ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Multi Resolution Pruning Based Co-location Identification in Spatial Data

Mrs.V.Prema[1], R.Selvasudhan[2]

Assistant Professor, Department of Computer Science, Valliammai Engineering College, Chennai, India[1]

PG Student, M.E (CSE), Department of Computer Science, Valliammai Engineering College, Chennai, India[2]

ABSTRACT—In this paper we put forward a plant prediction system with advanced clustering and improved colocation mining. Spatial data differs from the other forms of data by the fact that the neighbouring objects will have noteworthy effect to the object under consideration. Thus mining the data item and its co-location pattern together becomes more vital. In our work, we suggest an advanced method of plant prediction scheme by two steps. Initially, clustering the location areas into three types according to the nature of the location by considering the GIS (Geographic Information System) attributes and clustering the plant species also according to thegeographical location which suits the existence of the plant. These clustering is done by modifying traditional k-means clustering algorithm by altering its repeated iteration process into single iteration and repeating the same clustering by considering multiple attributes associated with the location and plant species. Finally, a combinatorial spatial co-location algorithm is used to mine the co-locations and a plant prediction system is designed in which, for a given plant species, prediction of suitable co-locations which has the highest supporting environment to grow and a set of plant species which has the highest probability of co-existence is determined. Experimental results shows the prediction to be more effective in computation time and accuracy, particularly while updating the database dynamically with the new entries as the computation and prediction is limited to the initially clustered dataset rather than the complete database.

KEYWORDS: GIS, Spatial co-location, plant prediction, iteration less k-means

I. INTRODUCTION

Data mining has wide applications in areas like medical, statistical, satellite and other related databases. In recent times, mining of geographical information from spatial databases for extraction of useful patterns has been an area of keen consideration for researchers. Hence, Data mining methods are applied to the spatial related data leads to the development of many spatial analysis and spatial feature prediction applications. We propose a plant predictions system which suggests the best possible location [1] for a plant to survive and the plant species which are having the highest probability of existing with the concerned plant species.

Merging of Geographical Information System attributes with the data mining technologies created new advancement in the agricultural researches. [2]Processing of agricultural datasets with the GIS system and advanced clustering methods of data mining has led to new ideas in prediction system. Many data clustering algorithm have been developed in the field of data mining. In recent time like DBSCAN, K-Median and K- Medoid algorithms and other variants are popularly used. In our proposed approach we have used theK-Means algorithm in a modified way to cluster the location information from GIS and the plant species information from a database.

Traditional methods involve clustering of data items that have closely related attributes to the concerned data item into the prescribed cluster size and prediction of higher probable items within the threshold level.i.e., for a given plant species, clustering is done completely at both location dataset and the plant dataset to find its cluster of groups[1]. This method consumes more time when the database grows large and time needed to produce cluster become high. Our approach involves Land suitability measures to categorize the location into three types and also categorize the plant species by nature of growth area into the same three types of categories. Prediction is done with reference to those three types of clusters. K-means algorithm is used for the clustering process of the location datasets and the plant species datasets initially. A combined algorithm is used to mine the closely related data items in the prediction phase where the possible co-locations and the possible plant species are mined. Experimental results confirm the increased efficiency in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

the computation time of clustering and the increased accuracy in the predicted data items. Computation time efficiency is achieved as during the dynamic update of database, prediction process is limited to the clustered datasets rather than the complete dataset.

Section II of this paper describes the problem definition. Section III presents the related works of the paper. Section IV explains the proposed technique and its methods. Results and performance analysis is discussed in section V.Section VI shows the output screen shots of the proposed prediction system and conclusion is summarized as section VII.

II. PROBLEM DEFINITION

Prediction of co-existence of plant species in spatially located areas seems to be a vital measure for the agricultural researchers [3]. The common problem that is associated with the traditional prediction system of any database is the time efficiency of computation. Likewise, in plant database time taken for clustering increases along with the growth of the database by the addition of newer plant species causes the whole clustering process to be recomputed. Thus minimising the computation of the co-location areas and the [1] prediction of the plant species must be taken into account for efficient computations in larger databases. Design of such system should ensure that it should be able to overcome these problems.

III. RELATED WORKS

Recent papers of many researchers presents various works regarding the clustering process. Some of the significant clustering procedures are highlighted in this section.

Land Suitability Evaluation Method Based on GIS Technology Liquan Qu1,2, Yuanzheng Shao2, Lianpeng Zhang1 [2].The paper shows the method of land suitability evaluation based on GIS technology. This paper studies the method of examining the evaluation results and improving its application value. The important facts that are discusssd in this paper are the Collection and collation of basic data in research area. To improve the evaluation results while evaluating the application value, it is necessary to mine spatial data with reference to the evaluation results and other materials. In order to achieve that they get some thematic maps for their computation such as arable land potential area, and returning many types of land to their respective areas according to their suitability.GIS plays an important role in this process. These evaluation results can be used to provide land planning without any further modification.

CHEN Guang-xue, LI Xiao-zhou ,LI Xiao-zhou[14] they have studied clustering algorithms of area geographical entities based on geometric shape similarity. They have presented three important facts that is to be considered. Firstly, a similarity criterion of line segments shape. Secondly,a criterion of area geographical entities comprehensively utilizing distance and finally, geometric shape similarity. For clustering analysis of spatial area and geographical locations. Clustering algorithms which is based on these requirements are found to be more efficient

Jie Wang et al. [12] have presented clustering analysis of errors for railway transport. This was based on DBSCAN. They found problems prevailed in the process of mining errors for railway transport; an improved DBSCAN algorithm was presented. This paper describes the study to mine the driver errors in the railway system by examining the data.Their improved Algorithm enhanced DBSCAN in taking account of the boundary shared object of cluster using the average distance function of shared object. Hence it efficientlyprevents over-segmentation and produced more exact clustering results. Experimental results showed that this method is effective.

R. Agarwal and R. Srikant explains that approaches that are based on the apriori framework apply three steps to discover co-location patterns [3]. The first step of their method was to generate candidate patterns. The second step was to determine instances of candidate patterns, and then finally the last step is to calculate the prevalence of each pattern. The major differences betweenresearches in this aspect are the algorithms that they used to generate instances of patterns and also the method they used to measure the prevalence (as an interestingness measure) of patterns.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

IV. PROPOSED SYSTEM

A) The Input Section

Input consists of the databases. There are two databases used in the system. Location database which comprised of the data about the various spatial locations. [2]GIS(Geographical Information System) measures each geographical area by means of various attributes the major characteristic attributes include latitude and longitude of location, prevailing temperature, rainfall, soil type, soil fertility, ground water level, and pH value of the soil.

TABLE I. Sample Dataset of Location

Location name	Latitude	Longitude	Temperature	Rainfall	Soil fertility	Soil type	pH	slope
A	13	24	80	23	2.000	Clay	.33	30
B	15	25	67	24	2.533	Sandy	.37	23
C	17	23	78	13	2.333	Alluvial	.98	20
D	15	26	60	14	4.333	Red	.35	15

Likewise, the plant database also described by the above stated GIS measures like what is the optimal temperature, rainfall, and other attributes that the plant species require to grow without any deficiency [3]

TABLE II. SAMPLE PLANT DATASET

Plant name	Growing time(weeks)	Temperature	Rainfall	Soil fertility	Soil type	pH	Water level
A	24	80	23	2.000	Clay	.33	100
B	25	67	24	2.533	Sandy	.37	150
C	23	78	13	2.333	Alluvial	.98	200
D	26	60	14	4.333	Red	.35	230

B) GIS Categorization

Survey of statistical and climatic features in a particular land constitutes the GIS attributes of the concerned land. With reference to these attributes the land locations are categorized into the following three types namely, dry region, plain region and Mountain region [2].Thus every particular location is characterised by itsspatio-regional characteristics. The range of the attributes falling in this region are summarised in a table as follows. The same criteria can be inversely applied to the plant input which results in categorising the plant species by means of determining which type of area will suit for the existence of the plant species

TABLE III. GIS CHARACTERISTICS

.Region Style	Index System	
	Nature	Socio-Economic
Dry and drought region	Low average rainfall, accumulated temperature, elevation, surface type, soil type	Low water usage Less population density

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Plain , Basins	Moderate Average annual rainfall, annual accumulated temperature, elevation, soil type, soil depth, soil fertility, PH value	Medium or high population, Moderate vegetation and irrigation. Industrial possibilities
Mountain, Hills	High water accessibility, distribution of settlements and farming radius (only in sparsely populated areas), traffic conditions, input-output ratio (only in sparsely populated areas)	High water accessibility, Atmospheric suitability of existence Less or no population density

C) Algorithm-Initial Clustering

Iteration less K-means algorithm is used in the initial clustering of the land and plant databases. Traditional K-Means involves the idea of assigning the mean values as the number of clusters to be created. Repeated iterations are done and each time the mean value gets updated to the centre of the cluster and algorithm iterates itself till it gets the optimal cluster of datasets and there is no more possibility of assigning new cluster centres. In our case the mean values will be the values of the point of convergence of the three types of locations that is characterised by the various GIS attributes[3] that stated before and there will not be any repeated iterations and updating the mean values.

Input: D-Set of data items,
{F}Set of Spatial feature(attribute)
Min_t-minimum threshold(Colocation size),
Output:CAT-Initial Spatial Cluster(Identified inner cluster by iteration-less k-means algorithm)
Steps
While (i)
Foreach item i€ D, and i≠last
Do
For Spatial feature $F_x$in {F}
Clust[i]=Iteration-less_k-means(F_x(i))
Add next iteration value to the array
WhileF_x≠last
Endforeach
CAT=Mine frequent_item inset{clust[]}
Add i^{th}data item to the cluster CAT
End while

Assigning a data item to a cluster involves determining the distance between the data item and the cluster centres and grouping it with the nearest cluster centre. The distance is calculated using the following equation.

Where,

μ_i- Cluster centre

x_j- j^{th} data item of instance x

k-no of clusters

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

This algorithm gives the possible categorization of both location and plant species [1] under three categories namely drained, plain areas and hill region. The attributes taken into consideration are the temperature, rainfall, soil type, soil fertility, water level, pH, slope of the area.

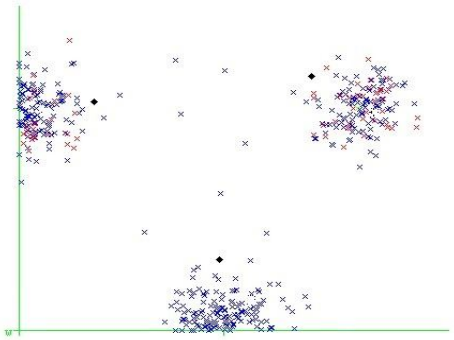

Fig 1. Iteration of Traditional K-means

The above figure shows the snapshot of one of the iteration of the traditional k-means algorithm, where the cluster centre is reassigned at the end of each iteration. As we use a single iteration in k-means algorithm, the cluster centre is fixed for each attribute and the possible clusters at the end of clustering of data items with reference to all attributes will produce a much more accurate cluster as shown below.

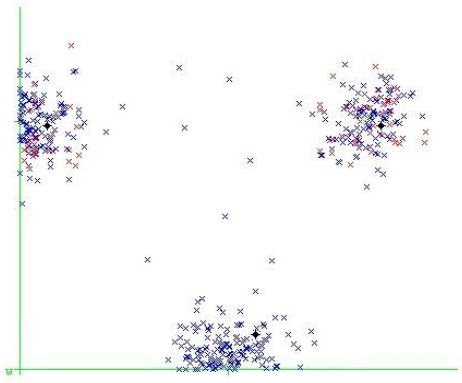

Fig 2. Iteration less K-means

D) Colocation prediction algorithm

After the clustering process, prediction process starts by examining the input value and comparing the values with the initial clusters and determining the inner cluster. Then the following combinatorial spatial co-location algorithm is used to mine the possible co-location

Input: I- Input data type to prediction algorithm,
{F}Set of Spatial feature for the data i(attribute)
Min_t-minimum threshold(Colocation size),
Output: A set of co-locations and a set of possible plant species which suits the plant existence and the co existence of other plant respectively.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Steps

Read input(i)

Do

For Every Spatial feature $F_x$for i

Clust[i]=Assign CAT

Add next iteration value to thearray

While$F_x \neq$last

Add i^{th}data item to the cluster CAT which occurs more

Foreachitem x in CAT(i)

If i is a location

{Co-location}= Minenearestneighbourhood(i,min_t)

Else if i is a plant species

If (prob>min_t)

Add plant species to {co-species}

Endif

End

The above algorithm is used to mine the co-locations and the co-existing plant species in a supportive environment. This algorithms efficiency depends upon the initial clustering algorithm, the k-means which is modified to have limited iterative capacity.

V. PERFORMANCE AND RESULTS

The proposed iteration less k-means algorithm for clustering the plant and location databases is tested with various datasets for the clustering efficiency. Apart from the prediction algorithm, the modified clustering algorithm itself shows slight improvement in the clustering process. When compared with the traditional k-means algorithm, the proposed algorithm has producedmore desired results. Weka explorer tool is used to analyse the performance of the traditional and the iteration-less k-means algorithm and the results are shown below.

Clusterer output

```
kMeans
======

Number of iterations: 1
Within cluster sum of squared errors: 7144.13510296457
Missing values globally replaced with mean/mode

Cluster centroids:
                          Cluster#
Attribute        Full Data         0          1          2
                    (5687)    (1311)     (1971)     (2405)
=============================================================
speaker                 3         3          3          8
utterance         24.1906   30.1754    26.0076    19.4391
frame              8.6147   12.6331     4.2085    10.0353
coefficient1       0.7548    0.4236     0.8655     0.8446
coefficient2      -0.5093   -0.3734    -0.2767    -0.7741
coefficient3       0.2178    0.1272     0.0632      0.394
coefficient4      -0.3049   -0.1385    -0.5432    -0.2003
coefficient5        0.231    0.4428     0.0567     0.2584
coefficient6      -0.2448   -0.2309     -0.345    -0.1703
coefficient7       -0.156   -0.0946    -0.0902    -0.2435
coefficient8      -0.0422   -0.2301     0.0902    -0.0484
coefficient9      -0.2113   -0.2405    -0.0919    -0.2933
coefficient10     -0.1803   -0.1512    -0.1951     -0.184
coefficient11     -0.0245   -0.0304    -0.0757     0.0207
coefficient12      0.0924    0.0931     0.1512     0.0438



Time taken to build model (full training data) : 0.11 seconds
```

Fig 3. Limited Iteration Results

Clusterer output

```
kMeans
======

Number of iterations: 24
Within cluster sum of squared errors: 5604.0396622193675
Missing values globally replaced with mean/mode

Cluster centroids:
                          Cluster#
Attribute        Full Data         0          1          2
                    (5687)    (1232)     (2072)     (2383)
=============================================================
speaker                 3         3          3          8
utterance         24.1906   33.332     25.8084    18.0579
frame              8.6147   13.2784     4.2481    10.0004
coefficient1       0.7548     0.215     0.8176     0.9793
coefficient2      -0.5093   -0.4568    -0.2929    -0.7246
coefficient3       0.2178    0.2318     0.0972     0.3155
coefficient4      -0.3049   -0.1108    -0.5724    -0.1727
coefficient5        0.231    0.4886     0.0559     0.2501
coefficient6      -0.2448   -0.2757    -0.3586    -0.1299
coefficient7       -0.156   -0.1222    -0.0907    -0.2303
coefficient8      -0.0422   -0.2263     0.1148    -0.0837
coefficient9      -0.2113    -0.291    -0.1105    -0.2579
coefficient10     -0.1803   -0.1102    -0.1977    -0.2013
coefficient11     -0.0245   -0.0221    -0.0725      0.016
coefficient12      0.0924    0.0405     0.1459     0.0727



Time taken to build model (full training data) : 2.15 seconds
```

Fig 4. Normal Iteration Results

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Similar run of various datasets proved the efficiency of clustering. The results of various runs of a plant dataset with 567 items and 7 GIS attributes is summarised as follows.

These results are achieved under normal system conditions and the standard tool environment for the weka data mining tool.

TABLE IV. SIMULATED ITERATION RESULTS

Algorithm	Run	Total Instances	Total Iterations	Attributes	Time Taken(ms)	Time per Iteration(ms)
Traditional k-means	1	5687	24	7	2050	85.41
	2	5687	25	7	2150	89.58
	3	5687	24	7	1980	82.5
Iteration less k-means	1	5687	1	7	560	80
	2	5687	1	7	576	72
	3	5687	1	7	539	77

VI. OUTPUT SCREENSHOTS

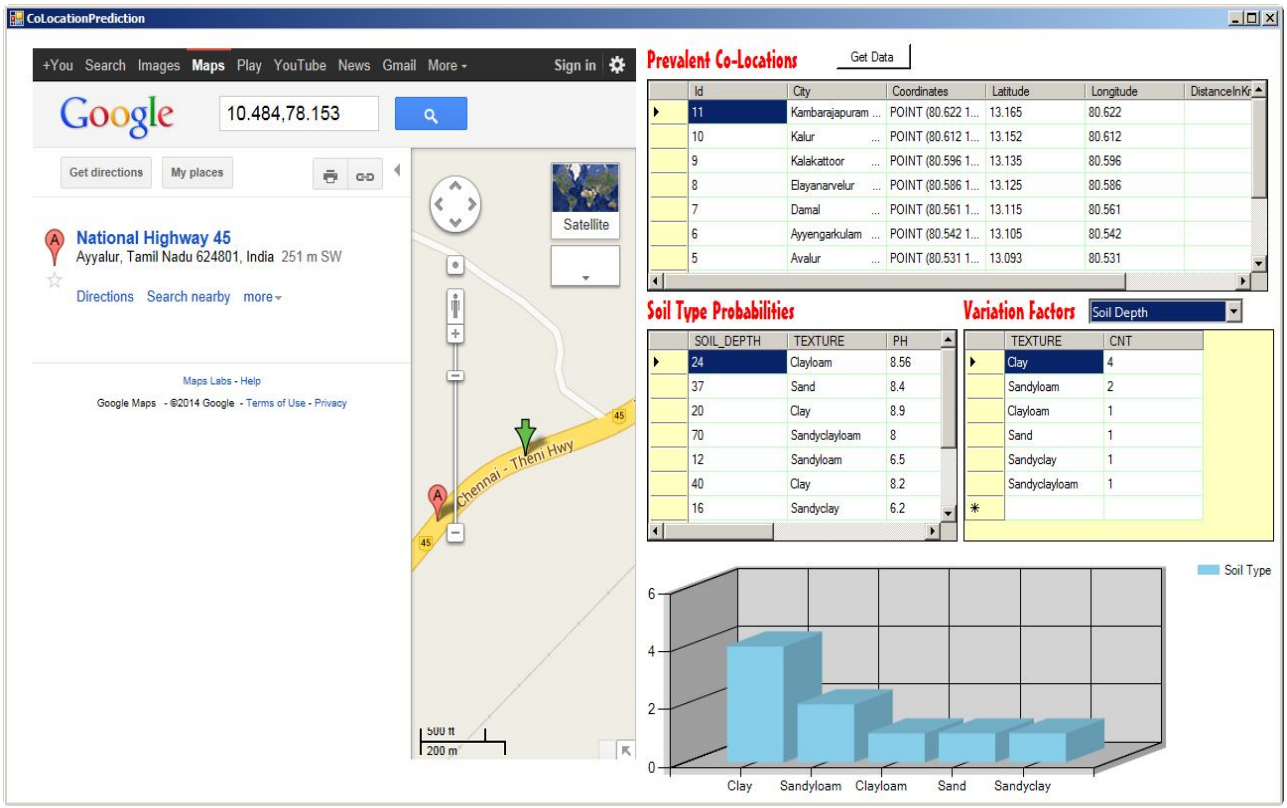

Fig 5. Initial Clustering

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

VII. CONCLUSION AND FUTURE WORKS

This paper studies the problem of pulling out the co-locations in a spatially uncertain datasets. It takes the plant and the location input as per the GIS measures of land categorisation. Then we use iteration-less k means algorithm to determine the initial clusters of the land and plant species. A combinatorial co-location algorithm is then used to mine the colocations. Experimental results have proved the increased accuracy in prediction and reduced computation overhead due to the reduction in cluster size. As future work in this area, mining of co-location in a efficient way will be researched. It also opens door for the possibility of mining the co-locations in various data models like fuzzy models and colour composition of an image and graphical models.

REFERENCES

[1].Y. Huang, S. Shekhar, and H. Xiong, "Discovering Co-Location Patterns from Spatial Data Sets: A General Approach," IEEE Trans.BKnowledge and Data Eng., vol. 16, no. 12, pp. 1472-1485, Dec. 2004

[2].Liquan Qu1,2, Yuanzheng Shao2, Lianpeng Zhang1," Land Suitability Evaluation Method Based on GIS Technology"

[3].A. Meenakshi, Dr. V. Mohan," Localized Matching Model for Plant Prediction Using Incremental Clustering" IEEE- Fourth International Conference on Advanced Computing, ICoAC 2012

[4] .C.C. Aggarwal and P.S. Yu, "A Survey of Uncertain Data Algorithms and Applications," IEEE Trans. Knowledge and Data Eng. (TKDE), vol. 21, no. 5, pp. 609-623, May 2009.

[5].T. Bernecker, H-P Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic Frequent Itemset Mining in Uncertain Databases," Proc. 15th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD '09), pp. 119-127, 2009.

[6].R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. Int'l Conf. Very Large Data Ba ses (VLDB), pp. 487-499, 1994.

[7].P. Agrawal, O. Benjelloun, A. Das Sarma, C. Hayworth, S. Nabar, T. Sugihara, and J. Widom, "Trio: A System for Data, Uncertainty, and Lineage," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 1151-1154, 2006.

[8].M. Ester, H.-P. Kriegel, and J. Sander, "Knowledge Discovery in Spatial Databases," Proc. 23rd German Conf. Artificial Intelligence (KI '99), (Invited Paper), vol. 1701, pp. 61-74, 1999.

[9].Y. Huang, H. Xiong, S. Shekhar, and J. Pei, "Mining Confident Co- Location Rules without a Support Threshold," Proc. ACM Symp. Applied Computing, pp. 497-501, 2003.

[10].C.C. Aggarwal et al., "Frequent Pattern Mining with Uncertain Data," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 29-37, 2009.

[11].Y. Huang, J. Pei, and H. Xiong, "Mining Co-Location Patterns with Rare Events from Spatial Data Sets," Geoinformatica, vol. 10, no. 3,pp. 239-260, 2006.

[12].Jie Wang, Hongpingshu, Guangsong Yan, "Clustering analysis of manipulate errors for railway transport based on DBSCAN", International Conference on Computer Science and Service System (CSSS), pp.3080-3082, 2011.

[13].CHEN Guang-xue, LI Xiao-zhou ,LI Xiao-zhou "Clustering Algorithms for Area Geographical Entities in Spatial Data Mining"2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010)