Spatial and Spatiotemporal Clustering Algorithms in Data Mining: a detailed Review

Mustafa DANACI A [1], Maliki MOUSTAPH [2]

1. Bilgisayar Mühendisliği Bölümü
Erciyes Üniversitesi
danaci@erciyes.edu.tr

2. Bilgisayar Mühendisliği Bölümü
Erciyes Üniversitesi
4010940020@erciys.edu.tr

Abstract

This study aims to present a review of past and recent studies in the field of special and spatiotemporal data clustering algorithms used in the field of spatial and space-temporal data mining. this review has been achieved through a deep detailed analysis of many articles. the methodology used was to study each selected article in detail, its method, implementation, and result and then proposed various ideas for better improvement related to each work.

Keywords: spatial, spatiotemporal data, clustering, data mining

1. Introduction

Clustering is one of the data mining techniques that can be used for grouping data without supervision to obtain hidden information from a spatiotemporal database [2]. The review method for this study will be to explore available article in different aspect such as the causes and the importance of the problem solved in each study, the hardness of the problem study, the solution proposed for the problem, how the proposed solution has been evaluated and how could be our approach to improve the proposed solution. our review will be to explore these aspects for spatial data clustering followed by the studies achieved for spatiotemporal data.

Spatial data consists of data that has a spatial component. Spatial objects can be made up of points, lines, regions, rectangles, surfaces, volumes, and even data of higher dimensions which includes time [1]. And the spatial component is implemented with a specific location attribute such as address or implicitly done by partitioning the database based on location. Geographic Information systems (GIS), biomedical applications including medical imaging, agricultural science, etc. produces a large volume of spatial data. In short Spatial data means data related to space (Güting, 1994). The special data consist of geometric information and can be discrete or continue.

2. Review related to spatial data clustering

Data mining (DM) is a field of Computer engineering also known as knowledge discovery (KN) that analyses and explores databases to extract useful information for classification. This study is also applied to spatial and spatiotemporal databases. Spatial data mining uses the relation of the special neighborhood such as topological, distance, and direction relations which are defined by the explicit location and extension of special object implicitly. Thus, special characterization, classification, and special anomaly discovery and special trend analysis require mining algorithms for special data mining. This section will discuss the spatial clustering algorithms and the second section will extend to the review of spatiotemporal clustering algorithms.

Considering available challenge tasks in discovering clusters in a spatial database especially when shape, size, and density of clusters vary a lot and the existing approach of algorithms requires sensitive parameters and well-separated clusters for successful mining. To try solving issues faced by the traditional clustering algorithm their research was based on the density-based clustering as according to researchers, this is the most promising algorithm.

To propose an alternative optimized solution for this problem, Xinging Zhou et al [2] have proposed DENSS, A multi-density clustering algorithm based on similarity for a dataset with density variation. The importance of their study was to solve existing clustering problems which is still faced in the KN field as when applying object categorization, the differences in sizes, densities, or even in shape make the process difficult for clustering algorithms and affect clustering results.

The DENSS performs clustering based on the similarity of neighbor distribution and the number of shared neighbors for two objects. Synthetic and real-world datasets were used for testing the effectiveness of the algorithm, and it has been compared with seven clustering algorithms. The experiment's result revealed that DENSS is a rapid, effective, and self-adaptive multi-density clustering algorithm that can remarkably handle datasets of any shape, density, and scale. Furthermore, DENSS is quite robust against noise. For better evaluation and better improvement of the DENSS, we could suggest experimenting with it on the dataset used to evaluate other algorithms.

Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in DM and machine learning (ML). An existing approach like CLARANS for large spatial clustering databases required previous domain knowledge for determining input parameters, arbitrary shape, and good efficiency on large databases is also required for clustering. In addition to these, those algorithms do not provide solutions for a combination of the requirement. Thus, to come up with solutions to these issues, DBSCAN was developed by Ester M. et al [3].

This algorithm was proposed to achieve clustering with few parameters. DBSCAN algorithm uses basically two parameters (eps and minPoints). DBSCAN groups together point that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions. The problem of this algorithm is that it cannot handle the density variation in data. For this reason, we this algorithm could be improved by considering additional parameters such as cluster density parameters for better

classification and this density parameter could be also related to the size and structure of the dataset.

Spatial clustering of seismic events in mines has been a challenging problem in research and has been widely reported in the literature (Marty Hudyma 2008). Clustering data with variety in density is a problem. DBSCAN has been a good choice of algorithms used to clustering dataset with equal density and with good performance in canceling noises but has problems in clustering data with unequal density. In DBSCAN, after 95% of clustering the remaining clusters are considers noise. Taking this issue in consideration, Mohammad M et al. (2018) [4] proposed an adaptative algorithm capable of identifying clusters with varying in density. The proposed adaptative algorithm is a modification of original DBSCAN so that it can adapt the values of Eps and MinPts based on the density distribution of the clusters. To test the performance of their new adaptative method, the state-of-the-art synthetic datasets were used for experiments. As a result, the ADBSACAN outperforms conventional like DBSCAN and can clusters data with varying densities but with limitations.

Adaptative DBSCAN has been experimented in dataset with only tree clusters of varying density and with random values of Eps, ε, and Minpts which makes it less generalized. We could suggest expanding the experimentation to a larger dataset with more variety and optimize the adaptative algorithm with more accuracy and try to automatize parameter variations based on the input data.

Most of the special data clustering algorithms perform well with good results. However, their performance often depends heavily on user-specified parameters [5] but this may be a problem in the practical tasks of data clustering example for image segmentation. Other related algorithms proposed in this field are k-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), EM (Expectation Maximization), and NCuts (Normalized Cuts), AP (Affinity Propagation) algorithm to identify cluster centers and members by means of passing affinity messages among data, etc.

To remove the dependence of clustering results on user-specified parameters, Jian H. et al. [5] investigated the characteristics of existing clustering algorithms and present a parameter-free algorithm based on the DSets (dominant sets) and DBSCAN. The approach was based on histogram equalization transformation to the pairwise similarity matrix of input data and make DSets clustering results independent of user-specified parameters. Then, they extend the clusters from DSets with DBSCAN, where the input parameters are determined based on the cluster's extension from DSets automatically. By merging the merits of DSets and DBSCAN, their algorithm can generate the clusters of arbitrary shapes without any parameter input. For testing the Dsets-DBSCAN, it was compared with Dests and DBSCAN algorithms separately and respectively and with other state-of-the-art algorithms in data clustering and image segmentation experiments. Their results showed that the proposed Dsets-DBSCAN approach is an effective parameters-free algorithm that performs better than other parameter turning algorithms when applied to both data clustering and image segmentation application. However, it has been noticed that in natural image segmentation there do not exist a unique, correct result as the segments are relatively small and the number of segments is large because only single partition is created.

We could suggest that to augment the performance and accurate result of this algorithm, other techniques should be integrated into it to generate multiple partitions during the process of image segmentation.

With the aim of improving DBSCAN for variably densities clustering, Safaa O and Esraa S(2019)[5] proposed another study as an enhanced DBSCAN using data partitioning technique. Their study technique was divided into two steps, the first step is the separation (partitioning) technique that separate data into sparse and dense regions and the other is the sampling technique that produces data with only one density distribution. Within the proposed algorithm process, one of the important steps is the data separation step where the data is separated into two groups based on the spatial proximity relationship using the k-nearest neighbors' method and not the Eps-neighborhood method.

As discussed by Liu et al (Liu et al., 2012), Distance relations such as k-nearest neighbors and Eps-neighborhood are two types of method of relations to find spatial proximity relationships among objects data. To experiment and evaluate their approach two-dimensional synthetic datasets are used as all the synthetic datasets have the characteristics of difference in shape, in size, and the difference in density that was the focus of their study. The results of the experiments realized on synthetic data proved that the proposed algorithm enhances the performance of DBSCAN on used data with different densities with good accuracy. However, this algorithm should be implemented in other real-world datasets for better comparison and generalization of the approach.

As DBSCAN performance is very sensitive to the parameters set, WENHAO L. et al. have proposed a new method to optimize the DBSCAN parameter with their study. They presented a multi-verse optimizer algorithm that can quickly find out the highest clustering accuracy of DBSCAN and find the interval of Eps corresponding to the highest accuracy. They design a new mechanism for the variable update of MVO to reach the range of Eps efficiently and quickly. For the experiment of the study, three artificial datasets were selected with differences in number and category of samples they contained. The inner-class distance and shape of the three clusters were also made different to make the verification of the performance of the improved MVO algorithm more reasonable. To verify the performance of the MVO, they also used the GA, PSO, MFO, and SOS to optimize the DBSCAN parameters. During the experiment, each optimization model searches the DBSCAN parameters multiple times independently and selects the optimal results.

The additional future work we would like to propose for this study is that it would be great if the parameter range search of DBSCAN algorithm can be related to nature and type of data using real-world datasets because artificial datasets were used to experiment with the approach in this study. Probably this might improve more the accuracy of the proposed optimization technique.

3. Review related to spatiotemporal data clustering

As mentioned in [8] Event clustering involves the discovery of groups that are close to each other with respect to space and time, and possibly share other non-spatial attributes (Kisilevieh et al.2010). In our previous section, we reviewed studies that mostly focused on clustering spatial data but, in this section, our review will discuss spatiotemporal studies.

We could realize from the last reviews that many types of research used the DBSCAN for solving spatial data-related problems. As the well-known DBSCAN (one of the notable density-based algorithm) does not consider the information contained within the object to define density during clustering. That could be the reason why Mohd et al [9] have proposed an approach where they use an extended DBSCAN with a density-based algorithm for spatiotemporal data as this approach provide a good result when applied to spatial clusters. The problem related to their study is how clustering could transform spatiotemporal data into useful information by considering other features such as object attributes. The aim of their study was to propose another method of clustering spatiotemporal data object by considering object attributes.

They conduct their study by defining the attribute-based function and modifying the definition of core objects for clustering. To define density in the dataset, a generalized concept using the attribute aspect of the relative importance of an object has been proposed and used. The algorithm produces multiple number of clusters as it takes into consideration the object's attributes and other precious intrinsic information contained within the object during clustering, therefore able to identify hidden and useful semantic-based clusters. their experimental results showed that their approach has the capability to identify intrinsic information based hidden clusters and does not present other relative problems encoded by the extended algorithms. To evaluate or improve the performance of the study we could suggest experimenting with the algorithm in other types of collected spatiotemporal databases for better generalization of the approach performance.

As related in [8], the ST-GRID spatiotemporal algorithm was proposed by Wang et al. (2006) as an approach that employs the partitioning of the spatial and temporal dimensions into a multi-dimensional grid with different precision, allocation of patterns into grid cells, and finally, extraction as well as merging of dense spatiotemporal regions into clusters. The study concept was about Analysis of sequences of seismic events problem. This work proposed an alternative algorithm capable to compensate lacks available in ST-GRID as it was a clustering method for the data object in a grid but not for point to point which did not improve its accuracy. So, to provide enhancement for both in time complexity and accuracy, The AGRID has been also proposed to treats data objects both in a grid and the smallest element in the clustering result, so it will enhance the time complexity and at the same time preserve the accuracy by Zhao et al. (2011).

The ST-GRID algorithm was based on a neighborhood searching strategy and relies on a sorted k-dist graph for the determination of the input parameter. But based on this technique, Fitrianah D et al [10] presented the ST-AGRID approach, based on AGRID+ algorithm with seven-step such as partitioning, computing distance threshold, calculating densities, compensating densities, calculating density threshold (DT), clustering and removing noises. Their research aimed to propose a grid density clustering algorithm for spatiotemporal data that is based on the adaptation of the grid density clustering algorithm. For experiment purpose of their study, the ST-AGRID is applied to the fishery data and the result of the clustering was processed with thresholding to discover the spatial-temporal distributions of the potential fishing zones based on the seven steps. To evaluate their approach, the execution time of their algorithm and the previous ones were compared and showed that the time execution for ST-AGRID was better than AGRID+, 8.41 second for daily temporal aggregate, 13.93 seconds for weekly temporal aggregate and 38.87 seconds for monthly temporal aggregate [10]. We would apply the ST-AGRID algorithm to the same database used in for ST-GRID and AGRID+ algorithms to demonstrate its robustness over the previous related studies.

Polygons serve an important role in the analysis of geo-referenced data as they provide a natural representation for certain types of objects, such as city blocks, city neighborhoods, and pollution hotspots. Polygons can serve as models for clusters as well and can be used by a broad set of applications, such as air pollution prevention, healthcare study, and urban planning [11]. To solve polygon-based clustering problem, Sujing W et al [11] proposed a new spatiotemporal clustering algorithm named ST-SNN (Spatiotemporal Shared Nearest Neighbor clustering algorithm) and ST-SEP-SNN (Spatiotemporal Separated Shared Nearest Neighbor clustering algorithm) which are both based on generic density-based clustering algorithm Shared Nearest Neighbor (SNN).

The proposed study aims to cluster overlapping polygons that can change their locations, size, and shapes over time. The SNN is a well establish algorithm which can find clusters of different size, shapes, and densities in high dimensional. It defines the similarity SNN defines the similarity between pairs of points in terms of how many nearest neighbors the two points share. Then they extended the SSN to cluster spatiotemporal polygons by redefining the spatiotemporal similarity between polygons considering both spatial and temporal similarities. The experimental results show that both ST-SNN and ST-SEP-SNN algorithms can find interesting spatiotemporal patterns from ozone pollution data after being evaluated in the case study involving ozone pollution events in the Houston-Galveston-Brazoria.

Khaing P. [12] proposed a distance measure to compute the spatial similarity between trajectories based on both geographical features and semantic features of motion approach. It has been realized the urgent need for specialized systems, techniques, and algorithms to analyze big spatiotemporal data generated sensors and GPS technologies. Object trajectories analysis is a very important trend in data mining filed as clustering trajectory data is a way to mine hidden information behind moving object sampling data, such as understanding trends in movement patterns and has gained high popularity in geographic information and other applications. this study tries to contribute to approaches for finding clusters moving objects with share the same moving pattern and infer the future locations of a moving object from its similar trajectories.

As the location of moving objects changes and the speed and semantic features vary, it is difficult to define the distance between them, and clustering their trajectory is a big challenge. The existing approaches first filter the trajectories by semantic similarity and then detect the geographic similarity, therefore two trajectories that are far from each other might have a very high similarity score if their semantic similarity is high. To present an approach for this issue, this research has proposed a new trajectory distance measurement is proposed to measure the distance score between trajectories from various trajectories data by considering both geographic and semantic features of moving objects. Instead of flat clustering, which is not suitable for trajectory mining, hierarchical learning has been adopted to experiment with the study with new distance function as it can support

unsupervised learning. The also proposed a designed clustering framework for moving objects trajectories to find out the groups of similar paths from big spatiotemporal data. The clustering quality of the proposed method is validated by means of external and internal validation criteria and is practically evaluated by TDrive datasets which are real trajectory dataset.

Most of the trajectory clustering related research mainly focused on spatial position changing of moving objects and a density-based spatiotemporal trajectory clustering algorithm has been proposed by Zhiyuan et al [13]. The problem related to this study is about trajectory clustering. The metric of space and time distance between points and points is relatively simple. Some trajectory pattern mining methods are to express trajectories as a series of points and use time intervals between points to measure the change of temporal dimension. But spatiotemporal analysis from the perspective of linear trajectory and pay attention to the change of object movement process is still a need.

Zhiyuan et al went from the hypothesis that if the right way to measure the space and time distance between linear trajectories is known, then a similar idea could be used to solve the spatiotemporal clustering problem of trajectories. Base on this perspective in their paper the trajectory feature points are extracted first using the curve edge detection method, then the trajectory is divided into sub-track segments according to the trajectory feature points. The density-based clustering algorithm is finally applied to cluster according to the temporal and spatial similarity between sub trajectories.

As traditional vector-based clustering algorithms are inadequate in many cases for trajectory dataset and within the scope of try to finding solutions for such issue for trajectory data clustering, Gaffney and Smyth (1999) [14] develop a clustering technique on the basis of a mixture model for continuous trajectories. They represented trajectories as functional data and their algorithm was based on a principled method for probabilistic modeling of a set of trajectories as individual sequences of points generated from a finite mixture model consisting of regression components. They employ the expectations-maximization (EM) algorithm to group the Gaussian noise and generate objects from a core trajectory. For experiment purposes, they demonstrated the ability of the approach on both simulated and real data sets and were seen to outperform well than naïve k-means and Gaussian mixture models.

Alon et al. (2003) [15] in their work proposed a model-based approach for clustering time-series data in which the cluster representative is expressed by hidden Markov models (HMMs) to estimate the transitions between successive positions. The EM framework is used for parameter estimation of the model. This research problem is to cluster similar object motions and the estimation of the motion time series model. As these methods could help in discovering clusters of similar motion sequences and enable pattern discovery, anomaly detection, modeling, summarization, etc. for this said their study focused on the problem of finding groups, or clusters of similar object motions within a database of motion sequences, and estimating motion time series models based on these groups. It was mentioned in this paper the existing issue in previous approaches for HMM-based clustering where k-means formulation have been employed but with some drawback that in each iteration, each sequence can only be assigned to a single cluster, and then only those sequences that are assigned to a particular cluster are used in the re-estimation of its HMM parameters. And this led to problems when there is not particularly good separation between the underlying processes that generated the groups of time series data.

After the implementation of the HMM-based algorithm using the HMM toolbox, classification accuracy was used to measure the validity of their clustering testing results. The simulated data were also used to compare performance between the k-means and EM-based approaches for clustering with HMM. The proposed approach lacks a robust method to handle outliers.

Proposing an approach that could discover sub-trajectories will be useful in many applications for analyzing some special regions of interest.

To solve this problem, A partition-and-group framework is proposed by Lee et al. (2007) [16] to cluster trajectories. Related to this problem, algorithms like k-means, BRICH, DBSCAN OPTICS, and STRING were reported in the literature. The logic used in this study was to partition trajectory into a set of line segments, and then, groups similar line segments together into a cluster. The trajectory Clustering (TRACLUS) algorithm works in two phases: partitioning and grouping. In the partitioning phase, the trajectory is divided into a collection of line segments using the principle of minimum description length (MDL). The grouping phase employs a density-based line segment clustering approach, which is based on the DBSCAN algorithm (Lee et al. 2007). The line segment clustering algorithm employs distance function that makes use of three components: the perpendicular distance, parallel distance, and angular distance between two-line segments. To evaluate the proposed algorithm and show its effectiveness, two different data sets were used (the hurricane track data set and the animal movement dataset). Their heuristic method for parameter value selection has been shown to estimate the optimal parameter values quite accurately. To measure the clustering quality of the algorithm, a simple quality measure for a ballpark analysis was defined and the Sum of Squared Error (SSE) was used. In addition to the SSE, the noise penalty was also considered to penalize incorrectly classified noises.

The detection of human activities on daily traffic is the problem interested by Huang et al. (2019)[17] in their studies as this kind of research help understanding human mobility and activity pattern and that can provide solutions for urban issues and traffic congestions. To solve this issue, they proposed a spatiotemporal clustering-based approach to detect human activities in daily traffic congestions from geo-tagged tweets. There is a lack of methods for information extraction from social media in available studies.

Due to the limitation of traffic information extraction from social media data, their study aims to develop an approach that allows exploring the potential influence of human activities on daily traffic congestions through linking human activities derived from geotagged tweets to the daily traffic conditions. Space, time, and semantics are associated with human activity. The DBSCAN-based approach is applied to identify the space and time of activity. This approach is helpful in urban planning and policymaking, as well as traffic event detection using social media data.

For experiments purpose, a case study of geotagged tweets posted from Toronto, Canada, and road travel speed representing traffic conditions of Toronto are used. And the results of the study showed that entertainment-related activities are more likely to appear during evening peak hours, while it seems morning rush hours are less sensitive to human

activities. In addition, it indicates that the activities involved in international events have a long-term impact on urban traffic.

This approach also can be implemented in other cities using the same type of datasets. The proposed approach and findings in this work lay a foundation for urban planning, traffic event detection, and urban policymaking using low-cost social media data. For optimal improvement of the study, we could suggest applying the same case in other countries under development like Africa countries to test in daily activities of people living in these countries could also be related to traffic congestions.

The clustering of dynamic spatiotemporal process data of the brain is another geo-referenced issue. Doborjeh and Kasabov (2015) [18] proposed a clustering method for dynamic spatiotemporal brain data. The method is based on the NeuCube spiking neural network (SNN) architecture. This study aimed to propose a method to analyze STBD as this kind of method could help to trace complex patterns and understand the process of the brain that generates data.

In NeuCube model-based used in this research, the spatiotemporal relationship between STBD streams is learned and simultaneously created as clusters. The intensity of spike communication within SNN cube is considered as the spatiotemporal similarity measure in the formation of clusters. The clusters reflect the dynamic spatiotemporal process of the brain. The proposed approach has been built based on a specific scheme simultaneously such as brain process-STBD-3D NeuCub model-3D Neucube Model clustering- Analysis of STBD- Analysis of spatiotemporal brain process. The authors experimented with the method using functional magnetic resonance imagery (fMRI) as a benchmark application. Our suggestion for more improvement of the study is how can it be used to analyze different types of brain processes and classify them for further learning analysis.

Liu et al. (2018) [19] proposed the dual-constraint spatiotemporal clustering approach (DcSTCA) to mine marine clustering patterns using long-term marine remote sensing products. Because of the variation of marine anomalies, their spatiotemporal clustering patterns have become a challenge. Marine anomaly variations have multi-dimensional attributes and are spatiotemporally continuous and existing methods for their clustering are not sufficient. The approach works in three phases. A spatiotemporal grid cube is generated based on spatial connectivity and the time evolution process of marine anomaly variations. In the second phase, the spatiotemporal density and clustering cores are obtained using the space, time, and thematic attributes. In the third phase, the spatiotemporal clustering patterns are constructed as per the density connectivity of spatiotemporal neighbors and clustering cores. The results of the tests improved performance in terms of effectiveness compared to the ST-DBSCAN algorithm suggested by Birant and Kut (2007).

In the field of geo-referenced time series clustering, it still needs to find out a suitable technique for treating spatial and temporal components of data. Proposing an approach to detect incident anomalies in temporal time series data as this is useful in a variety of applications. Izakian and Pedrycz (2013) [20] propose a framework to detect amplitude and shape anomalies in time series. In this study, the anomalies in time series are divided into two categories such as amplitude anomalies and shape anomalies. The proposed approach generates a set of sub-sequences of time series using a fixed length sliding window after FCM clustering is used to reveal the structure of time series. Dissimilarity is measured using reconstruction criteria.

The original representation of the time series is used to detect amplitude anomalies, and the autocorrelation representation of time series is used to detect shape anomalies. To measure the dissimilarity of each subsequence to different cluster centers a reconstruction criterion is used and the calculated reconstruction error has been considered an anomaly score. For detecting anomalies in amplitude, the original representation of time series is considered, while for shape anomalies an autocorrelation representation of time series was used [20]. The approach lacks in estimating an anomaly score based on the nature of the data and the application. So, we would suggest experimenting this framework with other real-world data and compare newly recorded scores.

By using the extension of the micro clustering approach, Li et al. (2004) [21] presented a study to solve the identification of moving micro clusters. Find a solution to this problem is important as due to the advances in positioning technologies, the real-time information of moving objects became increasingly available and so boosted new perspective for database research. the problem of clustering moving objects was conducted to catch interesting pattern changes during the moving process and probably provide better insight into the essence of the mobile data points. The micro-clustering was employed to handle the spatial-temporal regularities of moving objects and handle large amounts of data. They proposed the concept of a moving micro cluster by extending the micro clustering approach to spatiotemporal data originally presented by (Zhang et al. 1996) where the introduced concept micro-clusters indicate a group of data that are so close to each other that they are likely to belong to one cluster.

They proposed efficient algorithms to keep moving micro clusters geographically small. In case the segments of different trajectories occur in similar time intervals, they are grouped within a given rectangle. They evaluate their approach through experimentations which have been conducted in 2D words with a size of 600x600 and by using a collection of synthetic datasets generated by their data generator. The efficient algorithm proposed here can keep the moving micro-clusters geographically small. The approach lacks the ability to discover interesting clusters of various forms. To perfectly improve this approach additional techniques should be integrated to extract additional features of clusters such as their forms.

Extracting important information such as places in single trajectory is a problem context of the study achieved by Palma et al. (2008) [22] in which they proposed a spatiotemporal clustering approach called clustering-based stops and moves of trajectories (CB-SMoT), which uses the speed of a trajectory as the criterion to assign stops. The significant places where the user spends a considerable amount of time appear as clusters of locations in the trace Although this is basically a clustering problem, the popular clustering algorithms are not quite right for this particular problem for three reasons: (1) the number of clusters is an a priori parameter; (2) the generated clusters include unimportant locations; and (3) the clustering algorithms require a significant amount of computation.

The incremental clustering approach where location measurements are clustered together was used as the main technical algorithm. After discussing the two principal ways to express locations in location-aware systems such as

coordinates-based and landmark-based systems, they describe an algorithm for extracting significant places from a trace of coordinates as in traces, significant places are the regions where many location measurement samples are clustered together. The algorithm identifies these clusters from the trace automatically. CB-SMoT first creates the cluster of the slow speed part, after which the technique matches the clusters with appropriate geographic places. CB-SMoT follows the same principle as DBSCAN; first, it looks for core points and then expands them by aggregating other points in the neighborhood.

The proposed algorithm has been evaluated experimentally with real traces collected from Place Lab (Bill S.et al. 2003), a location system that uses WiFi access points' beacon messages to determine a user's location. For the effectiveness of the system, some criteria for the evaluation were also decided such as how the places are well identified and the results are measured in terms of accuracy, erroneousness of the identification of the places, and completeness. The initial experimental results of the algorithm showed that their algorithm extracts the most significant places successfully. The approach addresses only spatial and speed-based semantics and lacks in handling other types of semantics in trajectories, such as acceleration. For this reason, we could suggest that other spatiotemporal techniques should be integrated for better performance and more automation and future places prediction for the user.

4. Conclusion

This study presents a detailed and comprehensive review of researches conducted in the application and development of clustering algorithms to solve problems related to spatial and spatiotemporal data. The first section discusses the clustering algorithm for spatial data while the second section exposes works on spatiotemporal data. This study is conducted in detail to understand the context problem of each paper, its importance, the proposed approach of each, and its related work. Each paper has been evaluated and we have also proposed some tips and suggestions for their improvement in case it available.

5. References

[1] Bindiya M. Unnikrishman A. Poulouse Jacob "Spactial Clustering Algorithms – An Overview" *Asian Journal of Computer Science and Information Technology (AJCSIT)* 2013.

[2] Xinging Zhou et al. "A Multi-density Clustering Algorithm Based on Similarity for Dataset with Density Variation", *IEEE*, 2019.

[3] M. Ester, H.P. Kriegel, and X. Xu, "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise" *International Conference on Knowledge Discovery & Data Mining*. 1996, pp. 226-231

[4] Mohammad M. et al. "ADBSCAN: Adapatative Density-Based Spatial Clustering of Applications with Noise for Identifying Clusters with Varying Densities". International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), IEEE 2018

[5] Jian H. et al. "Dset-DBSCAN: A Parameter-Free Clustering Algorithm" *IEEE* 3026.

[6] Safaa O., Esraa S. "Data Partitioning Technique to Enhance DBSCAN Clustering Algorithm". *Journal of Babilon University/Pure and Applied Science*. 2017.

[7] WENHAO L. et al. "A New DBSCAN Parameters Determination Method Based on Improved MVO" *IEEE* 2019.

[8] Mohd Y et al. "Spatiotemporal clustering: a review" *ResearchGate* 2019

[9] Mohd Y et al. "Density based Algorithm for Spatiotemporal Data" ICCIS, *IEEE* 2019.

[10] Fitrianah D. et al "ST-AGRID: A Spatio Temporal Grid Density Based Clustering and Its Application for determining the Potential Fishing Zones" *International Journal of Software Eng. and Its Applications*, 2015.

[11] Suhing W et al. "New Spatiotemporal Clustering Algorithms and their Application to Ozone Pollution" *IEEE International Conference on Data Mining Workshop* 2013.

[12] Zhiyuan C et al. "Density based spatiotemporal trajectory clustering algorithm" *IEEE IGARSS* 2018

[13] Khaing P., Than N. Aung "Distance-based Clustering of Moving Objects' *Trajectories form Spatiotemporal Big Data" IEEE, ICIS* 2018.

[14] Shiyuan et al., "Density based spatio-temporal trajectory clustering algorithm" *IGARSS IEEE* 2018.

[15] Scott Gaffney and Padhraic Smyth "Trajectory Clustering with Mixures of Regression Models" *Research Gate*, 1999.

[16] Alon J. et al., "Discovering Clustering in Motion Time-Series Data" *Computer Society Conference on Conference on Computer Vison and Pattern Recognition*, 2003.

[17] Lee J, Jiawei H. and Whang K., "Trajectory Clustering: A partition-and-Group Framework" *SIGMOD*, 2007.

[18] Huang et al., "An exploration of the interaction between urban human activities and daily traffic conditions: A case study of Toronto, Canada". 84:8-22. https://doi.org/10.1016/j.cities.2018.07.001.

[19] Deborjeh M. and Nikola K., "Dynamic 3D Clustering of Spatio-Temporal Brain Data in the NeuCube Spiking Neural Network Architecture on a Case Study of fMRI Data", 191-198. *Springer, Cham*. DOI: 10.1007/978-3-319-26561-2_23.

[20] Liu J et al., "Dual-Constraint Spatiotemporal Clustering Approach for Exploring Marine Anomaly Patterns Using Remote Sensing Products", *Journal of Selected Topics in Applied Earth Observation and Remote Sensing, IEEE* 2018.

[21] Izakian H. and Witold P. "Anolmaly Detection in Time Series Data using a Fuzzy C-means Clustering", *IEEE*, 2013.

[22] Li Y, Han J, Yang J (2004) "Clustering moving objects", *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD* '04 617–622. doi: 10.1145/1014052.1014129

[23] Palma AT, Bogorny V, Kuijpers B, Alvares LO (2008). "A clustering based approach for discovering interesting places in trajectories," in *ACMSAC. New York, NY, USA: ACM Press*, 863–868.