Find the user visiting throughout Google Map and cluster based on most frequent visiting region using k means algorithm

[1]Taruna Sehgal, [2]Er. Yogesh Kumar
[1]Researcher, [2]Assistant Professor
[1]Bhai Gurdas Institute of Engineering and Technology,
[2]Bhai Gurdas Institute of Engineering and Technology

Abstract— **In this paper we are finding the correction location of user based on latitude and longitude is big challenge in spatial database. Clustering is an efficient technique to group together data set to obtain accurate maps. We discuss the k-means clustering algorithm in detail with Spatial Data Mining and how this method can be used. To address these challenges, spatial data mining and geographic knowledge discovery has emerged as an active research field, focusing on the development of theory, methodology, and practice for the extraction of useful information and knowledge from massive and complex spatial databases. Results confirm that k-means clustering can be used to obtainmost visited region of user throughout getting the latitude and longitude and search in google map database.**

***Index Terms*— Clustering, Spatial Data Mining, latitude and longitude**

INTRODUCTION

Modern technology now permits improved acquisition, distribution, and utilization of geographic or geospatial data (Craglia, 2006). There are many popular web based mapping technology. By using the technology, people are able to get any information based on earth coordinates. The technology are very sufficient for personal and daily use, but there is a need to utilize it for corporate or governmental data needs.Geospatial information is critical to promote economic development and improve stewardship of natural resources. The use the mapping technology for corporate or governmental need will involve the preparation of the data and the relationship between geographic data in a very detailed level (e.g. to support the interests of plantation management, taxation, city layout, quality of roads, quality of rivers, etc.).

I. DATA MINING

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. A side from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inferences interesting, metrics, complexity considerations, post-processing, discovered structures, visualization, and online updating.

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amount of data, not the extraction of data itself. It also is a buzz word and is frequently applied to any form of large-scale data or information processing (collection, extraction ,warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence. The popular book "Data mining: Practical machine learning tools and techniques with Java"(which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

II. SPATIAL DATA MINING

Mining spatial co-location patterns [is an important spatial data mining task. A spatial co-location pattern is a set of spatial features that are frequently located together in spatial proximity. To illustrate the idea of spatial co-location patterns, let us consider a sample spatial data set, as shown in Fig. 1. In the figure, there are various spatial instances with different spatial features that are denoted by different symbols. As can be seen, spatial feature + and × tend to be located together because their instances are frequently located in spatial proximity. The problem of mining spatial co-location patterns can be related to various application domains. For example, in location based services, different services are requested by service subscribers from their mobile PDA's equipped with locating devices such as GPS. Some types of services may be requested in proximate geographic area, such as finding the nearest Italian restaurant and the nearest parking place. Location based service providers are very interested in finding what services are requested frequently together and located in spatial proximity. This information can help them improve the effectiveness of their location based recommendation systems where a user requested a service in a location will be recommended a service in a nearby location. Knowing co-location patterns in location based services may also enable the use of pre-fetching to speed up service delivery. In ecology, scientists are interested in finding frequent co-occurrences among spatial features, such as drought, EI Nino, substantial increase/drop in vegetation, and extremely high precipitation. The previous studies on co-location pattern mining emphasize frequent cooccurrences of all the features involved. This marks off some valuable patterns involving rare spatial features. We say a spatial feature is rare if its instances are substantially less than those of the other features in a co-location. This definition of "rareness" is relative with respect to other features in a co-location. A feature could be rare in one co-location but not rare in another. For example, if the spatial feature A has 10 instances, the spatial feature B has 20 instances, and the spatial feature C has 10,000 instances. A is not considered a rare feature in the co-location {A, B} but it is considered a rare feature in co-location {A, C}. Of course, a feature with very small number of instances are often rare in many co-location patterns.

III. COMMON SPATIAL DATA-MINING TASKS

Spatial data mining is a growing research field that is still at a very early stage. During the last decade, due to the widespread applications of GPS technology, web-based spatial data sharing and mapping, high-resolution remote sensing, and location-based services, more and more research domains have created or gained access to high-quality geographic data to incorporate spatial information and analysis in various studies, such as social analysis (Spielman&Thill, 2008) and business applications (Brimicombe, 2007). Besides the research domain, private industries and the general public also have enormous interest in both contributing geographic data and using the vast data resources for various application needs. Therefore, it is well anticipated that more and more new uses of spatial data and novel spatial data mining approaches will be developed in the coming years. Although we attempt to present an overview of common spatial data mining methods in this section, readers should be aware that spatial data mining is a new and exciting field that its bounds and potentials are yet to be defined.Spatial data mining encompasses various tasks and, for each task, a number of different methods are often available, whether computational, statistical, visual, or some combination of them. Here we only briefly introduce a selected set of tasks and related methods, including classification (supervised classification), association rule mining, clustering (unsupervised classification), and multivariate geovisualization.

IV. SPATIAL ASSOCIATION RULE MINING

Association rule mining was originally intended to discover regularities between items in large transaction databases (Agrawal, Imielinski,& Swami, 1993). Let I = {i1, i2, ...,im} be a set of items (i.e., items purchased in transactions such as computer, milk, bike, etc.). Let D be a set of transactions, where each transaction T is a set of items such that T # I. Let X be a set of items and a transaction T is said to contain X if and only if X # T. An association rule is in the form: X) Y, where X I; Y I and X \ Y ¼ £. The rule X) Y holds in the transaction set D with confidence c if c% of all transactions in D that contain X also contain Y. The rule X) Y has support s in the transaction set D if s% of transactions in D contain X [Y. Confidence denotes the strength and support indicates the fre- quencies of the rule. It is often desirable to pay attention to those rules that have reasonably large support (Agrawal et al., 1993). Similar to the mining of association rules in transactional or relational databases, spatial association rules can be mined in spatial databases by considering spatial properties and predicates (Appice, Ceci, Lanza, Lisi, &Malerba, 2003; Han &Kamber, 2001; Koperski& Han, 1995; Mennis& Liu, 2005). A spatial association rule is expressed in the form A) B [s%, c%], where A and B are sets of spatial or non-spatial predicates, s% is the support of the rule, and c% is the confidence of the rule. Obviously, many possible spatial predicates (e.g., close_to, far_- away, intersect, overlap, etc.) can be used in spatial association rules. It is computationally expensive to consider various spatial predicates in deriving association rules from a large spatial datasets. Another potential problem with spatial association rule mining is that a large number of rules may be generated and many of them are obvious or common knowledge. Domain knowledge is needed to filter out trivial rules and focus only on new and interesting findings. Spatial co-location pattern mining is spiritually similar to, but technically very different from, association rule mining (Shekhar& Huang, 2001). Given a dataset of spatial features and their locations, a co-location pattern represents subsets of features frequently located together, such as a certain species of bird tend to habitat with a certain type of trees. Of course a location is not a transaction and two features rarely exist at exactly the same location. Therefore, a user-specified neighborhood is needed as a container to check which features co-locate in the same neighborhood. Measures and algorithms for mining spatial co-location patterns have been proposed (Huang, Pei, &Xiong, 2006; Lu &Thill, 2008; Shekhar& Huang, 2001).

V. DECISION TREE

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represents classification rules.

In decision analysis a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of 3 types of nodes:
Decision nodes - commonly represented by squares
Chance nodes - represented by circles
End nodes - represented by triangles

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities.

Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods.

VI. PROBLEM FORMATION

In my Research work I will develop an android smart phone application in which I will get the

movement of every user using Google map then it pass to the decision tree algorithm where the rules will apply to extract the pattern of region and then pass to the k -means cluster which check that region in clusters if the pattern has already in clusters then that particular region will store in clusters otherwise k means algorithm create a new cluster based on pattern and then put the region on that clusters.

In this research based on regions we will also extract the closeness of user on that particular region using latitude and longitutude value which will pass to the google map api then breadth first ,depth first algorithm will run and check the region based on latitude and longitutude which will return the respected positions of users those users who are more close to each other based on distance on that particular user they can see their name and their sex code also.

By finding the closeness of user of the particular region it helps to the user to know the other user for their interest of visiting regions

VII. LITRATURE SURVEY

[1] 2014 Wen-Yuan Zhu, Wen-ChihPeng [1]In this paper the problem of discovering movement-based communities of users, where users in the same community have similar movement behaviors. Note that the identification of movement-based communities is beneficial to location-based services and trajectory recommendation services. Specifically, I will propose a framework to mine movement-based communities which consists of three phases: 1) constructing trajectory profiles of users, 2) deriving similarity between trajectory profiles, and 3)discovering movement-based communities. In the first phase, I will design a data structure, called the Sequential Probability tree (SP-tree),as a user trajectory profile. SP-trees not only derive sequential patterns, but also indicate transition probabilities of movements. After that propose two algorithms: BF (standing for breadth-first) and DF (standing for depth-first) to construct SP-tree structures as user profiles. To measure the similarity values among users' trajectory profiles, further I will develop a similarity function that takes SP-tree information into account. In light of the similarity values derived, we formulate an objective function to evaluate the quality of communities. According to the objective function derived, propose a greedy algorithm Geo-Cluster to effectively derive communities. To evaluate my proposed algorithms, i have conducted comprehensive experiments on two real data sets. The experimental results show that proposed framework can effectively discover movement-based user communities.

[2] 2014 ShamimRipon [2] In this research Semantic web offers a smarter web service which synchronizes and arranges all the data over web in a disciplined manner. In data mining over web, the accuracy of selecting necessary data according to user demand and pick them for output is considered as a major challenging task over the years. This paper proposes an approach to mapping data over the web 3.0 through ontology and access the required data via an intelligent agent. The agent provides all the searched data related to user query from which user can find desired information. When the user does not have sufficient search parameter, knowledge can be perceived from the information provided by the agent. The derivation of such unknown knowledge from the existing can be achieved by semantic web mining. I present an intelligent agent-based web mining model where users' query is being searched by following existing traditional way, e.g. by Google. The intelligent agent checks the searched data and derives only those are the semantically related to users search parameter. A work-in-progress case study of University Faculty Information presented to examine the effectiveness of the proposed model.

[3] 2011 Anil Sharma*1 , Suresh Kumar2 , Manjeet Singh3[3] In this research Semantic Web Mining is the outcome of two new and fast developing domains: Semantic Web and Data Mining. The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. Data Mining is the nontrivial process of identifying valid, previously unknown, potentially useful patterns in data. Semantic Web Mining refers to the application of data mining techniques to extract knowledge from World Wide Web or the area of data mining that refers to the use of algorithms for extracting patterns from resources distributed over in the web. The aim of Semantic Web Mining is to discover and retrieve useful and interesting patterns from a huge set of web data. This web data consists of different kind of information, including web structure data, web log data and user profiles data. Semantic Web Mining is a relatively new area, broadly interdisciplinary, attracting researchers from: computer science, information retrieval specialists and experts from business studies fields. Web data mining includes web content mining, web structure mining and web usage mining. All of these approaches attempt to extract knowledge from the web, produce some useful results from the knowledge extracted and apply these results to the real world problems. To improve the internet service quality and increase the user click rate on a specific website, it is necessary for a web developer to know what the user really want to do, predict which pages the user is potentially interested in. In this paper, various techniques for Semantic Web mining like web content mining, web usage mining and web structure mining are discussed.

[4] 2014 Abhishek Yadav#1 Gaurav Srivastava#2[3] In this research Semantic Web Mining combines two fast developing research areas: Semantic Web & Web Mining. In this relation, the research intension is to improve on the one hand, Web mining methods with new needs of semantic strategies and on another hand new strategic rule to make it fast and accurate. With tremendous development of WWW, it is making web experience more time spending to user. Hence semantic web mining has become necessary to apply some strategy so that valuable knowledge can be extracted and consequently returned to the user. Data extraction strategies and techniques when applied with web mining will provide a new way result to user query. Clustering will help to provide better satisfaction to user query with less surfing time.

[5] 2012 Kalyani M Raval [4] Data mining is a process which finds useful patterns from large amount of data The process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions - Simoudis 1996 This data mining definition has business flavor and for business environments. However, data mining is a process that can be applied to any type of data ranging from weather forecasting, electric load prediction, product design, etc. Data mining also can be defined as the computer-aid process that digs and analyzes enormous sets of data and then extracting the knowledge or information out of it. By its simplest definition, data mining automates the detections of relevant patterns in database.

[6] 2012 Neelamadhab Padhy1 , Dr. Pragnyaban Mishra 2 , and Rasmita Panigrahi3 [5] In this paper I have focused a variety of techniques, approaches and different areas of the research which are helpful and marked as the important field of data mining Technologies. Each place of operation may generate large volumes of data. Corporate decision makers require access from all such sources and take strategic decisions .The data warehouse is used in the significant business value by improving the effectiveness of managerial decision-making. In an uncertain and highly competitive business environment, the value of strategic information systems such as these are easily recognized however in today’s business environment, efficiency or speed is not the only key for competitiveness. This type of huge amount of data’s are available in the form of tera- to peta-bytes which has drastically changed in the areas of science and engineering. To analyze, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields.

VIII. CONCLUSION

By studying several application of k-means algorithm in both side data mining and pattern recognition where its use as clustering method with data mining giving a promised results and using as segmentation result when it used in pattern recognition and also give a very good result ,so that's mean it's an efficient algorithm in both state. This paper imparts more number of applications of the data mining and also o focuses scope of the data mining which will helpful in the further research. We have developed an algorithm which search the location from google map using latitude and longitude very fast.

REFRENCES

[1] Abugov D, “Oracle Spatial Partitioning: Best Practices (an Oracle White Paper),’’ Oracle Inc., 2004.

[2] Abraham, T. and Roddick, J.F. 1997.‘Discovering meta-rules in mining temporal and spatio-temporal data’. In Proc. Eighth International Database Workshop, Data Mining, Data Warehousing and Client/Server Databases (IDW'97), Hong Kong. Springer-Verlag.30-41. Abraham, T. and Roddick, J.F. 1999.‘Incremental meta-mining from large temporal data sets’.

[3] ‘Active Data Mining’.InProc.First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Quebec, Canada.AAAI Press, Menlo Park, California. 3-8. Bettini, C., Wang, X.S. and Jajodia, S. 1996. ‘Testing complex temporal relationships involving multiple granularities and its application to data mining (extended abstract)’.

[4] DeWitt D J, Ghandeharizadeh S, Schneider D, et al. “The Gamma Database Machine Project,” IEEE Transaction on Knowledge and Data Engineering, 2(1), 1990, pp. 44 -62.

[5] Enviromental Systems Research institute, inc., 1999, Modeling our world the ESRI Guide to Geodatabase design. ESRI Press, [2] Environmental Systems Research Institute, Inc., 1999, Managing ArcSDE™ Services, ESRI Press.

[6] Faloutsos C, Bhagwat P. “Declustering using fractals,” Proceedings of the second international conference on

Parallel and Distributed Information Systems, San Diego, California, United States, 1993, pp. 18- 25.

[7] Faloutsos C, Roseman S., "Fractals for Secondary Key Retrieval," In Proceedings of the 8th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, Philadelphia, Pennsylvania, United States, 1989, pp. 247-252.

[8] Ghandeharizadeh S, DeWitt D J. "Hybrid-range partitioning strategy: a new declustering strategy for multiprocessor databases machines," Proceedings of the sixteenth international conference on Very large databases, Brisbane, Australia, 1990, pp. 481- 492.

[9] Hua K A, Lee C. "An adaptive data placement scheme for parallel database computer systems," Proceedings of the 16th International Conference on Very Large Databases, 1990, 1990, pp. 493-506.

[10] In Computers and Biomedical Research. 15, 164-187. Chakrabarti, S., Sarawagi, S. and Dom, B. 1998.'Mining surprising patterns using temporal description length'.

[11] In Proc. Twenty-Fourth International Conference on Very Large databases VLDB'98, New York, NY. Morgan Kaufmann. 606- 617. Chen, X. and Petrounias, I. 1998.'Language support for temporal data mining'.InProc.Second European Symposium on Principles of Data Mining and Knowledge Discovery, PKDD'98, Springer-Verlag, Berlin.282-290. Chen, X., Petrounias, I. and Heathfield, H. 1998.'Discovering temporal association rules in temporal databases'.In Proc. International Workshop on Issues and Applications of Database Technology (IADT'98), 312-319.

[12] In Advances in Database Technologies, Proc.First International Workshop on Data Warehousing and Data Mining, DWDM'98. Y. Kambayashi, D.K. Lee, E.-P. Lim, et al. (eds.), Lecture Notes in Computer Science 1552, Springer-Verlag, Berlin.41-54.Agrawal, R. and Psaila, G. 1995.

[13] Li J Z, Srivastava J, Rotem D. "CMD: A Multidimensional Declustering Method for Parallel Database Systems," In Proceedings of the 18th International Conference on Very Large Data Base Conference, Canada,1992, pp. 1-14.

[14] Moon B, Jagadish H V, Faloutsos C, et al., "Analysis of the clustering properties of the Hilbert space-filling curve," IEEE Transactions on Knowledge and Data Engineering, 13(1), 2001, pp. 124 -141.

[15]Zhao C Y, Meng L K, Lin Z Y., "Spatial Data Partitioning Towards Parallel Spatial Database System," Geomatics and Information Science of Wuhan Univeristy, vol. 31(11), 2006, pp. 391-394. (in Chinese)