Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 3 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Clustering on Spatial Data Sets using Extended Linked Clustering Algorithms

By K. Lakshmaiah

Sir Vishveshwaraiah Institute of Science & Technology

Abstract- Various clustering algorithms (CA) have been reported in literature, to group data into clusters in diverse domains. Literature further reported that, these CAs work satisfactorily either on pure numerical data or on pure categorical data and perform poorly on mixed numerical and categorical data. Clustering is the process of creating distribution patterns and obtaining intrinsic correlations in large datasets by arranging the data into similarity classes. The present work pertains to reviewing the available research papers on clustering spatial data. In a web perspective, a detailed inspection of grouped patterns and their belonging to well known characters will be very useful for evolution of clusters. The review work is split into spatial data mining, clustering on spatial data sets and extended linked clustering. This review work will enable the researchers to make an in depth study of the till date research work on above areas and will pave way for developing extended linked clustering algorithms with a view to find number of clusters on mixed datasets to produce results for several datasets.

Keywords: *extended linked clustering, optimum criterion, coherence, databases, spatial datasets.*

GJCST-C Classification: *I.5.3*

CLUSTERINGONSPATIALDATASETSUSINGEXTENDEDLINKEDCLUSTERINGALGORITHMS

Strictly as per the compliance and regulations of:

Clustering on Spatial Data Sets using Extended Linked Clustering Algorithms

K. Lakshmaiah

Abstract- Various clustering algorithms (CA) have been reported in literature, to group data into clusters in diverse domains. Literature further reported that, these CAs work satisfactorily either on pure numerical data or on pure categorical data and perform poorly on mixed numerical and categorical data. Clustering is the process of creating distribution patterns and obtaining intrinsic correlations in large datasets by arranging the data into similarity classes. The present work pertains to reviewing the available research papers on clustering spatial data. In a web perspective, a detailed inspection of grouped patterns and their belonging to well known characters will be very useful for evolution of clusters. The review work is split into spatial data mining, clustering on spatial data sets and extended linked clustering. This review work will enable the researchers to make an in depth study of the till date research work on above areas and will pave way for developing extended linked clustering algorithms with a view to find number of clusters on mixed datasets to produce results for several datasets. This work is likely to assist in deciding which clustering solution to use to obtain a coherent data solution for a particular character experiment. Further it could be used as an optimal tool to guide the clustering process towards better and character interpretable meaningful solutions. The major contribution of present work is to present an in depth literature review of research in the areas of Spatial data mining, clustering on spatial data sets and extended linked clustering with a view to assist researchers to develop optimum extended linked clustering and to develop optimum extended linked clustering algorithms for clustering process, towards better and character interpretable meaningful solutions.

Keywords: extended linked clustering, optimum criterion, coherence, databases, spatial datasets.

I. Introduction

Data mining (DM) is the process of effectively extracting data in the form of knowledge discovery, which provides useful and helping guide for information processing that can be utilized in varieties of applications [1]. Different types of DM techniques are augmented for application in the fields of science, research, medicine etc. Data bases (DB) comprise of terabytes, which are capable of storing huge mine of heterogeneous data. For effective DM pre mining and post mining procedures are prescribed. DM models consist of predictive models and descriptive models. AIS algorithm was the first association rule mining algorithm. The main limitation of this is that, it requires large space and is there by not efficient. Apriori algorithm is more efficient than AIS. Spatial data mining (SDM) is the process of arriving at potentially helpful patterns from large spatial data sets. Spatial data base can be best understood as the data relating to price range of the houses with nearby spatial features like beaches, different geographical regions in a city etc. Image spatial data base deals with image database dealing entirely with pictures or images [2]. The algorithms used for SDM include generalization – based methods. This is based on the concept of data from more than a few evidences from a concept level to its higher concept level and tracking knowledge from the widespread database. Collecting data and using it for knowledge discovery are two independent factors. It is often said that, we are data rich, but information poor. We require tools for automatic summarization of data, discovery of patterns in the data for analyzing and interpretation [3]. Many techniques are being used for data mining with Geographic Information Systems (GIS) for carrying out spatial data analysis of geographic data. The two approaches in common use are, first comes first learning on spatial data bases, where as the second is based on spatial statistics [4]. In SDM we must consider the spatial relations between objects. SDM is used in geo marketing, environmental studies, risk analysis etc. another equally important area in SDM is visual data mining (VDM) which applies visual human perception for mining large data sets. This mostly comprises of presenting huge data simplified to a graphical format [5], resulting in discovery of valuable patterns in very large data bases. In the fields of clustering and visualization pixel maps, a new way of displaying dense point sets are often used. However in SDM the issues and challenges involved need careful analysis [6]. The voluminous data and its handling and analysis poses a major challenge in SDM. Another challenge is extraction of implicit knowledge not explicitly stored in spatial data bases. The Introduction of computer capabilities and emergence of I.T, have led to enormous amount of data relating to science and engineering [7], which is normally made available through internet. This has led to transforming many areas from data poor to data rich stage.

Clustering algorithms (CA) are proposed by researchers in the fields DM [8]. The advantage is that, the clustered data is easy to understand and normally it does not confine to the shapes of the clusters. Various algorithms on clustering namely DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN etc are proposed. In simple

Author: e-mail: klakshmaiah78@gmail.com

words clustering is grouping of objects or data into meaningful sub classes. Among various CA, density based algorithms are more efficient in identifying clusters with varied densities. However there exist constraints on application of CA in DM which will be presented in subsequent sections. Recent research in clustering includes, proposing Adaptive flocking algorithms for spatial clustering [9]. This works on use of new Swarm Intelligence (SI) techniques. Many authors have presented huge survey on application of above algorithms in clustered SDM [10]. Most of the clustered SDM algorithms were applied to clustered SDM in the fields of spatial cancer databases [11]. The various assumptions and requirements needed for applying the clustered SDM are [12] huge data to exist for applying algorithms, and the developed algorithms should be capable of handling irregular shapes, insensitive to bulky amounts of noise etc. Clustering algorithms are modified as two density based spatial clustering algorithms, especially when very huge large databases are to be handled [13]. The accomplishments and research needs of spatial data mining focus on location prediction, spatial outlier detection, co-location mining etc [14]. Researchers have also proposed and presented efficient and effective clustering methods for DM [15]. Novel methods based on delaunay triangulation were tried in the area of clustered SDM [16].

A recent development in mining spatial structural data mining is link mining (LM). This technique deals with mining richly structured data sets, where the objects are linked in some way [17]. The links include certain patterns, which could not be analyzed in traditional DM techniques. Web and hypertext mining, social networks, security and law enforcement data bases, bibliographic citations etc are best mined using linked clustered algorithms. LM is an emerging area and is an instance of multi-relational data mining. LM encompasses a range of tasks including descriptive and predictive modeling. With the introduction of link concepts, new issues such as number of links, types of links, inferring the existence of links etc arise. In the area of applying LM to web page DM, the algorithms used are based on citation relation between web pages.

Linked data base mining model of bibliographic description [18] is derived from ideas based on schema bib extended group and was found useful in above mining task. Though linked clustered SDM is an emerging area of research in DM, approaches to visualizing linked data is [19] assuming greater dimensions. These studies are mostly confined to semantic web community. The main issue involved is the lack of technical knowledge and an understanding of the semantic technology in the use of web data. This linked data base mining is an extended concept of linked clustered SDM and may be termed as Extended Linked Clustered SDM (ELC SDM). These applications result in further challenges in the areas of mobile devices and the reduction in cost of producing sensors. When a Uniform Resource Identifier (URI) is de referenced, a response is returned and is characterized by an extended hypertext markup linked clustered representation of the resource, managing the life cycle of linked data. Extended linked data with LOD2 stack forms the latest research in the field of extended linked clustered SDM [20]. The LOD2 stack is an integrated distribution of aligned tools, supporting the life cycle of linked data from extraction via enrichment, interlinking and fusing. Recent applications of ELCSDM are seen in educational linked data bases.

In the area of Web databases related to education, highlighting how such extended links form a globally addressable network of resources for education [21] is very important. Adopting the linked DM to extended linked DM needs a minor integration effort, to improve the global cohesion of education networks. DM techniques have now emerged to the extent of application of extended linked clustered SDM techniques, for arriving at meaningful and useful conclusions from the huge vast data base.

II. Literature Review

The literature review is considered from three aspects. First one deals with general methods of data mining. The second is on clustering spatial data sets and third is on linked clustering with Spatial reference to extended linked clustering. The detailed survey is given below.

Amitkumar patnaik etal[1], worked on different types of data mining techniques for powerful data mining ranging from commercial to scientific applications. Their studies included the areas of warehouse and online analytical processing, along with various data mining models. They have tried different data fields.

M. Hemaltha etal [2], made extensive survey on knowledge discovery in spatial data mining. Spatial data mining is the process of discovering, motivating and obtaining previously unknown, but potentially helpful patterns from large spatial datasets. Extracting interesting and useful patterns from spatial datasets is more tricky than extracting the parallel patterns from established numeric and definite data, due to the complexity of spatial data types, spatial relationships, and spatial auto correlations. They focused on the sole features, that distinguish spatial data mining from traditional data mining. Major activities and research needs in spatial data mining research were discussed by them. They listed the applications and techniques, issues and challenges on spatial data mining. They concluded that, spatial data mining is a promising field with rich research results and many challenging issues.

Neelamadhab padty etal [3], made extensive work on data mining applications and future scope, and published their work. They focussed on variety of techniques among different areas. Their work concentrated on MNC data collection and mining the data, collected from different places and countries. They have used data ware house, by improving the effectiveness of managerial decision making. They mainly worked on mining huge data. Their paper concentrated on, number of applications of data mining, which formed a basis for further research in this area.

Karine etal [4], made extensive survey of spatial data mining methods from available data bases and developed statistical point of views. Their work included review of data mining methods, combined with GIS for conducting spatial analysis of geographic data. Their conclusions included, listing of main differences between the two common approaches, namely first come first learning, and spatial statistics and also the elements they have, in common.

Danial etal [5], developed pixel based visual mining of geo spatial data. In many application domains, data is collected and referenced by geo-spatial location. Spatial data mining, or the discovery of interesting patterns in such databases, is an important capability in the development of database systems. A noteworthy trend includes increasing size of data sets in common use, such as records of business transactions, environmental data and census demographics. These data sets often contain millions of records, or even far more. This situation creates new challenges in coping with scale. For data mining of large data sets to be effective, it is also important to include humans in the data exploration process and combine their flexibility, creativity, and general knowledge with the enormous storage capacity and computational power of today's computers. Visual data mining applies human visual perception to the exploration of large data sets. Presenting data in an interactive and graphical form often fosters new insights, encouraging the formation and validation of new hypotheses to the end of better problem-solving and gaining deeper domain knowledge. They gave a short overview of visual data mining techniques especially for analyzing geo-spatial data. They provided examples for effective visualizations of geo-spatial data in important application areas such as consumer analysis and census demographics.

Krzysztof etal [6], did extensive work on spatial data mining and reviewed the progress made so far along with associated issues and challenges. Since huge amount of data exists in various applications, analysis of this huge data far exceeds human ability. Data mining is extended to spatial data bases. They have summerised recent works on spatial data mining from spatial data generalization to spatial data clustering, mining spatial association rules etc. They concluded that spatial data mining is a promising field, with fruitful research results and many challenging issues.

Jiawei etal [7], identified research challenges for data mining in science and engineering .With the advent of IT and CSE fields fast developing ,data is collected and stored in a massive scale. This data is made available globally through networks. This has led to, developing data rich data base, calling for new data intensive methods, to conduct research in science and engineering. Their work focused on issues including (1) information network analysis, (2) discovery, usage, and understanding of patterns and knowledge, (3) stream data mining, (4) mining moving object data, RFID data, and data from sensor networks, (5) spatio temporal and multimedia data mining, (6) mining text, Web, and other unstructured data, (7) data cube-oriented multidimensional online analytical mining, (8) visual data mining, and (9) data mining by integration of sophisticated scientific and engineering domain knowledge.

M. Parimala etal [8], made a survey on density based clustering algorithms for mining large spatial data bases. Density based clustering algorithm is one of the primary methods for clustering in data mining. The clusters which are formed, based on the density are easy to understand and it does not limit itself to the shapes of clusters. They gave a detailed survey of the existing density based algorithms namely DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN and DBCLASD based on the essential parameters needed for good clustering algorithms. They analyzed the algorithms, in terms of the parameters essential for creating meaningful clusters.

Gianluigi etal [9], developed an adaptive flocking algorithm for spatial clustering. This algorithm was based on the use of new swarm intelligence techniques (SI). SI is a new emerging area where a problem can be solved by using a set of biologically inspired agents exhibiting a collective intelligent behaviour. They have applied this algorithm to two synthetic data sets and its performance was comparable with other algorithms.

Jiawei etal [10], brought out current status and future of frequent data mining, in their research paper. Frequent pattern mining has been a focused theme in data mining research for over a decade. Abundant literature has been dedicated to this research and tremendous progress has been made ranging from efficient and scalable algorithms for frequent item set mining in transaction databases, to numerous research frontiers, such as sequential pattern mining, structured pattern mining, correlation mining, associative classification, and frequent pattern-based clustering, as well as their broad applications. They provided a brief overview of the current status of frequent pattern mining and discussed a few promising research directions. They made it clear that, frequent pattern mining

Global Journal of Computer Science and Technology (C) Volume XVIII Issue III Version I Year 2018

research has substantially broadened the scope of data analysis and will have deep impact on data mining methodologies and applications in the long run. However, there are still some challenging research issues that need to be solved before frequent pattern mining can claim a corner stone approach in data mining applications.

Rituchauhan etal [11], worked on data clustering methods for discovering clusters in spatial cancer databases. They outlined data analyzing tools and data mining techniques to analyze medical data as well as spatial data. The spatial data base is formed by grouping the objects into clusters. Their study focused on discrete and continuous spatial study, focused on discrete and continuous database on which clustering techniques are applied to form clusters. Classical clustering and hierarchical clustering on the spatial data sets to generate efficient clusters formed their main work. Their experimental results were reported which exhibited certain facts that are evolved and cannot be otherwise retrieved from raw data.

Sundararajh etal [12], studied on spatial data clustering algorithms in data mining. Heavy and huge databases have produced interests in the area of data mining. Useful information can only be obtained after clustering the data. Through this process, the hidden patterns or useful subgroups can be identified. They used spatial clustering approach for investigations. They developed fast working and effective algorithms for extraction of information, trends etc from the database. They presented the essential features of clustering algorithms which include scalability, ability to recognize irregular shapes, insensitive to bulky noises etc. The major contribution of their research work was to help researchers to come up with needy techniques to cluster the spatial data effectively.

Xin etal [13], compared density based spatial clustering algorithms for large datasets. The two density methods chosen by them were, density based spatial clustering algorithms and density based clustering algorithms. The two methods are described in detail and a comparison of algorithms was made.

Shastistekar etal [14], worked on spatial data mining, which is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets. They have identified the research needs of spatial data mining, in the areas of location, prediction, spatial outlier detection, co-location mining and clustering.

Rayman etal [15], developed effective clustering methods for spatial data mining. Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. They explored whether, clustering methods have a role to play in spatial data mining. To this end, they developed a new clustering method called CLAHANS, which is based on randomized search. They also developed two spatial data mining algorithms that use CLAHANS. Their analysis and experiments showed that, with the assistance of CLAHANS, these two algorithms are very effective and can lead to discoveries that are difficult to find with current spatial data mining algorithms. Furthermore, experiments conducted, to compare the performance of CLAHANS with that of existing clustering methods showed that, CLAHANS is the most efficient.

Insooking etal [16], used Delaunay triangulation for spatial data mining, with a view to discover significant pattern which may implicitly exist in huge data bases. They have used the SMTIN (spatial data mining by triangulated irregular network) method, which is based on Delaunay triangulation. Its advantages over the Previous ones were described, which include identification of sophisticated patterns and heirarchical structure of cluster distribution, knowledge of prior nature of distribution is not required, requirements of distribution is not required, requires less CPU processing time. It is not ordering sensitive and handles effectively outliers.

Lise etal [17], worked on key challenges on data mining. A key challenge for data mining is tacking the problem of mining richly structured datasets, where the objects are linked in some way. Links among the objects may demonstrate certain patterns, which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical models. Recently there has been a surge of interest in this area, fueled largely by interest in web and hypertext mining, and also by interest in mining social networks, security and law enforcement data, bibliographic citations and epidemiological records.

Carol [18], worked on developing OCLCs linked data models of bibliographic description.

This document describes a proposed alignment between BIBFRAME and a model being explored by OCLC with extensions proposed by the Schema Bib Extend project, a W3C-sponsored community group tasked with enhancing Schema.org to the description of library resources. The key result is that the two efforts are complementary except for some common vocabulary required for the most important entities and relationships. The analysis presented was prompted by the call at the end of the December 2012 BIBFRAME Early Experimenters Meeting, for a set of Point or Position papers that worked out technical issues and made recommendations for a number of sketchy, difficult, or controversial aspects of the BIBFRAME model. The description was based on a small dataset presented in the entity in the Appendix, and the analysis was based on a larger dataset derived from the application of a mapping algorithm from MARC to BIBFRAME on all of World Cat.org. This draft is being released as an OCLC report, but it was intended to be read as a working paper for the BIBFRAME community.

Krzysztof etal [19], worked on visualizing linked data, and approaches for achieving the same. Their survey covered large, distributed and interlined networks of information fragments contained within desparate data sets as provided by unique data publishers. The data was published in a format which was machine readable. This data was linked to other external data. They presented a survey of existing approaches for handling web enabled linked data.

Soren etal [20], did extensive work on managing the life cycle of linked data with LOD2 stack. The LOD2 Stack is an integrated distribution of aligned tools which support the whole life cycle of Linked Data from extraction, authoring/creation via enrichment, interlinking, fusing to maintenance. The LOD2 Stack comprises new and substantially extended existing tools from the LOD2 project partners and third parties. The stack is designed to be versatile, for all functionality web cleared interfaces, which enables the plugging in of alternative third-party implementations. The architecture of the LOD2 Stack is based on three pillars: (1) Software integration and deployment using the Debian packaging system, (2) Use of a central SPARQL end point and standardized vocabularies for knowledge base access and integration between the divergent tools of the LOD2 Stack, and (3) Integration of the LOD2 Stack user interfaces based on REST enabled Web Applications. These three pillars comprise the methodological and technological framework for integrating the very heterogeneous LOD2 Stack components into a consistent framework. In their article they described these pillars in more detail and gave an overview of the individual LOD2 Stack components. The article also included a description of a real-world usage scenario in the publishing domain.

Matheu etal [21], worked on assessing the educational linked data land -scape. They presented a preliminary study of available web datasets related to education, providing an overview of this area and, more importantly, highlighting how such linked datasets form a globally addressable network of resources for education. As expected, a certain level of heterogeneity was found. They also showed how a minor integration effort can improve the global cohesion of such networks of educational web data.

III. Issues and Challenges

1. Since vast huge data base is to be handled, it is very difficult to identify the initial parameters like number of clusters, shape and density of clusters.
2. Since the shape of clusters may be in random manner, discovery of clusters among arbitrary shapes poses a challenge.
3. A good efficiency is to be achieved among large data bases.
4. The various algorithms like a) density based spatial clustered algorithms (DBSCN) ,varied density based spatial clustered algorithm (VDBSCAN) etc have to be carefully written to achieve meaningful end results.
5. Analysing linked clustered data is mostly restricted to web community. The lack of technical knowledge limits users in their ability to interpret and make use of webpage data.
6. Careful analysis is to be made before presenting linked data visualization.
7. The extended linked cluster SDM is an extended version of linked clustered SDM and is arrived at through minor modifications of linked clustered data.

IV. Scope and Objective of Present Work

Though DM is not a new concept and has been in use since a long time, with the advent of I.T enabled computer services, huge databases are created and handling this voluminous data has called for newer techniques, revised algorithms in the field of DM. This resulted in analyzing clustered linked spatial data and the latest trend is handling web based URL through extended linked clustered spatial data. The scope of present work is to make a review of the research work carried out in above areas and to develop extended linked clustered spatial data mining (ELCSBM).

V. Formulation of Problem

Voluminous data collected and stored is drawn from different geographical areas, having or not having similarities and links among them. Handling this data meaningfully and efficiently calls for a systematic analysis of data through spatial clustering the data, grouping the same as per the links present among them and extending these procedures ultimately for web enabled data though extended linked clustered SDM, has now become the center of research in the field of data mining. Hence the formulation of the problem.

VI. Present Work

The present work consists of making a detailed literature review on data mining with special reference to SDM, clustered SDM, linked clustered SDM along with a detailed understanding of the various algorithms used on issue basis. The work ends up with listing modified algorithms to be used for extended linked clustered spatial data mining (ELCSDM) operations. The various results are discussed and the important conclusions are listed.

VII. Algorithm used

The clusters formed based on density of database can be analyzed with ease, without confining to shapes of clusters. The various density based algorithms in use are DBSCAN, VDBSCAN, DVBSCAN,

Global Journal of Computer Science and Technology (C) Volume XVIII Issue III Version I Year 2018

ST DBSCAN, DBCLASD etc, and are briefly discussed here under.

a. Density based spatial cluster algorithms with website (DBSCAN) discovers clusters with arbitrary shapes with minimum number of input parameters such as radius of the clusters and minimum points required inside the cluster. The related algorithm consists of
   i. Selecting an arbitrary point.
   ii. Retrieving all points which are density reachable from the arbitrary point.
   iii. If the point is a core point, a cluster is formed.
   iv. If it is a border point the next point is considered.
   v. The process is continued till at the points are processed.

This algorithm requires only two input parameters and discovers clusters of arbitrary shapes. It holds good for large SDB.

b. Varied density based spatial clustering algorithms with noise (VDBSCAN) detects clusters with varied density, where the DBSCAN fails, and is capable of selecting several values of input parameters. The related algorithm consists of
   i. Calculating and storing K-dist for each project and partitioning the k-dist point.
   ii. Calculating the number of densities.
   iii. Selecting the parameter for each density.
   iv. Scanning the data for different densities.
   v. Displaying the valid cluster with respect to the corresponding density.

This algorithm helps in finding meaningful clusters having varied densities.

c. Density based algorithm for discovering density varied clusters in large spatial data bases (DVBSCAN) is a pioneer density based clustering algorithm which detects clusters with different shapes and sizes, but fails to detect clusters with varied densities that exist within the clusters.

This algorithm is capable of handling local density variations that exist within the clusters.

d. Distributed based clustering algorithm for mining large spatial data bases (DBCLASD) is a new clustering algorithm which is capable of detecting clusters with arbitrary shapes without calling for input parameters. The efficiency of this algorithm in handling huge database is satisfactory.

Link mining is a newly emerging research area in data mining and is mostly used in hypertext, web mining. It is a multi relational DM technique specializing analysis of links present in the spatial cluster data bases. Link mining does a range of tasks such as descriptive and predictive modeling. To perform these operations link mining requires new data mining algorithms dealing with predicting the number links, predicting type of link between two objects, finding co-reference and subgraph patterns. The algorithms that are commonly used for linked clustered spatial data mining (LCSDM) are given here under.

i. Select hypertext and webpage classification, which has its roots in information retrieval community.
ii. Define the features of the links to be searched for, in the web data base.
iii. Obtain the links and their characteristics.
iv. Identify the incoming and outgoing links.
v. Label the category of the web page, based on the features of the link.
vi. Use the link information such as anchor text and neighboring text around each link and obtain categorization results.

A modified approach based on extended linked clustered SDM (ELCSDM) to mine data present in hypertext and link mining combines techniques from inductive logic programming with statistical learning algorithm to construct features for related documents.

The algorithm for ELCSBM is presented here under.

i. Instead of using words in a hypertext document, make use of anchor text, neighbouring text, capitalized words and alphanumeric words.
ii. Using above, propose a combined model for text classification to form links and clusters.
iii. Select the features of the links to be searched for, in the converted web data.
iv. Define the features of the links in the web data.
v. Identify the incoming and outgoing links.
vi. Label the category of the webpage based on the features of the link.
vii. Use the link information such as anchor text and neighboring text around each link and obtain categorization results.

A suitable machine language can be chosen and coding can be written, to run the same to visualize the results.

VIII. Results and Discussions

1. A detailed review of literature dealing with classical data mining, spatial data mining, clustered spatial data mining, linked clustered spatial data mining is presented in literature review.
2. The various algorithms used for DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN, DBCLASD etc are briefly discussed and presented with their merits and demerits.
3. The algorithms used for identifying linked spatial DBM are discussed separately along with their merits and demerits, to mine useful information from the web based huge data bases.
4. The latest trend in DM relating to web data is, extended/modified linked clustered spatial data

mining (E/MLCSDM). The related algorithms for this technique are also presented.

5. The advantage of this modified/extended algorithms is that, more meaningful correlations and results can be obtained using these extended algorithms.

IX. Conclusions

The major contribution of present work is to review and understand the "as on today research status" on DM starting from classical DM to E/MLCSDM. The proposed algorithms for E/MLCSDM can be altered based on issues , and suitable coding can be written, which when run ,gives useful and meaning full results.

References Références Referencias

1. Amit Kumar Patnaik, "Data Mining and Its Current Research Directions", Information Sciences, Volume 181,Issue 7, 1 April 2011, Pages 1264-1284
2. Dr. M. Hemalatha, "A Recent Survey on Knowledge Discovery in Spatial Data Mining", International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011.
3. Neelamadhab Padhy, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.
4. Karine Zeitouni, "A Survey of Spatial Data Mining Methods Databases and Statistics Point of Views", PRiSM Laboratory-University of Versailles 45, avenue des Etats-Unis - F-78 035 Versailles Cedex.
5. Daniel A. Keim, "Pixel Based Visual Mining of Geo-Spatial Data", AT&T Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971, USA.
6. Krzzsztof, " spadial data mining : progress and challenges survey paper", school of computer science,simon fraser university, Burnaby B.C., Canada.
7. Jiawei Han, "Research Challenges for Data Mining in Science and Engineering", University of Illinois at Urbana-Champaign.
8. M. Parimala, "A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases", International Journal of Advanced Science and Technology Vol. 31, June, 2011.
9. Gianluigi Folino, "An Adaptive Flocking Algorithm for Spatial Clustering", Via Pietro Bucci cubo 41C c/o DEIS, UNICAL, 87036 Rende (CS), Italy.
10. Jiawei Han, "Spatial Clustering methods in data minings: A survey", Burnaby, BC. Canada V5A 1s6.
11. Ritu Chauhan, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010.
12. Sundararajan S, "A Study on Spatial Data Clustering Algorithms In Data Mining", International Journal of Engineering And Computer Science Volume1 Issue 1 Oct 2012 Page No. 37-41.
13. Xin Wang, "A Comparative Study of Two Density-Based Spatial Clustering Algorithms for Very Large Datasets", Department of Computer Science, University of Regina, Regina, SK, Canada S4S 0A2.
14. Shashi Shekhar, "Spatial Data Mining", Department of Computer Science and Engineering, University of Minnesota 4-192, 200 Union ST SE, Minneapolis, MN 55455.
15. Raymond T. Ng, "Efficient and Effective Clustering Methods for Spatial Data Mining", Proceedings of the 20th VLDB Conference Santiago, Chile, 1994.
16. In-So Kang, "A Spatial Data Mining Method by Delaunay Triangulation", Kumjeong-Gu, Jangjeon-Dong, Pusan, Korea 609-735.
17. Lise Getoor, "Link Mining: A New Data Mining Challenge", Dept. of Computer Science/UMIACS University of Maryland College Park, MD 20742.
18. Carol Jean Godby, "The Relationship between BIBFRAME and OCLC's Linked-Data Model of Bibliographic Description: A Working Paper", Senior Research Scientist OCLC Research.
19. Aba-Sah Dadzie, "Approaches to Visualising Linked Data: A Survey", Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom.
20. Soren Auer, "Managing the Life-Cycle of Linked Data with the LOD2 Stack", LOD2 Project ?, c/o Universitat Leipzig, Postfach 100920, 04009 Leipzig, Germany.
21. Mathieu d'Aquin, "Assessing the Educational Linked Data Landscape", Paris, France, ACM 978-1-4503-1889-1. Web Sci'13, May 1 – May 5, 2013.