normalize value for each instance is calculated using 6. $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \mu^2$ 7. The calculated values provide x axis for an instance 8. Transpose the created D_n 9. Repeat step 2 on transpose D_n 10. this process generate the values for y axis 11. plot all (X_i, Y_i) in 2d vector space

5. IMPLEMENTATION AND RESULTS

This section of the presented document provides the results and the performance of the designed algorithm, with respect of K-mean clustering algorithm and the previously modified algorithm. The performance of the algorithm is estimated in terms of accuracy, error rate and memory uses. To estimate results there are 5 different datasets are used and the relevant performance of the system is given below.

5.1 Accuracy

Here to evaluate the performance of the proposed system is calculated using the n cross validation method. The overall classification accuracy is given below and evaluated using the below given formula, the listed accuracy of the system are the best performance during different experiments.

$$\% \text{ accurecy } = \frac{Total\ correctly\ classified}{total\ values\ to\ classify} X100$$

Table 5.1 shows the accuracy of system

Data set	K-mean	Density cluster	Proposed
Laustralian dataset30	76.34	78.28	79.91
breast cancer dataset	82.42	85.27	89.72
diabetes data sets30	63.9	67.38	70.27
iris dataset TF30	100	100	100
bupa dataset	73.29	78.62	85.21

The accuracy after finding outliers are improved effectively and in addition of that the proposed algorithm is utilized geometric distance based algorithm, this can produces more accurate cluster formation.

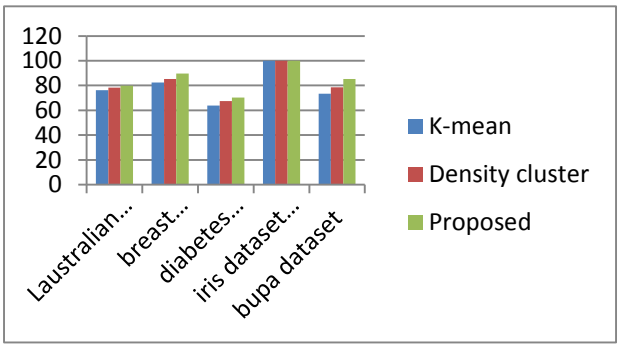

Figure 5.1 shows the accuracy of system

5.2 Error Rate

Error rate is a scale which provides the information how much amount of objects are incorrectly classified during experiments. Due to experiments we found that as the size of database increase the error rate increases, which is evaluated using the given formula.

$$\%\ error = \frac{Total\ incorrectly\ classified}{Total\ Number\ of\ objects} X100$$

Table 5.2 shows the error rate of the system

Data set	K-mean	Density cluster	Proposed
Laustralian dataset30	23.66	21.72	20.09
breast cancer dataset	17.58	14.73	10.28
diabetes data sets30	36.1	32.62	29.73
iris dataset TF30	0	0	0
bupa dataset	26.71	21.38	14.79

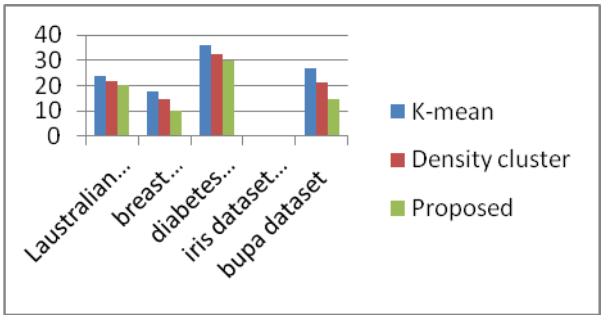

Figure 5.2 shows the error rate of the system

5.3 Memory Uses

That is defined as the memory resources consumed during the execution of the system, here the consumption of the resources in given in KB.

Table 5.3 shows the memory consumed

Data set	K-mean	Density cluster	Proposed
Laustralian dataset30	26901	27183	29211
breast cancer dataset	28198	27189	28291
diabetes data sets30	28747	27191	29289
iris dataset TF30	27817	29811	29832
bupa dataset	28891	28381	28389

The proposed algorithm is consumes more memory resources than other two algorithms thus in future work improvement is required to reduce the cost of resources consumed during algorithm processing.

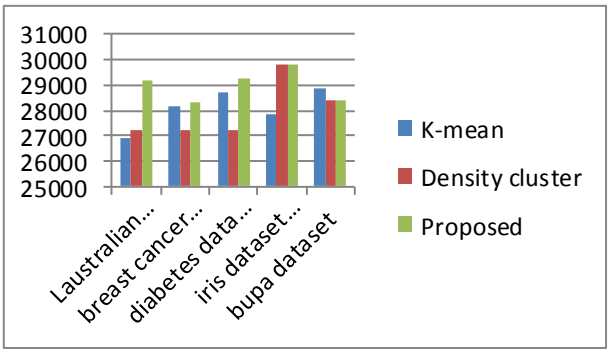

Figure 5.3 shows the memory consumed

5.4 Model Building Time

The total time required to develop data model is known as model build time, which is estimated using the time

difference between initialization of algorithm to the results producing of the system.

Table 5.4 shows the build time

Data set	K-mean	Density cluster	Proposed
Laustralian dataset30	2.19	2.83	2.11
breast cancer dataset	1.98	2.13	2.01
diabetes data sets30	4.72	4.19	3.35
iris dataset TF30	1.27	2.67	2.2
bupa dataset	2.91	2.8	2.2

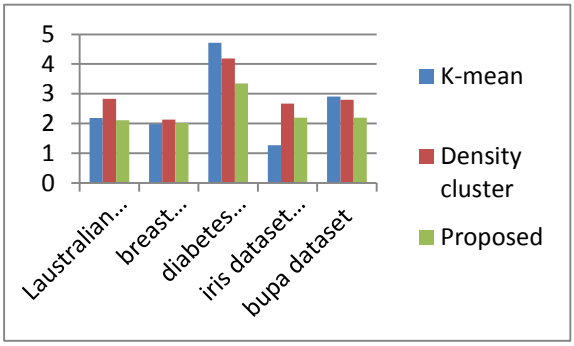

Figure 5.4 shows the build time of the system

6. CONCLUSION AND FUTURE WORK

The proposed algorithm for finding the accurate and efficient data clusters is implemented successfully. The proposed technique is provides a way to design clustering algorithm based on the geometric distance calculation method. Here for each data instances a similarity is computed to form data clusters. To get this idea various kinds of clustering schemes are studied which are listed under section of 2.1 and 2.3 in addition of this k-mean clustering algorithm is studies in this proposed work.

After literature study that is observed that there are some improvements are required to handle deficiencies of the traditional k mean clustering. A similar effort in the domain of k-mean clustering is found in [4] and other similar, where outlier detection is proposed to find dense and clear clusters this method improves the clustering quality and provides clear and dense clusters at the time of cluster analysis. To improve more this scheme a geometric distance based clustering schemes is proposed and implemented by us.

After implementation the experimental results shows the following advantage of the proposed algorithm.

Performance parameters	Remark
Accuracy	After implementation of the proposed technique of clustering the accuracy of the system is improved.
Error rate	The error rate is decreases at the time of cluster analysis
Memory	Memory consumption is normal but in comparison of traditional algorithm consumes more resources thus improvement is required
Model building time	Time required to form clusters is normal and much nearer then the previously proposed systems.

According to the above given performance and results analysis proposed technique is able produce clear and accurate clusters of data based on the geometric analysis. Overall performance of the proposed system is adoptable but required some improvements for decreasing the memory cost.

Clustering can be implementable with verity of applications, such as huge data analysis, image processing, biometric applications, feature extraction and others. The proposed method is simple to implement and efficient in clustering formation. But that is only works with the numerical data analysis. In future that is extended to perform over categorical data analysis too. Results observation shows that the overall performance is quite adoptable but due to memory resource some improvements are required in future work.

References

[1] Juntao Wang, Xiaolong Su, "An improved K-Means clustering algorithm", 978-1-61284-486-2/111$26.00 ©2011 IEEE.

[2] Tan, Steinbach, Kumar, "Data Mining Cluster Analysis: Basic Concepts and Algorithms", 2004.

[3] C.C.Aggarwal A Human-Computer Interactive Method for Projected Clustering [.I]. IEEE Transactions on Knowledge and Data Engnieering, 16(4), 448-460, 2004.

[4] JiaweiHan, Micheline Kamber Data Mining Concepts and Techniques [M] Beijing: Mechanical Industry Press, 2005 185-218.

[5] J Dong, M. Qi. K-means Optimization Algorithm for Solving Clustering Problem. Knowledge Discovery and Data Mining.

2009:52-55.

[6] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, etal. LOF: Identifying Density-Based Local Outliers [J]. Proc of ACM SIGMOD Conf, 2000, 2(29)93-104.