ENHANCED NEAREST GROUP QUERY OPTIMIZATION WITH DATA ANALYTICS IN GEOLOCATION DATA

[1] K. MAHALAKSHMI, [2] V.VAISHNAVI,
[1] Assistant Professor, [2] UG Research Scholar,
[1] Department of B.Com (Business Analytics),

[1] PSGR Krishnammal College for Women Coimbatore, Tamilnadu, India.

Abstract: Spatial data is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial dataset Extracting interesting and useful patterns from Google spatial datasets is extracting the corresponding patterns from traditional numeric and categorical data thanks to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. Spatial data is about instances located during a physical space. When spatial information becomes dominant interest, spatial data processing should be applied. Spatial data structures can facilitate spatial mining. Standard data mining algorithms are often modified for spatial data mining, with a considerable part of pre-processing to require under consideration of spatial information. Initially the set of knowledge points, containing the keyword information of the query object and therefore the query keyword should tend by the User. By Group Nearest query, each nearest point matches a minimum of one among the query keywords of the User. Next, the user wants to rank the selected locations with respect to the sum of distances to nearest interested facilities. As a result, the best location can be obtained from the minimized summed Distance calculation.

***Key words*: Spatial dataset, nearest group query, Spatial mining.**

I. INTRODUCTION

The main objective is to find the best user location .By Group Nearest query, each nearest point matches at least one of the query keywords of the User, the user has to rank the selected locations with respect to the sum of distances to nearest interested facilities. The major objective is to find out the cluster of nearest points or location in the spatial dataset. The Coimbatore spatial dataset will be collected, the dataset will contain details of latitude and longitude of locations. It shows in location in x axis and y axis. Using the dataset, we will be finding out the NEAREST GROUP QUERY it means we are going to fetch the location that the user satisfies all the multiples queries. Using classification algorithm K NEAREST NEIGHBOUR algorithm to cluster the nearest point or location of the attributes. To find out the nearest group query we use the EUCLIDEAN DISTANCE formula by using an existing algorithm.

II. RELATED WORKS

Support of highly performance queries on large volumes of spatial data becomes increasingly important in many application domains, including geospatial problems in numerous fields, location based services, and emerging scientific applications. The emergence of massive spatial data is the proliferation of cost effective and ubiquitous positioning technologies, development of highly resolution imaging technologies, and contribution from an oversized number of community users [1].

Data Classification is the conscious and leads to allocate the level of sensitivity to the information because it is being created, amended, enhanced, stored, or transmitted. The classification of any property should be determined by the extent to the information which is controlled and secured and it's additionally supported its value in terms of worth as a business asset. The classification of all property is indispensable if a corporation is to differentiate between the little value, and which is very sensitive and confidential. When the information is stored whether it is received, created or amended it should be classified at an appropriate sensitivity level. Systems must be used to catch keywords and terms used in classification [2].

A distance can find the objects within a certain distance of a given object. In general the spatial attributes are classified into three major relations they are distance relation, direction relation and topological relation. Topological relation is always non spatial data, so it can require spatial mapping to convert non spatial to spatial data [3].

The recent explosion is the amount of spatial data that requires the specialized systems to handle the massive spatial data. In this paper, we discuss the features and components they should be supported during the system to handle the big spatial data. We review the recent ad that the planet of massive spatial data requires these four components, namely, language, indexing, query processing, and visualization [4].

Spatial Association rules are implications of a single set of data by others. For e.g. in Rajasthan the average income for a person living near rural is Rs.15, 000. It discovers uncovering relationships from spatially related dataset and won't describe the patterns of the database. It's used to find the occurrence of an event Y in the neighbourhoods of another event X in spatiotemporal data [5].

Spatial data mining is the technique to seek out the knowledge from huge geospatial dataset for extracting unknown, necessary spatial relationships, trends or patterns, not stored explicitly in spatial databases [6].

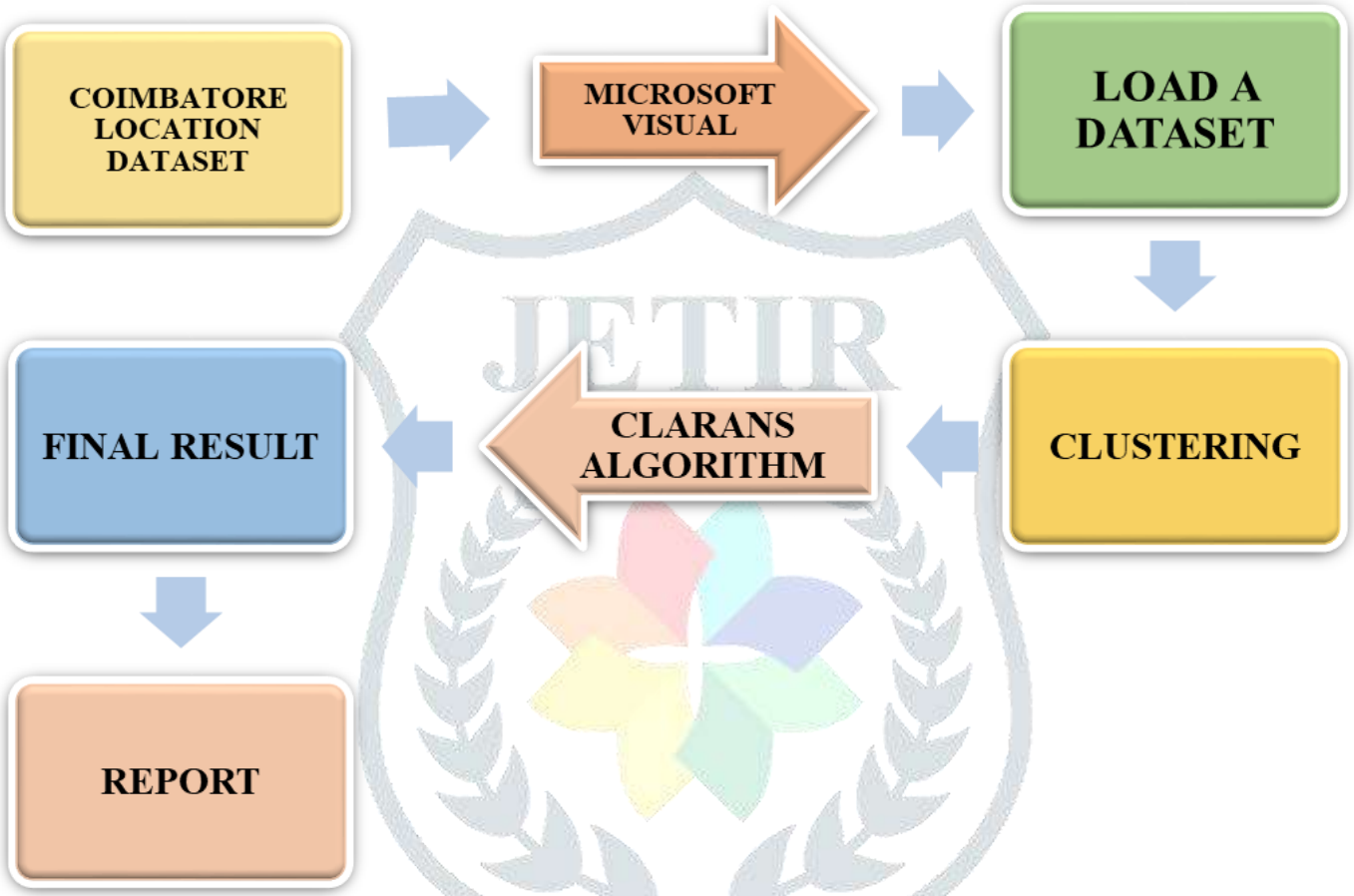
A new algorithm is developed for achieving the efficient classification of knowledge with non a-priori information available about the amount of groups. A performance index is defined the minimising it results in appropriate clustering of the given data. Examples are given to illustrate the procedure, whose convergence is guaranteed. The proposed method, which is not based towards the clusters of any particular shape or size, is compared with the two other clustering techniques [7].

III. OBJECTIVES OF THE STUDY

The present study has framed the following objectives. They are

1. The Dataset is collected based on Coimbatore location.
2. The EUCLIDEAN DISTANCE formula is applied on existing feature algorithms to find the nearest location.

IV. METHODOLOGY



The following Algorithms are used:

- KNN algorithm (k nearest neighbour)
- Clarans algorithm (existing)
- ProMISH algorithm (machine learning)

Here we use the geo-spatial data that contains information about specific locations on earth's surface. As input, the Coimbatore city data base was taken with the following attributes in it, category, place with their dimensions (i.e.) latitude and longitude .Then based on the users requirement we perform multiple queries to the database, the clustered data are formed. Next, we experimented with a couple of data mining algorithms namely existing algorithm (Clarans) and machine algorithm (ProMISH). To find the nearest group query by using Existing algorithm. We have collected data based on Coimbatore location.

IV.1. EUCLIDEAN DISTANCE

The EUCLIDEAN DISTANCE formula is applied on existing feature algorithms to find the nearest location. Beyond its application to distance comparison, squared Euclidean distance is of central importance in statistics, where it is used in the method of least squares, a standard method of fitting statistical estimates to data by minimizing the average of the squared distances between observed and estimated values [10]. The addition of squared distances to each other, as is done in least squares fitting, corresponds to an operation on (unsquared) distances called Pythagorean addition [11]. In cluster analysis, squared distances can be used to strengthen the effect of longer distances [9]. Squared Euclidean distance does not form a metric space, as it does not satisfy the triangle inequality

EUCLIDEAN DISTANCE is computed using the following formula

$$D(p, q) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \ldots + (p_n-q_n)^2}$$

The primary objective of the study is to find out the nearest distance by using spatial dataset and multiple queries. Initially the set of Data points, containing the keyword information of the query object and the query keyword should be given by the User. By Group Nearest query, each nearest point matches at least one of the query keywords of the User. Next, the user wants to rank the selected locations with respect to the sum of distances to nearest interested facilities. As a result, the best location can be obtained from the minimized summed Distance calculation. Clustering Large Applications based upon Randomized Search.

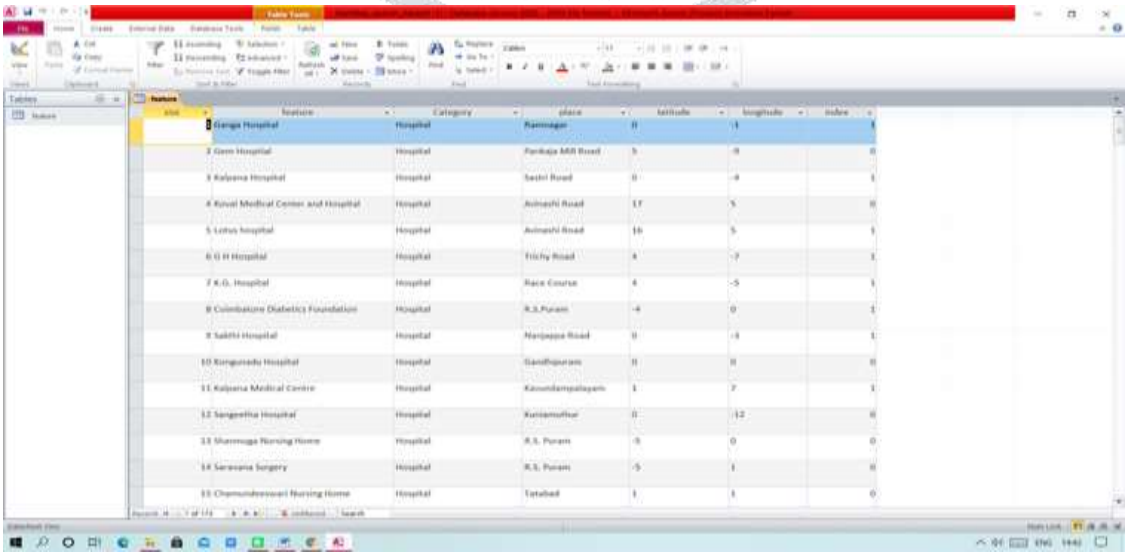
IV.2. CLARANS ALGORITHM

CLARANS are also called as existing algorithms. In CLARANS, the process of finding k medoids from n objects is viewed abstractly as searching through a certain graph. In the graph, a node is represented by a set of k objects as selected medoids. Two nodes are neighbours if their sets differ by only one object. In each iteration, CLARANS considers a set of randomly chosen neighbour nodes as candidate of new medoids. We will move to the neighbour node if the neighbour is a better choice for medoids. Otherwise, a local optima is discovered. The entire process is repeated multiple times to find better [8].

IV.3. TOOLS AND TECHNIQUES

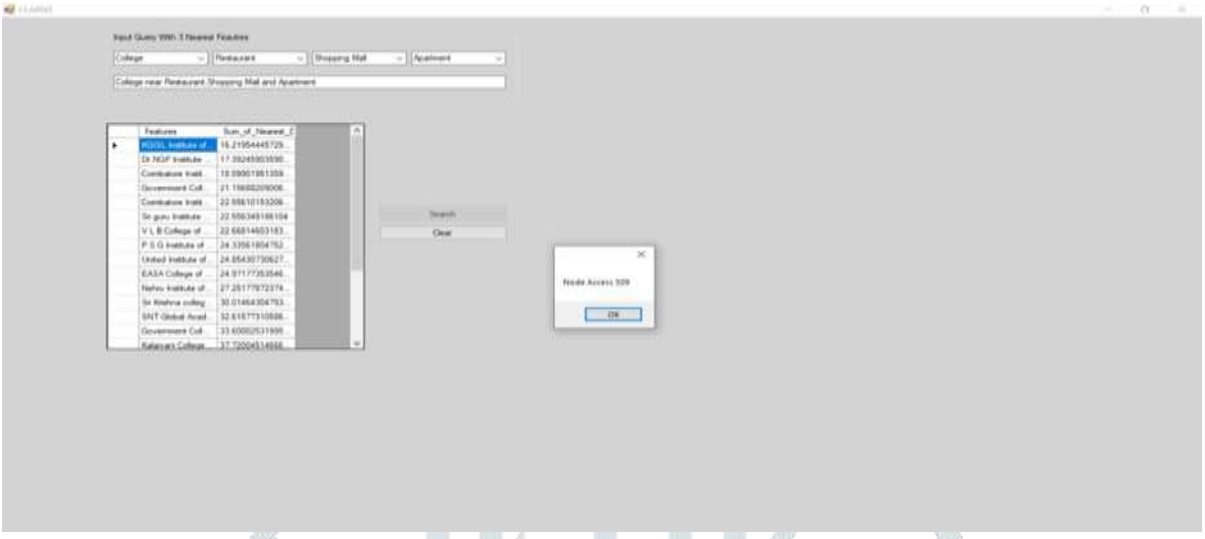
Frameworks have become increasingly popular, through them reuse of design as well as code is achieved for object oriented systems. One relatively new framework is the .NET framework from Microsoft. The .NET framework is part of the larger .NET space. It includes the Common Language Runtime, a large number of partially interfaced, partially class-based frameworks packed into assemblies, and a number of tools. .NET is an open platform for enterprise and web development and it is not bound to a particular programming language. This paper starts with a description of the concept of frameworks. Next we try to cover some of the pieces of .NET framework but due to the extensive size of the .NET not all parts can be covered. The framework perspective of .NET is analysed and we try to focus on the Object Oriented aspects while still covering enough technical parts to let the reader learn about .NET features. We are not trying to paint the .NET features as unique and the only choice on the market, nor do we try to compare .NET as a whole with its competitors. However we can conclude that the .NET framework has advantages over many other frameworks we encountered in the past [12].

IV.4. DATASET



The dataset is named as SPATIAL DATASET enumerated with various locations of the nearest places. The dataset consists of attributes such as FEATURE, CATEGORY, PLACE, LATITUDE, LONGITUDE and INDEX. The feature gives information about the place name ,the category attribute gives information about what kind of the place, the place attribute gives information about the location of the selected place , the latitude and longitude gives the exact geographical location of the selected place.

V. RESULT



(Clarans algorithm result)

The above dataset is classified by an existing algorithm to find out the Euclidean distance. Select multiple features in the input query with three nearest features and click search to list the nearest group query location of features. By Group Nearest query, each nearest point matches at least one of the query keywords of the User. Next, the user wants to rank the selected locations with respect to the sum of distances to nearest interested facilities. As a result, the best location can be obtained from the minimized summed Distance calculation.

VI. CONCLUSION

Different methods of data mining in spatial databases have been outlined in this paper, which has shown that these methods have been developed by two very separate algorithms: Existing algorithm and Machine Learning algorithm. In this project is to find out the nearest group query by using spatial dataset location of multiple queries by using .net tool. By using an existing algorithm the Euclidean distance was found. The Euclidean distance field was used to provide additional spatially relevant predictors to the environment commonly used for mapping. By using an existing algorithm we found out the nearest group query point in the spatial dataset.

REFERENCES:

[1] Aji, A. , Wang, F. , Vo, H. , Lee, R. , Liu, Q. , Zhang, X. , & Saltz, J. (2013). Hadoop-GIS: A high performance spatial data warehousing system over MapReduce. Proceedings of the VLDB Endowment , 6 (11), 1009–1020.10.14778/2536222MM.33.

[2] Craig Wright, in The IT Regulatory and Standards Compliance Handbook, 2008.

[3] Egenhofer, M. 1994. Spatial SQL A Query and Presentation Language. IEEE Transactions and Data Engineering 6, pp.86–95

[4] Eldawy, A. , & Mokbel, M. F. (2015a). The era of big spatial data. Paper presented at the 31st IEEE International Conference on Data Engineering Workshops, Seoul, South Korea, April 13–17.

[5] Mennis, J., and Liu, J. W. 2005. Mining Association Rules in SpatioTemporal data: An Analysis of Urban Socioeconomic and Land Cover Change. Transactions in GIS, 9(1), pp.5-17.

[6] Shekhar, S., Zhang, P., Huang, Y., and Vatsavai, R.R., (2003): Trends in spatial data mining.

[7] Umesh, R. M. (1988). A technique for cluster formation . Pattern Recognition, 21(4), 393–400. doi:10.1016/0031-3203(88)90052-0.

[8] R. Ng and J. Han. CLARANS: A Method for Clustering Objects for Spatial Data Mining. IEEE TRANS. KNOWLEDGE AND DATA ENGINEERING, 2002.

[9] Spencer, Neil H. (2013), "5.4.5 Squared Euclidean Distances", Essentials of Multivariate Data Analysis, CRC Press, p. 95, ISBN 978-1-4665-8479-2.

[10] Randolph, Karen A.; Myers, Laura L. (2013), Basic Statistics in Multivariate Analysis, Pocket Guide to Social Work Research Methods, Oxford University Press, p. 116, ISBN 978-0-19-976404-4.

[11] Moler, Cleve and Donald Morrison (1983), "Replacing Square Roots by Pythagorean Sums" (PDF), IBM Journal of Research and Development, 27 (6): 577–581, CiteSeerX 10.1.1.90.5651, doi:10.1147/rd.276.0577.

[12] Patrik Törnros, Lisa Walterfeldt, Mälardalen Published 2004.