Journal of critical reviews 54 Journal of Critical Reviews FOCUSED INFORMATION CRITERION BASED PARTITIONED ITERATIVE X-MEANS DICE CORRELATION CLUSTERING FOR BIG GEO-SOCIAL DATA

Article *in* Journal of Critical Reviews · May 2021

DOI: 10.31838/jcr.07.06.13

2 authors:

M. Anoop
Alpha Arts and Science College

Sripriya P.
Vels University

Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 6, 2020

Review Article

FOCUSED INFORMATION CRITERION BASED PARTITIONED ITERATIVE X-MEANS DICE CORRELATION CLUSTERING FOR BIG GEO-SOCIAL DATA

M. Anoop[1], P. Sripriya[2]

[1]Research Scholar, Department of Computer Applications, Vels Institute of Science, Technology & Advanced Studies, Chennai, Tamil Nadu, India. profanoopcs@rediffmail.com
[2]Professor, Department of Computer Applications, Vels Institute of Science, Technology & Advanced Studies, Chennai, Tamil Nadu, India. sripriya.phd@gmail.com

Received: 13.02.2020 *Revised: 15.03.2020* *Accepted: 17.04.2020*

Abstract
Geo-social data is location-based social media data which is generated by people on social network (i.e. face book, twitter etc.,) that is related to specific locations. There are lot of social users who are generates very large amount of data called "Big Data" that is difficult to be analyzed and make real-time decisions. Few research works have been designed for clustering geo-social data using different techniques. However, clustering performance of conventional algorithms was not higher to exactly find frequently visited location of users in social network when taking big geo-social dataset as input. In order to overcome such drawbacks, a Focused Information Criterion based Partitioned Iterative X-means Dice Correlation Data Clustering (FIC-PIXDCDC) Method is proposed in this work. The FIC-PIXDCDC Method groups the similar geo-social data with higher accuracy and lesser time. In FIC-PIXDCDC method, geo-social data (i.e., user, location and time) from Weeplaces dataset is initially taken as an input. After obtaining input, FIC-PIXDCDC method chooses number of clusters and centroids randomly. Then, FIC-PIXDCDC calculates dice correlation between each input geo-social data and cluster centroids. Subsequently, FIC-PIXDCDC method applies Focused Information Criterion to construct optimal number of clusters for a given big dataset. This process of FIC-PIXDCDC method is repetitive until no deviation in cluster centroids. Accordingly, FIC-PIXDCDC method group's interrelated geo-social data together with higher accuracy and lower time to precisely discover location information of frequently visited users in social network. Experimental evaluation of FIC-PIXDCDC method is carried out on factors such as clustering time, clustering accuracy, error rate with respect to number of geo-social data.

Keywords: Cluster Centroid, Dice Correlation, Focused Information Criterion, Fréchet Mean, Frequent Visited Users, Geo-Social Data.

DOI: http://dx.doi.org/10.31838/jcr.07.06.13

INTRODUCTION

Clustering is an essential area in data mining that partition the data into groups where the points in same cluster are similar while the points in different clusters are dissimilar. Clustering identifies the applications in pattern identification, image analysis, information retrieval and bioinformatics. The huge development of Geo-Social Networks (GeoSNs) brings interesting data to perform the clustering process. In GeoSNs such as Gowalla, Foursquare, and Facebook, users gather the geographic locations and distribute them through operation termed checkin. A checkin is a triplet (user, position, time) modeling that user visited the place with point location at specified time.

The users of social networks are linked with their checkin point locations. Geo-social clustering is a straightforward when the set of communities is identified. A community is the group of users with similar interests in visiting the places. When the user visiting geo-social cluster increases chance of user visiting, then they are part of same community. Many researches were carried out their research on geo-social network. But, the clustering accuracy was not improved which increases the false positive rate of finding most user visited place with point location at specified time. To resolve the above said conventional issues, FIC-PIXDCDC method is proposed in this work. The objective of FIC-PIXDCDC method is to increases the clustering accuracy and reduces the clustering time of big geo-social data analytics.

Density-based spatial clustering of applications with noise (DBSCAN) algorithm was introduced in [1] for consumer clusters discovery with geo-tagged social network information. However, clustering performance of geo-social data was not efficient. Density-based Clustering Places in Geo-Social Networks (DCPGS) [2] was designed to find the social connections between users. However, DCPGS was not effective and the temporal dimension failed to get better quality of clusters.

A novel algorithm was introduced in [3] to analyze the data streams with interrelated components from clusters with varied covariance matrices. However, the clustering time was not reduced by using designed clustering algorithm. A powerful clustering method termed MUFOLD-CL was introduced in [4]. Though clustering accuracy was improved, computational cost was not minimized

An in-memory computing design on heterogeneous CPU-GPU clusters called GFlink was introduced in [5] for large data. But, the error rate was not reduced using GFlink. The computational overhead was not minimized. Subject-Verb-Object Semantic Suffix Tree Clustering (SVOSSTC) was presented in [6] to reduce the time needed for grouping twitter data with higher accuracy. However, the ratio of number of twitter data that are exactly clustered was not enough.

A survey of different techniques designed for big data analytics of geo-social media was analyzed in [7]. A large-scale location-based social network was analyzed in [8] to find the impact of human geo-social interaction patterns with lower false positive rate. But, computational complexity of this algorithm was very higher. Spatio-temporal context-aware event representation was introduced in [9] to discover connections and related patterns among countries. However, time and space complexity were remained open issue.

Advanced computing model was presented in [10] to attain higher throughput by examining huge amount of geo-social network information. But, finding location of most visited users in social network was not accurate. Relative study on analyses and inference of geo-social media to find real-time decisions in big-data was introduced in [11].

To addresses the above said existing issues, FIC-PIXDCDC method is proposed in this research work. The key contributions of proposed FIC-PIXDCDC method is explained in below,

- To get enhanced clustering performance for geo-social data when compared to state-of-the-art works, FIC-PIXDCDC method is introduced by using Focused Information Criterion and Dice Correlation Coefficient Measurement, Fréchet mean in Partitioned Iterative X-means Clustering algorithm. The proposed FIC-PIXDCDC method presents a fast and effective way to group unstructured data as compared to existing works using Focused Information Criterion and Fréchet mean as compared to existing works. This results in minimal error rate for efficient clustering of big geo-social data.
- To reduce the amount of time taken for clustering big geo-social data when compared to conventional algorithms, Dice Correlation Coefficient Measurement is used in proposed FIC-PIXDCDC method. On the contrary to state-of-the-art works, FIC-PIXDCDC method identifies the similarities between input geo-social data and cluster centroid depends on the locations and their semantics by Dice Correlation Coefficient Measurement. This supports for FIC-PIXDCDC method to effective big geo-social data clustering with a minimal amount of time.

The rest of paper is created as follows. In Section 2, the detailed process of FIC-PIXDCDC method is explained using an architecture diagram. Section 3 describes the experimental settings. The comparative result analysis of proposed FIC-PIXDCDC method is discussed in Section 4. Section 5 shows the literature survey. Finally, the paper concluded in section 6.

FOCUSED INFORMATION CRITERION BASED PARTITIONED ITERATIVE X-MEANS DICE CORRELATION DATA CLUSTERING METHOD

The Focused Information Criterion based Partitioned Iterative X-means Dice Correlation Data Clustering (FIC-PIXDCDC) Method is introduced with aiming at enhancing the clustering performance of big geo-social data. On the contrary to traditional works, FIC-PIXDCDC method is proposed by combining the Focused Information Criterion and Dice Correlation Coefficient Measurement in Partitioned Iterative X-means Clustering algorithm. The FIC-PIXDCDC is designed by using concepts of k-means clustering. The FIC-PIXDCDC method is developed used for clustering analysis of big geo-social data in which similar location data is grouped based on Focused Information Criterion on the contrary to state-of-the-art works.

The designed FIC-PIXDCDC method partitioned the collection of input big geo-social data in a given dataset into number of clusters 'x' according to Focused Information Criterion. In proposed FIC-PIXDCDC, Focused Information Criterion is utilized to determine which groups a certain object (i.e. input big geo-social data) really belongs to with a minimal amount of time complexity. On the contrary to conventional clustering algorithms, FIC-PIXDCDC method used Dice Correlation Coefficient Measurement and Focused Information Criterion in order to accurately cluster geo-social data in input big data. The architecture diagram of FIC-PIXDCDC method is presented in below Figure 1.

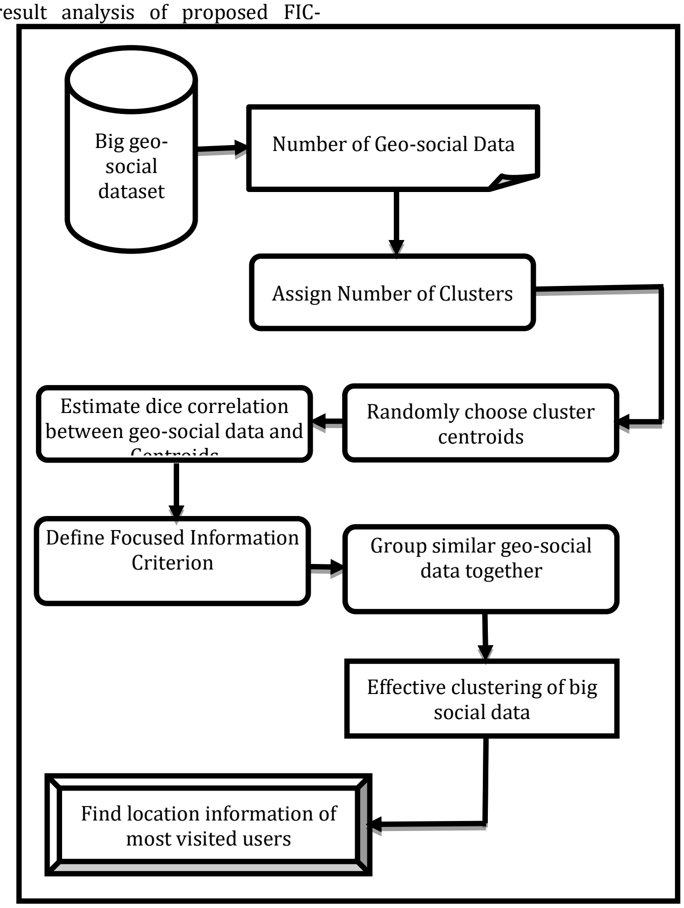

Figure 1: Architecture Diagram of FIC-PIXDCDC method for Clustering Big Geo-Social Data

Figure 1showsblock diagram of FIC-PIXDCDC method to efficiently carry out big geo-social data clustering process. As demonstrated in the above figure, FIC-PIXDCDC method initially gets number of geo-social data 'd_i' in given big dataset as input. Followed by, FIC-PIXDCDC method selects number of clusters and consequently defines number of cluster centroids arbitrarily. Next, FIC-PIXDCDC estimates dice correlation **(identify similarities)** between each input geo-social data and cluster centroids. Then, FIC-PIXDCDC method applies Focused Information Criterion **(does not assess the overall fit of candidate models but focuses attention directly on the parameter of primary interest with the statistical analysis,)** to form optimal number of clusters for a given big dataset. The above process of FIC-PIXDCDC method is continual until no variation in cluster centroids. From that, FIC-PIXDCDC method group's similar types of geo-social data in input dataset with a minimal amount of time consumption by using focused information criterion. By grouping of similar geo-social network data, FIC-PIXDCDC method significantly identifies location information of most visited users by geo-social network as compared to conventional works.

Let us consider input big geo-social dataset is represented as '$DS = d_1, d_2, \ldots, d_\varepsilon$' where '$\varepsilon$' denotes the total number of geo-social data. After taking input, Focused Information Criterion based Partitioned Iterative X-means Dice Correlation Data Clustering is carried out in this work. On the contrary to existing works, FIC-PIXDCDC method is designed because it consistently gives better clustering accuracy for both synthetic and real life dataset. In addition to that, FIC-PIXDCDC method also run very faster to find frequently visited location of users in social network. The flow processes of FIC-PIXDCDC method is depicted in below Figure 2.

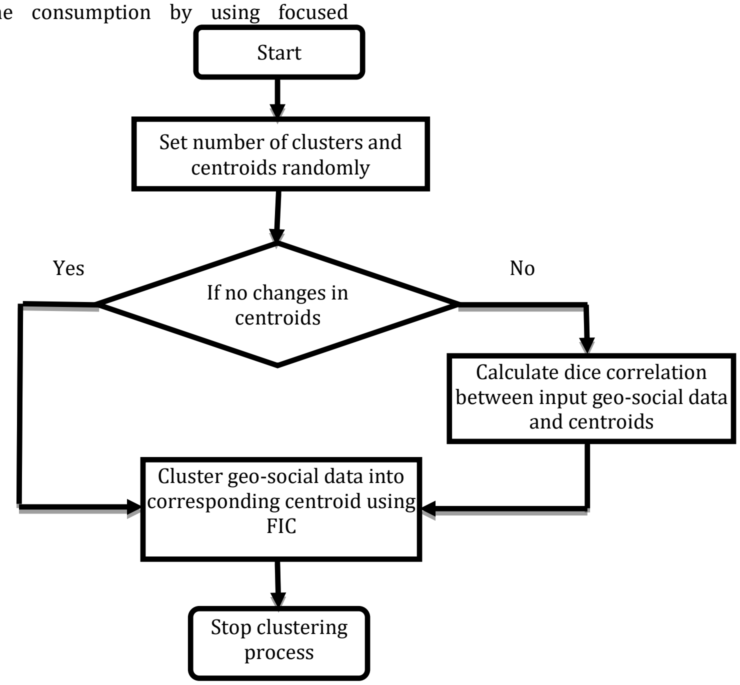

Figure 2: Flow Processes of FIC-PIXDCDC Method

Figure 2 presents flowchart of FIC-PIXDCDC method to increases the clustering accuracy of big geo-social data analytics for efficient prediction of frequently visited regions or areas of users. As illustrated in the above diagram, FIC-PIXDCDC process starts with random initialization of number of clusters 'x' and centroids 'τ'. After that, FIC-PIXDCDC method computes the similarity between each geo-social data and centroids. The conventional K-means clustering employed Euclidean distance to find out the distance between data and cluster centroids. By using distance calculation, the conventional K-means clustering does not give higher clustering accuracy to exactly determine frequently visited regions of users by social network. Therefore, a novel clustering algorithm called FIC-PIXDCDC method is introduced in this work with application dice correlation coefficient measurement to achieve better accuracy for grouping social network data. On the contrary to state-of-the-art clustering algorithms, FIC-PIXDCDC method applies dice correlation coefficient measurement which identifies the similarities between all input geo-social data 'd_i' and cluster centroid 'τ_j' using below,

$$\$\ (\boldsymbol{d_i}, \boldsymbol{\tau_j}) = \frac{\boldsymbol{d_i \cap \tau_j}}{|\boldsymbol{d_i}| * |\boldsymbol{\tau_j}|} \quad (1)$$

From the above mathematical representation (1), 'intersection symbols '$\cap$' denotes a mutual dependence between geo-social data 'd_i' and cluster centroid 'τ_j' whereas '$|d_i|$'and '$|\tau_j|$' refers the cardinalities between the geo-social data 'd_i' and cluster centroid 'τ_j'. The result of dice correlation coefficient '$\$\ (\boldsymbol{d_i}, \boldsymbol{\tau_j})$' value is always ranges between the '0' and '1'.

In FIC-PIXDCDC method, dice correlation coefficient is determined based on the locations and their semantics. **If the measured correlation value between the geo-social data is '1', places of two social network users are similar. When the correlation value between the geo-social data is '0', places of social network users are dissimilar.** By using this dice correlation coefficient value, FIC-PIXDCDC method significantly groups frequently visited regions or areas of users in social network with the application of Focused Information Criterion. On the contrary to the traditional clustering techniques, FIC-PIXDCDC method used focused information criterion in order to improve big geo-social data clustering performance.

In proposed FIC-PIXDCDC method, the focused information criterion (FIC) selects most appropriate geo-social data among a set of geo-social data for an each cluster centroids. **On the contrary to selection method such as the Akaike information**

criterion, Bayesian information criterion and the deviance information criterion, the proposed focused information criterion does not try to assess the overall fit of candidate models but focuses attention directly on the data of primary interest with the statistical analysis. This helps for FIC-PIXDCDC method for effective clustering of geo-social data. The focused information criterion utilized in FIC-PIXDCDC method is a condition for choosing geo-social data among collections of geo-social data in a given big dataset during the cluster formation process. The focused information criterion considers the dice correlation between each input geo-social data and centorids in order to precisely group related data together with a minimal time complexity. Thus, the goal of focused information criterion is to cluster the geo-social data which has maximum dice correlation value to particular centroid which is mathematically performed using below,

$$X\_Means\ Cluster = arg\max_{c} \sum_{j=1}^{x} \sum_{d_i \in c_i} \$\ (d_i, \tau_j) \quad (2)$$

From the mathematical formula (2), 'c_i' signifies the set of geo-social data that belong to cluster 'j'. With help of the above equation (2), FIC-PIXDCDC method groups the geo-social data to the cluster whose dice correlation value from the cluster centroid is higher of all the cluster centroid by using focused information criterion. Subsequently, cluster centroid is updated by considering the weighted average dice correlation value of geo-social data in that cluster.

On the contrary to conventional clustering, to accurately determine the centroid of each cluster during the every iteration, Fréchet mean is employed in FIC-PIXDCDC method which is a generalization of centroids to metric spaces which gives central tendency for a cluster of points. Let us consider '$d_1, d_2, .., d_n$' be number of geo-social data within cluster 'c_i'. For a data in cluster, new cluster centroid 'cd_i^*' is measured as weighted average dice correlation value of geo-social data in that cluster. From that, re-estimation of the new cluster centroid 'cd_i^*' is mathematically obtained as follows,

$$cd_i^* = \frac{\sum_{d_i=1}^{n} \$\ (d_i, \tau_j)}{n} \quad (3)$$

From the above mathematical equation (3), a_i represents the number of geo-social data in i^{th} cluster. This re-determination of cluster centroids in FIC-PIXDCDC method gives higher clustering results for efficient analytics of big geo-social data. This process of FIC-PIXDCDC method is recurrent until there is no variation in cluster centroids. From that, FIC-PIXDCDC method efficiently groups each geo-social data into a related cluster with enhanced accuracy and minimal amount of time.

The algorithmic process of FIC-PIXDCDC Method is shown in below,

// Focused Information Criterion based Partitioned Iterative X-means Dice Correlation Data Clustering Algorithm
Input: Number of geo-social data '$DS = d_1, d_2, .., d_\varepsilon$'
Output: Achieve higher clustering accuracy for big geo-social dataset
Step 1: Begin
Step 2: Consider 'x' number of clusters
Step 3: Randomly select number of cluster centroids 'cd_i'
Step 4: While (no change in cluster centroids 'cd_i') **do**
Step 5: For each geo social data 'd_i'
Step 7: Compute dice correlation between 'd_i'and 'cd_i' using (1)
Step 8: Define focused information criterion
Step 9: Group geo social data to corresponding cluster 'd_i' using (2)
Step 10: Re-determine cluster centroid 'cd_i^*' using (3)
Step 11: End For
Step 12: End while
Step 13: Identify frequently visited location information users in social network
Step 14: End

Algorithm 1: Focused Information Criterion based Partitioned Iterative X-means Dice Correlation Data Clustering

Algorithm 1 demonstrates the step by step process of FIC-PIXDCDC Method to attain better clustering performance for grouping related geo-social data together. As shown in above algorithmic process, at the beginning FIC-PIXDCDC Method assumes 'x' number of clusters and centroids randomly. Then, FIC-PIXDCDC Method evaluates the dice correlation between each input geo-social data and cluster centroids. Followed by, FIC-PIXDCDC Method clusters the similar geo-social together into corresponding clusters using focused information criterion and recalculating cluster centroid. The above process of FIC-PIXDCDC Method is repeated until there is no change in cluster centroids. Through an effective clustering of geo-social data, finally FIC-PIXDCDC Method accurately finds location information of frequently visited regions or areas of users in social network as compared to conventional works.

EXPERIMENTAL SETTINGS

To evaluate the performance, proposed FIC-PIXDCDC Method and conventional Density-based spatial clustering of applications with noise (DBSCAN) algorithm [1] and Density-based Clustering Places in Geo-Social Networks (DCPGS) [2] are implemented in Java Language using Weeplaces Dataset [22]. This dataset was obtained from popular location-based social network services e.g., Facebook Places, Foursquare, and Gowalla. Besides, this dataset comprises of 7,658,368 check-ins made by 15,799 users over 971,309 locations. In Weeplaces Dataset contains check-in history, their friends who also use Weeplaces, and other additional information regarding the locations. Here, check-in information considered as geo-social data which includes user, check-in-time, latitude, and longitude and location id.

From that, FIC-PIXDCDC Method takes 1000 to 10000 geo-spatial data from Weeplaces Dataset to conduct experimental process. The performance of proposed FIC-PIXDCDC Method is measured in terms of clustering accuracy, clustering time and error rate with respect to various number of geo-social data. The effectiveness of FIC-PIXDCDC Method is compared against conventional Density-based spatial clustering of applications with noise (DBSCAN) algorithm [1] and Density-based Clustering Places in Geo-Social Networks (DCPGS) [2].

RESULTS

In this section, the experimental result of proposed FIC-PIXDCDC Method is compared with two existing Density-based spatial clustering of applications with noise (DBSCAN) algorithm [1] and Density-based Clustering Places in Geo-Social Networks (DCPGS) [2] is presented. The efficiency of proposed FIC-PIXDCDC Method is analyzed using metrics such as clustering accuracy, clustering time and error rate with help of below table and graph.

1) Case 1: Performance Measure of Clustering Accuracy

In FIC-PIXDCDC Method, Clustering accuracy 'CA' calculates the ratio of number of geo-social data that are precisely grouped to the total number of geo-social data taken for conducting experimental process. The clustering accuracy is computed mathematically using below,

$$CA = \frac{\varepsilon_{AC}}{\varepsilon} * 100 \quad (4)$$

From the above mathematical representation (4), 'ε_{AC}' signifies number of accurately clustered geo-social data in which 'ε' point outs a total number of geo-social data. The clustering accuracy of big geo-social data is determined in terms of percentage (%).

Sample Calculation

- **Proposed FIC-PIXDCDC**: Number of geo-social data perfectly clustered is 860 and the total number of geo-social data is 1000. Then the clustering accuracy is acquired as follows,

$$CA = \frac{860}{1000} * 100 = 86\ \%$$

- **Existing DBSCAN:** Number of geo-social data properly clustered is 740 and the total number of geo-social data is 1000. Then the clustering accuracy is evaluated as follows,

$$CA = \frac{740}{1000} * 100 = 74\ \%$$

- **Existing DCPGS:** Number of geo-social data exactly clustered is 790 and the total number of geo-social data is 1000. Then the clustering accuracy is computed as follows,

$$CA = \frac{790}{1000} * 100 = 79\ \%$$

Both the proposed FIC-PIXDCDC and existing DBSCAN [1] DCPGS [2] Methods are implemented in Java Language by taking a diverse number of geo-social data in the range of 1000-10000 from input big dataset to estimate clustering accuracy. When carried outing the experimental process using 4000 geo-social data from big Weeplaces dataset, the proposed FIC-PIXDCDC method obtains 92 % clustering accuracy whereas traditional DBSCAN [1] DCPGS [2] acquires 77 % and 81 % respectively. From the above get experimental results, it is expressive that the clustering accuracy of big geo-social data using proposed FIC-PIXDCDC method is very higher when compared to other conventional works [1] and [2]. The clustering accuracy result of proposed FIC-PIXDCDC method is compared with two state-of-the-art works is demonstrated in below Table 1.

Table 1: Experimental Result of Clustering Accuracy

Number of geo-social data (ε)	Clustering Accuracy (%)		
	FIC-PIXDCDC	DBSCAN	DCPGS
1000	86	74	79
2000	88	79	83
3000	89	79	82
4000	92	77	81
5000	90	75	81
6000	89	75	80
7000	87	74	79
8000	87	72	77
9000	85	71	76
10000	84	70	75

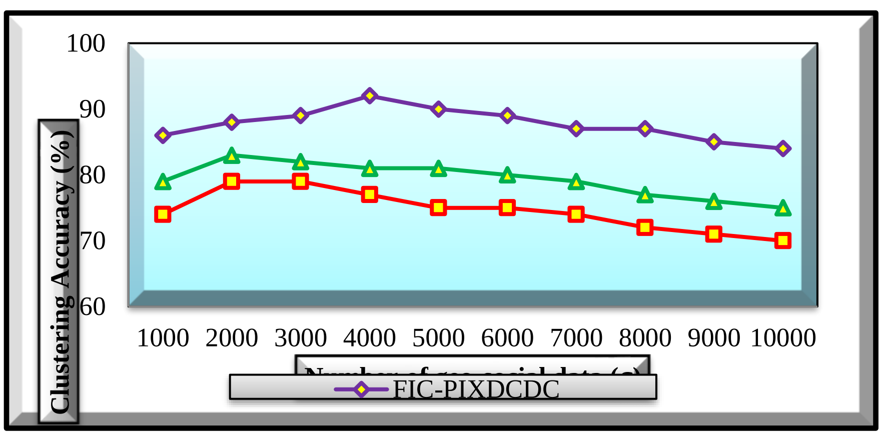

Figure 3: Comparative Result of Clustering Accuracy versus Different Number of Big Geo-Social Data

Figure 3 illustrates the impact of clustering accuracy with respect to diverse number of big geo-social data in the range of 1000 to 10000 using three methods namely proposed FIC-PIXDCDC and existing DBSCAN [1] DCPGS [2]. As presented in the above graphical representation, proposed FIC-PIXDCDC method gives higher accuracy to cluster related geo-social data together with increasing number of input geo-social data when compared to conventional DBSCAN [1] DCPGS [2]. This is owing to application of Focused Information Criterion and Dice Correlation Coefficient Measurement and Fréchet mean calculation in Partitioned Iterative X-means Clustering algorithm on the contrary to traditional works.

Proposed FIC-PIXDCDC method is a variation of k-means clustering that effectively performs cluster allocations through repeatedly attempting partition and keeping the optimal result until some condition is attained. From that, proposed FIC-PIXDCDC method increase the clustering performance of big geo-social data as compared to existing works. Hence, proposed FIC-PIXDCDC method exactly carried outs big geo-social data process. This helps for proposed FIC-PIXDCDC method to enhance the ratio of number of geo-social data that are correctly grouped when compared to other conventional works [1] and [2]. As a result, proposed FIC-PIXDCDC method achieves enhanced clustering accuracy to discover location information of frequently visited users in social network by 18 % as compared to DBSCAN [1] and 11 % when compared to DCPGS [2].

2) Case 2: Performance Measure of Clustering Time

In FIC-PIXDCDC Method, Clustering Time 'CT' determines the amount of time needed to group same type of geo-social data together. The clustering time is mathematically estimated using below formula,

$$CT = \varepsilon * t(CS) \quad (5)$$

From the above mathematical expression (5), '$t(CS)$' symbolizes a time utilized to cluster a single geo-social data and 'ε' refers to a total number of geo-social data considered for experimental evaluation. The clustering time of big geo-social data is computed in terms of milliseconds (ms).

Sample Calculation for Clustering Time

- **Proposed FIC-PIXDCDC**: the amount of time employed to cluster one geo-social data is 0.026 ms and the total number of geo-social data is 1000. Then the clustering time is mathematically calculated as follows,
$$CT = 1000 * 0.026 = 26\, ms$$
- **Existing DBSCAN:** the amount of time taken to cluster one geo-social data is 0.035 ms and the total number of geo-social data is 1000. Then the clustering time is mathematically evaluated as follows,
$$CT = 1000 * 0.035 = 35\, ms$$
- **Existing DCPGS:** the amount of time used to cluster one geo-social data is 0.037 ms and the total number of geo-social data is 1000. Then the clustering time is mathematically determined as follows,
$$CT = 1000 * 0.037 = 37\, ms$$

In order to measure time complexity of big geo-social data clustering, both the proposed FIC-PIXDCDC and traditional DBSCAN [1] DCPGS [2] Methods are implemented in Java Language by considering a various number of geo-social data in the range of 1000-10000 from input big Weeplaces dataset. When performing the experimental evaluation using 8000 geo-social data from big dataset, the proposed FIC-PIXDCDC method attains 50 ms clustering time whereas state-of-the-art works DBSCAN [1] DCPGS [2] get 59 ms and 61 ms respectively. Thus, it is clear that the clustering time of big geo-social data using proposed FIC-PIXDCDC method is very minimal as compared to other traditional works [1] and [2]. The performance result of clustering time using proposed FIC-PIXDCDC method is compared with two existing methods is depicted in below Table 2.

Table 2: Experimental Result of Clustering Time

Number of geo-social data (ε)	Clustering Time (ms)		
	FIC-PIXDCDC	DBSCAN	DCPGS
1000	26	35	37
2000	28	36	40
3000	33	42	48
4000	36	44	52
5000	39	47	48
6000	41	53	54
7000	46	56	57
8000	50	59	61
9000	54	63	65
10000	58	68	70

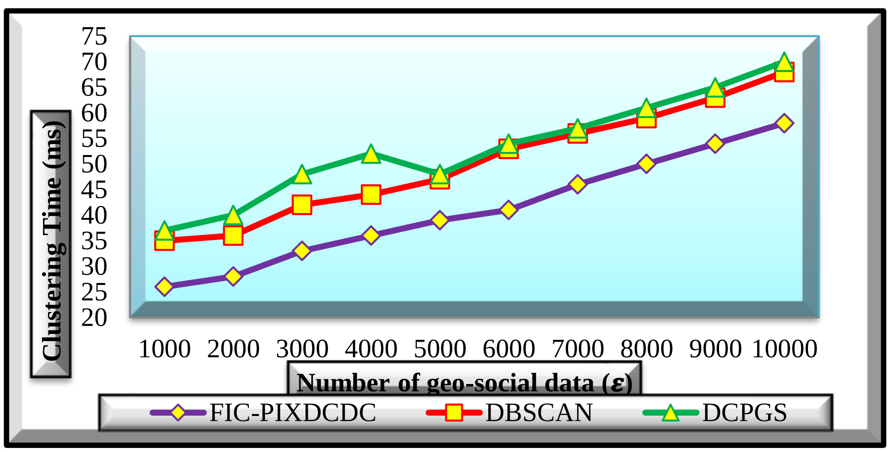

Figure 4: Comparative Result of Clustering Time versus Different Number of Big Geo-Social Data

Figure 4 demonstrates the impact of clustering time according to varied number of big geo-social data in the range of 1000 to 10000 using three methods namely proposed FIC-PIXDCDC and existing DBSCAN [1] DCPGS [2]. As shown in the above graphical depiction, proposed FIC-PIXDCDC method provides minimal amount of time to cluster related geo-social data together with increasing number of input geo-social data when compared to conventional DBSCAN [1] DCPGS [2]. This is because of application of Focused Information Criterion and Dice Correlation Coefficient Measurement and Fréchet mean calculation in Partitioned Iterative X-means Clustering algorithm on the contrary to state-of-the-art works.

By using the above concepts, proposed FIC-PIXDCDC method gives a fast and effective way to cluster unstructured data and also provides concurrency speeds up the process of model construction. In addition to that, proposed FIC-PIXDCDC method utilizes Focused Information Criterion that provides a mathematically sound measure of higher quality cluster for big geo-social data as compared to existing works. This assists for proposed FIC-PIXDCDC method to reduce the amount of time utilized to group same type of geo-social data into a different number of clusters when compared to other traditional works [1] and [2]. Hence, proposed FIC-PIXDCDC method attains minimal amount of clustering time to find out location information of most visited users in social network by 19 % as compared to DBSCAN [1] and 24 % when compared to DCPGS [2].

3) Case 3: Performance Measure of Error Rate

In FIC-PIXDCDC Method, Error Rate 'ER' computes ratio of number of geo-social data mistakenly clustered to the total number of geo-social data. The error rate is mathematically determined using below representation,

$$ER = \frac{\varepsilon_{WC}}{\varepsilon} * 100 \quad (6)$$

From the above mathematical formula (6), 'ε_{WC}' indicates a number of geo-social data wrongly clustered and 'ε' signifies a total number of geo-social data. The error rate of geo-social data is determined in terms of percentage (%).

Sample Calculation for Error Rate

- **Proposed FIC-PIXDCDC**: number of geo-social data mistakenly grouped is 140 and the total number of geo-social data is 1000. Then the error rate is mathematically obtained as follows,

$$ER = \frac{140}{1000} * 100 = 14\,\%$$

- **Existing DBSCAN:** number of geo-social data wrongly clustered is 260 and the total number of geo-social data is 1000. Then the error rate is mathematically acquired as follows,

$$ER = \frac{260}{1000} * 100 = 26\,\%$$

- **Existing DCPGS:** number of geo-social data incorrectly clustered is 210 and the total number of geo-social data is 1000. Then the error rate is mathematically measured as follows,

$$ER = \frac{210}{1000} * 100 = 21\,\%$$

For determining the error rate involved during the big geo-social data clustering process, both the proposed FIC-PIXDCDC and state-of-the-art DBSCAN [1] DCPGS [2] Methods are implemented in Java Language by using a different number of geo-social data in the range of 1000-10000 from input big dataset. When conducting the experimental work using 6000 geo-social data from big Weeplaces dataset, the proposed FIC-PIXDCDC method achieve 11 % error rate whereas existing works DBSCAN [1] DCPGS [2] attain 25 % and 20 % respectively. From the above acquired experimental results, it is descriptive that the error rate of big geo-social data clustering using proposed FIC-PIXDCDC method is very lower when compared to other conventional works [1] and [2]. The comparative result of clustering time using proposed FIC-PIXDCDC method and two traditional methods is presented in below Table 3.

Table 3: Experimental Result of Error Rate

Number of geo-social data (ε)	Error Rate (%)		
	FIC-PIXDCDC	DBSCAN	DCPGS
1000	14	26	21
2000	12	22	17
3000	11	21	18
4000	8	24	19
5000	10	25	19
6000	11	25	20
7000	13	27	21
8000	14	28	23
9000	15	29	24
10000	16	30	25

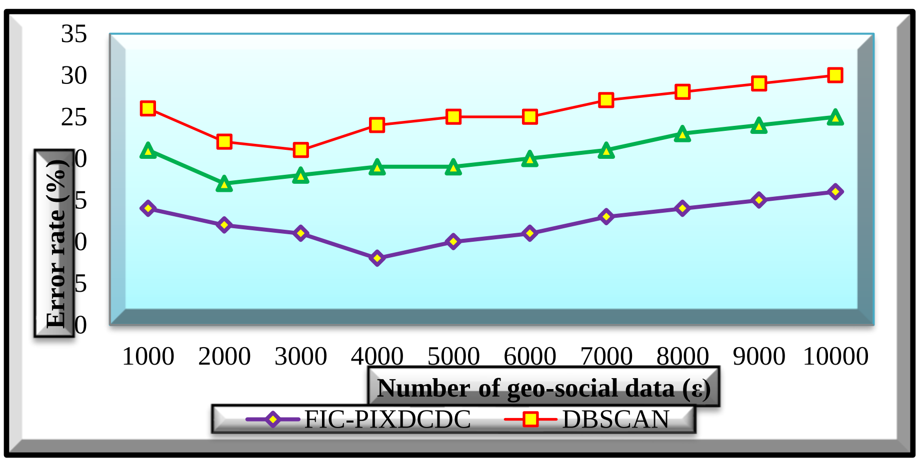

Figure 5: Comparative Result of Clustering Time versus Different Number of Big Geo-Social Data

Figure 5 presents the impact of error rate involved during the big geo-social data clustering along with dissimilar number of big geo-social data in the range of 1000 to 10000 using three methods namely proposed FIC-PIXDCDC and existing DBSCAN [1] DCPGS [2]. As demonstrated in the above graphical diagram, proposed FIC-PIXDCDC method presents lower error rate to correctly group interrelated geo-social data together with increasing number of input geo-social data when compared to traditional DBSCAN [1] DCPGS [2]. This is due to application of Focused Information Criterion and Dice Correlation Coefficient Measurement and Fréchet mean calculation in Partitioned Iterative X-means Clustering algorithm on the contrary to existing works.

By using the Focused Information Criterion, proposed FIC-PIXDCDC method considers the dice correlation between each input geo-social data and centorids to group interrelated data together with an enhanced accuracy. Accordingly, FIC-PIXDCDC method clusters the geo-social data to the cluster whose dice correlation value from the cluster centroid is higher of all the cluster centroid. This supports for proposed FIC-PIXDCDC method to minimize the ratio of number of geo-social data incorrectly clustered when compared to other state-of-the-art works [1] and [2]. Therefore, proposed FIC-PIXDCDC method gets minimal error rate for clustering big geo-social data and thereby determining location information of frequently visited users in social network by 52 % as compared to DBSCAN [1] and 40 % when compared to DCPGS [2].

LITERATURE SURVEY

Emotional maps based on social networks data were developed in [12] to examine cities emotional structure and determine their emotional similarity. A community detection algorithm was utilized in [13] for discovering travel region with help of location-based social network check-in information.

Density-based clustering and thread-based aggregation techniques was presented in [14] to identify unexpected behavior in a city. A Geo-visual analytic approach was introduced in [15] to finding geo-social connections in the international trade network.

An improved Density-Based Spatio–Textual Clustering was accomplished in [16] for analyzing social media with a minimal computational complexity. An effective framework was designed in [17] to identify the most popular place or venue in a given location depends on the tips given by user.

K-mean Clustering and Geocoding technique was employed in [18] to precisely find the latitude and longitude information of the user's friends. A novel framework was presented in [19] by using Geo-Self-Organizing Maps (GeoSOMs) to discover the similar areas of social interaction in cities.

Visual analytics of geo-social interaction patterns was developed in [20] to study the effectiveness of designing control approach. An novel data mining methodology was designed in [21] for analysis of social data sets and thereby solving natural challenges.

CONCLUSION

An efficient FIC-PIXDCDC method is proposed in this research work with the objective of increasing the clustering performance of big geo-social data with a minimal error rate. The objective of FIC-PIXDCDC method is attained with the application of Focused Information Criterion, Dice Correlation Coefficient Measurement, Fréchet mean calculation and Partitioned Iterative X-means Clustering algorithm on the contrary to traditional works. The designed FIC-PIXDCDC method increases the ratio of number of geo-social data that are properly grouped when compared to existing works. In addition to that, proposed FIC-PIXDCDC method minimize the amount of time needed to cluster same type of geo-social data into a diverse number of clusters when compared to other conventional works. Moreover, proposed FIC-PIXDCDC method decreases ratio of number of geo-social data inaccurately clustered to effectively identify location information of frequently visited users in social network when compared to other state-of-the-art works. Hence, proposed FIC-PIXDCDC method gives better performance in terms of accuracy, time and error rate for clustering big geo-social data as compared to existing works. The experimental result shows that the proposed FIC-PIXDCDC method provides better geo-social data analytics performance with an improvement of clustering accuracy and reduction of clustering time for large volume of geo-social data when compared to state-of-the-art works.

REFERENCES

1. Tianhui Fan. Naijing Guo. Yujie Ren, "Consumer clusters detection with geo-tagged social network data using DBSCAN algorithm: a case study of the Pearl River Delta in China", GeoJournal, Springer, Pages 1–21, September 2019
2. Dingming Wu, Jieming Shi, and Nikos Mamoulis, "Density-based Place Clustering using Geo-Social Network Data", IEEE Transactions on Knowledge and Data Engineering, Volume 30, Issue 5, Pages 838 – 851, May 2018
3. Giacomo Aletti and Alessandra Micheletti, "A clustering algorithm for multivariate data streams with correlated components", Journal of Big Data, Volume 4, Issue 48, Pages 1-20, 2017
4. Yun Wu, Zhiquan He, Hao Lin, Yufei Zheng, Jingfen Zhang and Dong Xu, "A Fast Projection-Based Algorithm for Clustering Big Data", Interdisciplinary Sciences, Computational Life Sciences, Springer, Pages 1-7, June 2018
5. Cen Chen, Kenli Li, Aijia Ouyang, Zeng Zeng and Keqin Li, "GFlink: An In-Memory Computing Architecture on Heterogeneous CPU-GPU Clusters for Big Data", IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Parallel and Distributed Systems, Volume 29, Issue 6, Pages 1275-1288, June 2018
6. Charlie Kingston, Jason R. C. Nurse, IoannisAgrafiotis& Andrew Burke Milich, "Using semantic clustering to support situation awareness on Twitter: the case of world views", Human-centric Computing and Information Sciences, Springer, Volume 8, Issue 22, Pages 1-31, 2018
7. Yogesh Sharma, Aastha Jaie, Heena Garg, Sagar Kumar, "A Review paper on Big Data Analytics of Geo-Social Media", International Journal of Computer Science Trends and Technology (IJCST), Volume 6, Issue 6, Pages 1-6, 2018
8. Wei Luo, Peng Gao, Susan Cassels, "A large-scale location-based social network to understanding the impact of human geo-social interaction patterns on vaccination strategies in an urbanized area", Computers, Environment and Urban Systems, Elsevier, Volume 72, Pages 78-87, November 2018
9. Vanessa Peña-Araya, Mauricio Quezada, Barbara Poblete & Denis Parra, "Gaining historical and international relations insights from social media: spatio-temporal real-world news analysis using Twitter", EPJ Data Science, Springer, Volume 6, Issue 25, Pages 1-35, 2017
10. M. Mazhar Rathore, Awais Ahmad, Anand Paul, Won-Hwa Hong, Hyun Cheol Seo, "Advanced computing model for geosocial media using big data analytics", Multimedia Tools and Applications, Springer, Volume 76, Issue 23, Pages 24767–24787, December 2017
11. S. Sindhuja, M.A. Thansirabanu, M. Sowndharya, C. Gokulapriya, "Comparative study on Efficient Analyses and Inference of Geo-social media to make Real-time decisions in Big-Data", International Journal of Scientific Research and Review, Volume 7, Issue 2, Pages 58-62, 2018
12. Soheila Ashkezari-Toussi, Mohammad Kamel, Hadi Sadoghi-Yazdi, "Emotional maps based on social networks data to analyze cities emotional structure and measure their emotional similarity", Cities, Elsevier, Volume 86, Pages 113-124, March 2019
13. Avradip Sen andLinus W. Dietz, "Identifying Travel Regions Using Location-Based Social Network Check-in Data", Front. Big Data, Volume 2, Issue 12, June 2019
14. Héctor Cerezo-Costasa, Ana Fernández-Vilas, Manuela Martín-Vicente, RebecaP. Díaz-Redondo, "Discovering geo-dependent stories by combining density-based clustering and thread-based aggregation techniques", Expert Systems with Applications, Elsevier, Volume 95, Pages 32-42, April 2018
15. Wei Luo, Peifeng Yin, Qian Di, Frank Hardisty, Alan M. MacEachren, "A Geovisual Analytic Approach to Understanding Geo-Social Relationships in the International Trade Network", PLoS ONE, Volume 9, Issue 2, Pages 1-12, February 2014
16. Minh D. Nguyen and Won-Yong Shin, "Improved Density-Based Spatio–Textual Clustering on Social Media", IEEE

Transactions on Knowledge and Data Engineering, Pages 1-26, November 2017

17. G. Vishnu Murthy, K. Priya Darshini, G.Balakrishna, "Analyzing Geo-social Media for Decision Making", International Journal of Engineering and Advanced Technology (IJEAT), Volume-8, Issue-3, Pages 564-569, February 2019
18. S. Hemamalini, K. Kannan and S. Pradeepa, "Location Prediction of Twitter User based on Friends and Followers", International Journal of Pure and Applied Mathematics, Volume 118, Issue 18, Pages 2817-2824, 2018
19. Achilleas Psyllidis,Jie Yang, And Alessandro Bozzon, "Regionalization of Social Interactions and Points-of-Interest Location Prediction with Geosocial Data", IEEE Access, Volume 6, Pages 34334-34353, 2018
20. Wei Luo, "Visual analytics of geo-social interaction patterns for epidemic control", International Journal of Health Geographics, Volume 15, Issue 28, Pages 1-16, 2016
21. Camila Maione, Donald R.Nelson, Rommel Melgaço Barbosa, "Research on social data by means of cluster analysis", Applied Computing and Informatics, Elsevier, Volume 15, Issue 2,, Pages 153-162, July 2019
22. Weeplaces dataset: https://www.yongliu.org/datasets/