Implementation of Rural Education Development System Using Spatial Data Mining

T. Sajana*, M. Hima Bindu, Ch. Haritha Rani and B. Sravya Sri

Department of Computer Science and Engineering, Koneru Lakshmaiah University, Vaddeswaram, Guntur-522 502, Andhra Pradesh, India

*Corresponding author.

Abstract

Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful pattern from large spatial datasets. Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. It is a highly demanding field because large amounts of spatial data have been collected in various applications ranging from remote sensing to geographical information systems, environmental assessment, computer cartography and planning. Recent studies on spatial data mining have extended the scope of data mining from relational and transactional databases to spatial databases. It shows that spatial data mining using clustering is a promising field also gives fruitful research results and many challenging issues.

Introduction

In the era of digital world, different kinds of data are being collected and accumulated at a dramatic phase and the massive amounts of the data are being gathered across a wide variety of fields. For the rapidly growing large and the multi-dimensional datasets the data mining techniques become necessary for extracting the useful knowledge. Many researchers have developed parallel versions of sequential data mining algorithms to cope with the large volumes of data. By the introduction of significant communication overhead makes them inefficient, these parallel versions may help to speed up the intensive computations. Two main steps are proposed to reduce the communication overheads DDM algorithm approach. In the first phase by creating local results on each node for the local data sets the data is usually distributed to execute the mining process. In the second phase the local results are aggregated to build it into global results. By the efficiency of aggregation phase only leads to the efficiency of DDM algorithm. For analysing these large and multi-dimensional datasets the distributed data mining with efficient aggregation phase have become necessary. For the geographical distributed data sets are appropriate for large scale distributed platforms such as clusters and grids. In the last few years distributed association rules and distributed classifications are proposed in the methods of DDM (Bogorny and Shekhar, 2010).

For the analysing large, heterogeneous and distributed data sets, only a few researchers concerns distributed clustering. Distributed clustering approaches based on the same 2-step process have proposed by recent researches i.e., by aggregating the local results perform partial analysis on local data at individual sites and then send them to the central site to generate the global models. By using same 2-step process we reduce the significantly the amount of information exchanged during the aggregation phase to generate the

automatically correct number of clusters. For this we can use any clustering algorithm to perform the analysis on local data sets. The efficient aggregation phase has been developed on spatial datasets it would be very efficient the result will be reduced by more than 98% of the data set.

Existing system

Two main phases are introduced in existing distributed data mining techniques (1) on local data at individual sites performing partial analysis; (2) by aggregating the local results and generating the global models. Local analysis may produce incorrect and ambiguous global data models. For these two steps are not independent since naïve approaches to the process (Guting, 1994). To minimize the effect of the local results on the global models but not only facilitates their integration, in order to take advantage of mined knowledge. One of the key factors will be affecting the outputs of these techniques. The data that will be collected from different locations by usage of different instruments will be having different formats, features and quality. Moreover centralised data mining techniques will not consider the issues of data driven applications for example scalability in both response time and accurate solutions distributions and heterogeneity. DDM approaches are using ensemble learning which in turn applies varies techniques for combining the results, example majority voting, waited voting and stacking. These approaches are well suited for performing on distributed platforms. For example the incremental algorithms which are used for discovering spatio-temporal patterns by breaking down the search space into a hierarchical structure by addressing its application can be easily optimised on hierarchical distributed system topology. Here, two categories are used one is parallel techniques which required dedicated machines and tools for purpose of communication between parallel processes although these are expensive. These techniques are based on combining that proceeds with a distributed data models on the execution platforms. As the data mining techniques are suffering from the scalability issues the amount of data is growing more in recent years. These will be major issue in the coming years. Although many solutions are proposed they are based only on small improvements for fitting into a particular data.

Clustering which is one of the fundamental techniques in data mining which combine data objects by taking information equipped the data that will described in objects and relationships (Lu and Han, 1993). The main purpose is to make use of similarity measure with in a cluster on differences between clusters for identifying structures in the data. The categories of cluster includes portioning and hierarchical. Different techniques of ongoing clustering algorithms that were given in the literature are proposed. Parallel algorithms are again classified into two categories the first includes the methods which will require different rounds of message passing which in turn requires synchronisation the second includes methods for building local clustering models and sending to a central site for forming global models. The parallel version for hierarchical clustering known as message passing clustering is dedicated to micro array data. Many parallel approaches require either several synchronisations constrain or the entire view of the dataset or both. In hierarchical and partitioning both have some weaknesses. In the majority of the cases k is not known in the k means algorithm in the partitioning class. This is one of the limitations in the partitioning class that is overcome by the hierarchical clustering to define in the stopping conditions in a straight forward manner. By using these two methods data are clustered in to some groups.

Geographical information system (GIS)

By using the GIS the information can locate, pattern, trend, optimum path, identification. Maximize the efficiency of planning and decision making. Information can be distributed in the form of maps, derived maps, statistics, charts etc. Minimize duplication. In GIS the data will be stored in different formats which are based on the data set which was given. It is mainly used to store the information by the features and characteristics. It stores information in the form of land highlights and their attributes. The elements are regularly named focuses, lines, or zones, or as raster pictures. On a guide city information could be put away as focuses, street information could be put away as lines, and limits could be put away as regions, while ethereal photographs or examined maps could be put away as raster pictures. Geographic Information Systems store data utilizing spatial lists that make it conceivable to distinguish the elements situated in any subjective district of a guide (Kazar et al., 2004). For instance, a GIS can rapidly distinguish and outline of the areas inside of a predefined sweep of a point, or the greater part of the lanes that gone through a locale. A standout amongst the most entrenched and well known GIS projects is ESRI's ArcGIS suite. ArcGIS was set up in 1969 initially as an exploration bunch concentrated on land use arranging activities (ESRI, 2012). From that point forward, ArcGIS has become exponentially and is presently the main business GIS programming which consolidates versatile, desktop,

server, and online stages. Up to this point, the permit levels in ArcGIS were sorted out by the naming plan Arc View, Arc Editor, and Arc Info, which indicated the measure of access to apparatuses and functionalities inside of the product. Presently, the distinctive levels are called Basic, Standard, and Advanced, individually, to make the permit levels more instinctive. The Basic level permits access to mapping and intuitive perception functionalities, the Standard level offers those and also multiuser altering and propelled information administration, and the Advanced level offers the greater part of the above, in addition, propelled investigation, top of the line cartography, and broad database administration conceivable outcomes (ESRI, 2012). These diverse levels permit the purchaser to pick the bundle that fits their necessities. Likewise, you can include expansions, for example, Spatial Analyst, 3D Analyst, and Geostatistical Analyst, which are sets of apparatuses for particular undertakings (ESRI, 2012; Brinkhoff et al., 1994). Expansions permit the client to include more propelled capacities without the requirement for moving up to an alternate permit. Be that as it may, one noteworthy issue with GIS (Shekhar et al., 1999) and other ESRI items are costly, which can frequently just be managed by substantial organizations, colleges, and government offices.

Table 1. Historical population of Vijayawada.

Year	Population	±%
1871	8,206	—
1881	9,366	+14.1%
1891	20,224	+115.9%
1901	24,224	+19.8%
1911	32,867	+35.7%
1921	44,159	+34.4%
1931	60,427	+36.8%
1941	86,184	+42.6%
1951	161,198	+87.0%
1961	234,360	+45.4%
1971	344,607	+47.0%
1981	543,008	+57.6%
1991	845,756	+55.8%
2001	1,039,518	+22.9%
2011	1,491,202	+43.5%

Proposed system

Mostly, in rural areas literacy system is low when compared to the urban areas. In urban areas the educational institutes are more because of that population is increasing in the cities. Now-a-days rural areas are also developing as the population is also increasing. But educational institutes are low in number. So many of the people are addicted to the urban culture and there are shifting to cities. To bring awareness among the people

who are living in the rural areas about the education this paper presents the historical report the population growth is rapidly increasing (Table 1).

In Vijayawada, the population rapidly increased from year to year as shown in the Fig. 1. In this paper the comparison starts from year 1800 to 2100. Estimated that population is vary when compared to the present growth.
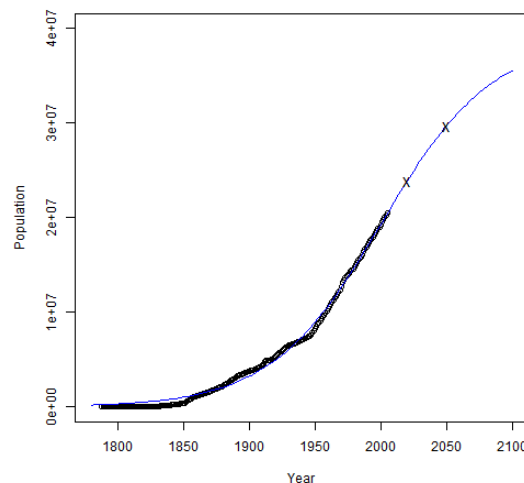

Fig. 1: Population graph in urban area –Vijayawada.

From the analysis of data, in ancient years population is very low; at that literacy rate is also too low. But years passed away awareness about the education among the people is started. In India the population is very high.

Table 2. Number of Mandals in District wise in Andhra Pradesh state.

Sl. No.	Name of District	No. of Mandals
1	Adilabad	52
2	Anantapur	63
3	Chittoor	66
4	East Godavari	58
5	Guntur	57
6	Kadapa	50
7	Karimnagar	57
8	Khammam	46
9	Krishna	49
10	Kurnool	54
11	Mahabubnagar	64
12	Medak	46
13	Nalgonda	59
14	Nellore	46
15	Niazamabad	36
16	Prakasam	56
17	Ranga Reddy	33
18	Srikakulam	38
19	Visakhapatnam	39
20	VIzianagaram	34
21	Warangal	50
22	West Godavari	46
	Total	**1,099**

From region to region the population is varying when compared with different districts in Andhra Pradesh state as shown in the Table 2.

Implementation

In the region of Vaddeswaram the population growth is low when compared to the urban area like Vijayawada. But now the population is increased by years passed away as shown in Fig. 2.
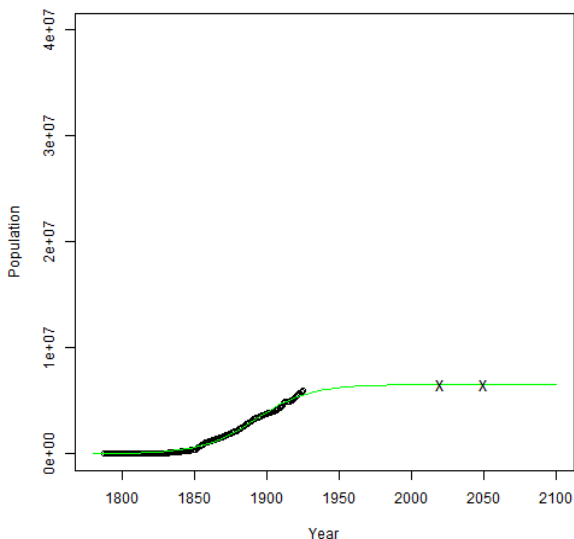

Fig 2: Population growth in rural area –Vaddeswaram.

Compared to Vijayawada the population is same while comparing with the area at the same time. The educational institutes are not sufficient for the population.
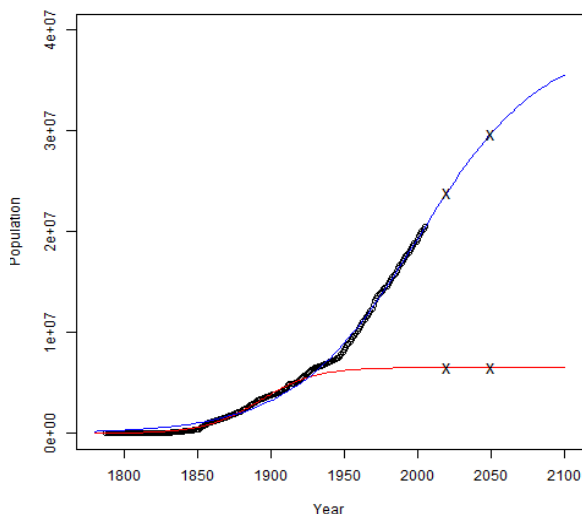
Fig 3: Population comparison graph between urban area-Vijayawada and rural area-Vaddeswaram.

By comparing the urban area -Vijayawada and rural area - Vaddeswaram the population growth increases but educational institutes in Vijayawada is increased with the population (Figs. 3 and 4).
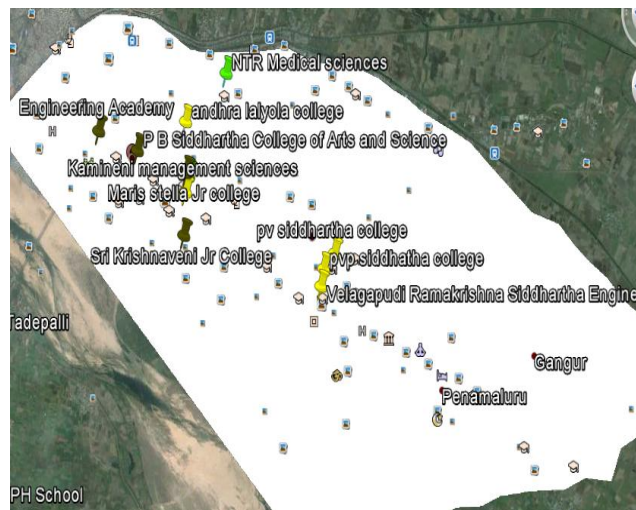

Fig 4: Educational institutes in urban area- Patamata region Vijayawada.

In Vaddeswaram, there is no increment in educational institutes with the growth of population as shown in the Fig. 5. Mainly based on the population of urban area like Vijayawada region there are many educational institutes are developing. Rural areas like Vaddeswaram region even though population is increasing still there are no educational institutes. Hence there is necessity to provide awareness about the educational institutes along with the population.
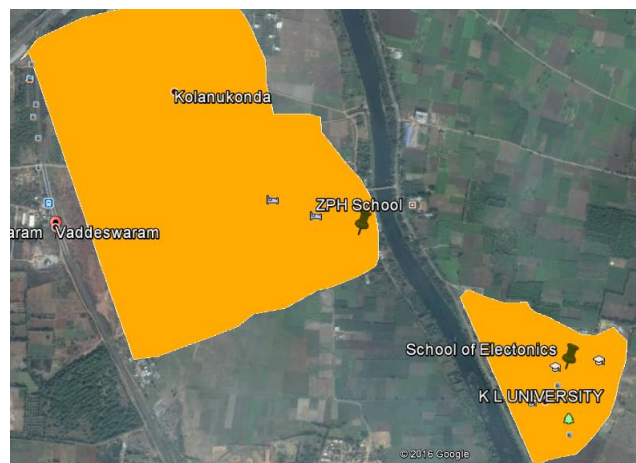

Fig 5: Educational institutes in rural area- Vaddeswaram village.

Consider the maps of Vijayawada and Vaddeswaram to the locations of educational institutes provided with the identification of particular point. Educational institutions

those are located in the urban area like Patamata area in Vijayawada region as shown in the Fig. 4 are very large in number but where as in rural area like Vaddeswaram only very few no of educational institutes are present.

Conclusion

Even though population is increasing in both rural and urban areas still there is increase in number of educational institutes are present in urban areas. But there is still necessity of establishing educational institutes in rural areas in fact rural areas environment is good for education system.

Conflict of interest statement

Authors declare that they have no conflict of interest.

References

Bogorny, V., Shekhar, S., 2010. Spatial and Spatio-Temporal Data Mining, IEEE 10th International Conference on Data Mining (ICDM), Sydney, NSW.

Brinkhoff, T., Kriegel, H.-P., Schneider, R., Seeger, B., 1994. Efficient multi-step processing of spatial joins. Proc. ACM SIGMOD'94 Int. Conf. Management of Data, Minneapolis, MN. pp.197-208.

ESRI, 2012. Algorithms used by Network Analyst. http://help. arcgis.com/en/arcgisdesktop/10.0/help/index.html#//00470 0000053000000

Guting, R.H., 1994. An introduction to spatial database systems. VLDB J. 3(4), 357-399.

Kazar, B.M., Shekhar, S., Lilja, D.J., Vatsavai, R. R., Kelley Pace, R., 2004. Comparing exact and approximate spatial auto-regression model solutions for spatial data analysis. In: Geographic Information Science (Eds.: Egenhofer, M. J., Freksa, C., Miller, H. J.). Third Int. Conf. Geographic Information Science (GIScience2004), October 2004, Adelphi, Maryland, USA. pp.140-161.

Lu, W., Han, J., 1993. Discovery of general knowledge in large spatial databases. Proc. Far East Workshop on Geographic Information Systems, Singapore, June 1993. pp.275-289.

Shekhar, S., Chawla, S., Ravada, S., Fetterer, A., Liu, X., Lu, C.T., 1999. Spatial databases — Accomplishments and research needs. IEEE Transac. Knowl. Data Engg. 11(1), 45-55.