Received: August 29, 2017

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Location Aware Social Networks User Profiling Using Big Data Analytics

Venu Gopalachari Mukkamula[1]* **Lavanya Nangunuri [1]**

[1]*Department of Computer Science and Engineering,*
Chaitanya Bharathi Institute of Technology, Hyderabad, India
* Corresponding author's Email: venugopal.m07@gmail.com

Abstract: Social Network Analytics (SNA), on the other hand, provides insights of many perspectives of the society through the sample users participated in social network. The exponential growth of the social network and correspondingly its data leads to the demand for big data computational environments. One such popular and useful big data in SNA is GIS (Geographical Information System) data that provides geographical location data of the record generated in social networks. In order to profile a user in social networks according to GIS data, the existing methodologies uses the centroid measures such as mean, median of the GIS data available for the user. These methods failed to mention and solve the serious issues such as cold user as well does not able to consider the weightage of the data item in analytics. The proposed method in this paper focuses to define the weightage of a data item in the available GIS data according to the application and also proposed a method to identify and solve cold user problem. The experiments carried on cutting edge technologies of the big data analytics shown significance of the proposed method in profiling user w.r.t. GIS data. The results of the proposed user profiling is measured with Within Sum of Squares (WSS) measure and compared over exiting profiling methodologies shown consistent improvement for any number of clusters formed over benchmark datasets.

Keywords: Big data, Location based social networks, User profiling, GIS data, Cold user problem.

1. Introduction

Big data analytics grabbed its significance in computing world to design solutions of advanced web services in various application domains such as E-commerce, E-tourism, Gene analysis, Social Networks etc. Due to the generation of huge amount of data for every second, the business organizations matters how to use big data to sustain their growth in competence. The exponential growth in the popularity of the social networks in the world from the past decade is generating huge data in turn leaving many challenges in analytics. The trend of mobile terminals and internet on mobile ease the users across world to access social network anytime deriving the role and group of the user. This made the interactions among users in social network became more comprehensive [1]. The GIS data attracted the generation with widespread use of location-aware applications such as Google maps, yelp local search etc. On the other hand location aware social networks such as QZone, Foursquare are generating the user's preference information along with the location information. Though there exists traditional search application using location data such as location based keyword queries [2], they failed to involve the personalized location aware data analytics at full extent [3]. The improvement in identifying the user's preferred region with the functionality features such as check-ins, residential access, business zones etc., definitely helps the user by the better services with the better understanding of the selected region. These identified user profiles according to the location could help even the applications such as crime detection, rare event analysis by considering the dense of the location. The architecture of the proposed GIS based analytics process on social networks is shown in figure 1.

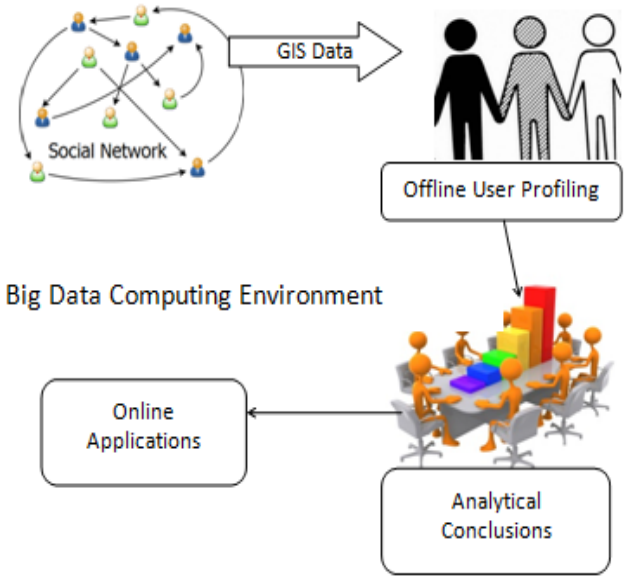


Figure.1 Architecture of GIS based analytics on social networks

This process has mainly three process steps as part of analytics, first one is user profiling in offline on the input GIS data along with the other information in social networks. This step said to be offline because of the key role of the historic data involved in the analytics that means in order to identify user interest patterns the user no need to be online. The second step in this process is drawing analytics conclusions which could involve other information such as similar users, demographic data and techniques such as machine learning. . The third step is the development of an online application such as recommender system, query engine with the objective of serving end user. This paper focuses on the user profiling step which generally suffers with cold start problem, which occurs when the historical information is low to analyse.

The exiting methodologies uses user profiling techniques such as mean distribution and median distribution are used to predict the location of the user making every past access record with equal weightage. Mean or median of both latitude and longitude of each user is calculated and the obtained value is considered as active location. The obtained location may not be the precise because the outliers affect the value and deviate from the actual location. These points may not give the actual position of the user leads to degradation in profiling. Because of privacy concerns a few users may not update their locations or few users may visit only a few locations with less frequency, becomes cold user, affects the accuracy of the clusters.

This paper focused to resolve the mentioned limitations by considering different weights for the access records and also modified the profiling strategy by means of mode measure instead of mean or median measure. This paper proposes:

a. A novel method to define the location centroid of the user resolving cold user problem.
b. Developed a big data computation process to improve the performance.

The main advantage of the proposed user profiling methodology is the quality maintained in pre-processing that in turn leads the quality user clusters. The experiments conducted on apache spark environment for two benchmark location based social networking datasets with different clustering techniques. The performance measured is compared and shown significant improvement in the proposed method.

The rest of the paper is organized as follows. The related work is presented in the next section and then the architecture of the proposed LBSN based user profiling is explained. There after results and analysis shows the experimentation part followed by conclusion.

2. Related work

Big data analytics is not just the processing of datasets with large volume but also the processing of datasets which increase in volume with great velocity embedding varied datatypes. The analytics sometimes may result in uncertainty and all these four V's-volume, velocity, variety, veracity is considered as characteristics of big data. Big data analysis helps to uncover the hidden patterns, draw unknown correlations, customer preferences, market trends. Because of its high yielding benefits many organizations are working on big data analytics. One of such domain where big data has got lot of importance is the data that is being generated from social networking websites [4].

Each day billions of posts, messages and likes are being sent from each social networking websites and it is estimated that by 2020 the amount of data generated in a minute with reach 44 zeta bytes which is equal to 44 trillion Gigabytes. All these statistics led the business organizations to focus on social network analysis [5]. Social network analytics involves three stages namely capture, understand and present. Capture stage involves collecting data from different sources; pre-process the data and then extracting relevant information from the data. The second step as prescribed is to understand the extracted information. It involves removing noisy data, performing some advanced analytics such as

social network analysis, opinion mining, trend analysis, sentiment analysis. The final step is presentation of the analysis [6]. It involves summarizing and evaluating the findings from the above step and presenting those findings.

This social network analysis helps in analysing a social network graph and to understand its underlying structure, connections, and theoretical properties as well as to identify the relative importance of different nodes within the network. Structure of social network graph consists of nodes and edges where nodes represent the users and edges represent the relation between the users. Social network analytics is used to model a social network dynamics and its growth. This in turn helps to monitor business activity. Social network analysis is a predominant technique for identifying salient influencers in advertising or marketing campaigns in different social networking websites. They are used to identify sub communities within a larger community.

When speaking about marketing campaigns, they work successfully when focusing on the users is done based on their location because it is probable that people of same location have similar life style and so similar preferences. One of such social networking sites are called as location based social networking websites. The availability of huge amounts of geographical and social data on Location Based Social Networks (LBSN) provides an unprecedented opportunity to study the behaviour of human mobility through data analysis in a spatial-temporal-social context, enabling a variety of location based services, from mobile marketing to disaster relief [7] proposed a model for the development of destination choice model using location based social networking website. This model did not consider the conventional issues such as data sparsity in analysing the social network data. Since LBSN involves lot of check-ins from its users, the active locations of the users have to be traced for the effectiveness of developing applications such as recommendations, analysing behaviour of a particular spatial locality. The proposed technique in [8] has explained recommendations in LBSN based on spatio-temporal data. A GIS based framework called spatial Hadoop is used, which provides options like grid file, R tree and R+ tree to handle spatial data [9]. Dart is another spatial analysing system that solves spatial tasks like K- nearest neighbours and geometric mean distribution for social media analytics[10]. The geometric mean distribution use in Dart would not pin point a location that user can be profiled which challenges the quality of user profiling. Hadoop-GIS utilize global partition indexing and customizable on demand local spatial indexing to achieve efficient query processing [11]. Hadoop-GIS is integrated into Hive to support declarative spatial queries with an integrated architecture. There also a method to find the active locations of the user [12]. But the active locations of the user found never focus on cold user issue. In [13], authors proposed a clustering method using Hadoop-GIS which deals with huge amount of data and so numerous calculations, the performance comes into picture.

The quality of the centroid measure of the user in all the mentioned systems is not focused much and none of the existing system considered cold user problem in profiling a user which may leads to the degradation of the analytics based on these profiles. The proposed location based user profiling is different from existing approaches for its way of dealing GIS data of the users in profiling.

3. Proposed user profiling in LBSN

The framework of the proposed methodology for user profiling is shown in figure 2. The frame work initially takes the social network raw data set as input. The data set under gone with pre-processing in order to make flexible for the analysis. The process is applied as individual entities w.r.t. user and selected as personalized data. This also extended to define the active location of the user. Then the clustering technique will group similar users.

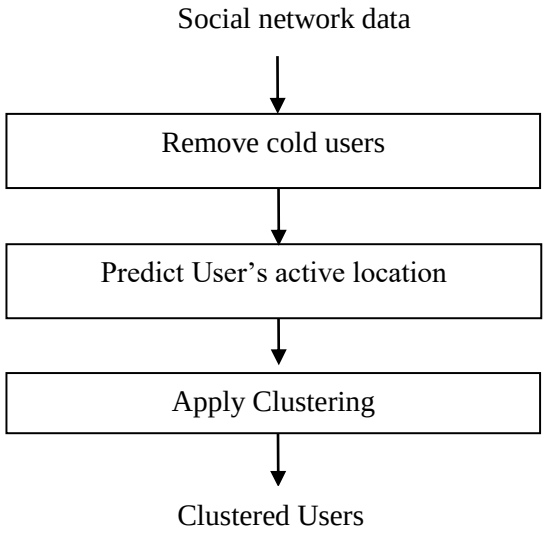


Figure.2 Proposed framework for clustering with GIS data

3.1 Cold user removal

The collected dataset undergoes pre-process. The irrelevant attributes if any in the dataset are removed. Individual users are grouped against the locations they visited along with its count. The users whose visit count is less than the minimum frequency value are considered as cold users and removed. Thereby the dataset consists of active users with attributes of user id, latitude and longitude values.

3.2 Predict user's active location

In this method all the locations of a single user are traced out. Let the user be 'x' who visited 'n' locations and the locations are indicated by latitude (Lat_x) and longitude (Lon_x).

$Lat_x = \{lat_1, lat_2, lat_3, \ldots\ldots\ldots\ldots lat_n\}$
$Lon_x = \{lon_1, lon_2, lon_3, \ldots\ldots\ldots\ldots lon_n\}$

Now the active location of the user is obtained by calculating the mean of all these locations.
Where

$$Mean_{lat} = \sum_{i=1}^{n} \frac{lat_i}{n} \quad (1)$$

$$Mean_{lon} = \sum_{i=1}^{n} \frac{lon_i}{n} \quad (2)$$

Mean of both latitude and longitude of each user is calculated and the obtained value is considered as active location. The active locations of the users are tabulated with user-id, latitude and longitude as attribute values. But the disadvantage of this method is that the obtained location may not be the precise because of the outlier's effect on the value and causal deviation from the actual location which may not give the accurate position of the user. On the other hand, mode of the location access record, that is the frequency of visits to the particular location, may serve better in finding centroid of the user. For each individual entity, count of each visited location is calculated and the location with highest count is considered as the active location.

$$\forall Users\ Mode\ (U_i) = max\ \{(lat_j, lon_j, count)\}$$

where

$$1 \leq i \leq N, 1 \leq j \leq k$$

N→Number of users in the pre-processed dataset
K→Number of locations of i^{th} user.

The active location of the users with user id, latitude and longitude values and count of the location are tabulated which acts as the final dataset of profiled users.

Step 1: Load the dataset into Hadoop Distributed File System (HDFS).

Table = load (dataset) as U_{id}, V_{id}, loc_{lat}, loc_{lon}

Step 2: Find the count of user visiting a specific venue

$\forall$ ***table,*** *Count_tab*
=Generate (U_{id}, V_{id}, count (U_{id}, V_{id}))

Step 2: Remove cold users who have visited a venue less than min_value

$\forall$ ***count_tab,****active_users*
=Generate (U_{id}, Vid,count(U_{id},V_{id})>min)

Step 3: Create a table which has active users with attributes user_id, latitude, longitude and count of visit

$\forall$ ***count_tab,*** $\forall$ ***table****, Join_tab*
=Generate (U_{id}, loc_{lat}, loc_{lon}, count)

Step 4: Calculate the location centroid of each user using 'mode' metric

$\forall$ ***Join_tab.Uid,****mode_tab*
*=Generate (**(U_{id}, loc_{lat}, loc_{lon}, Mode (U_i))*
= Max {(lat_j, lon_j, count)})

where

$$1 \leq i \leq N, 1 \leq j \leq k$$

N→Number of users in the pre-processed dataset
K→Number of locations of i^{th} user.

Step 5: Tabulate the values of user id and its centroid location

$\forall$ ***mode_tab,*** *final_data*
=Generate (U_{id}, loc_{lat}, loc_{lon}, Mode (U_i))

Step 6: Cluster the users using k-Means and bisecting k-means clustering techniques.

Thus the proposed framework could use active locations of the user for profiling user.

3.3 Clustering

Clustering is an unsupervised technique that groups the data points based on the similarity or distance measures applicable to the data set. The proposed framework used two popular clustering techniques named k-means and bisecting k-means for experimentation. The process of k-means clustering is an iterative process that assigns the data points to one of the cluster. Assume that there are N multi-dimensional data points $P = \{p_1,p_2,p_3 \ldots p_N\}$ are participating in the clustering where each data point p_i has to be assigned into one of the $K=\{C_1,C_2,C_3 \ldots C_K\}$ number of clusters. The centroid of the cluster C_i is represented by $\bar{C}_i$, which is the mean of the data points exists in that cluster upto that iteration as shown in the following equation.

Each cluster centroid is initially assigned with random data points or also may be assigned as proposed in [15]. For each iteration of the clustering technique, a data point pi is assigned to a cluster C_j if the distance from centroid of the cluster C_j to the point p_i is minimum among all the clusters. The proposed work used Euclidean distance measure to find the distance between cluster centroid and data point.

$$C_i \leftarrow arg\ {}^{min}_{k} \|p_k - \bar{C}_k\|^2 \tag{3}$$

$$\bar{C}_k = \frac{\sum_{i=1}^{n} f(C_i,k) \cdot p_i}{\sum_{i=1}^{n} f(C_i,k)} \tag{4}$$

Where
$1 \le i \le n, 1 \le k \le K$

$$f(C_i,k) = \begin{cases} 1\ if\ C_i == j \\ 0\ otherwise \end{cases} \tag{5}$$

$n \leftarrow$ Number of data points in the data set
$K \leftarrow$ Number of clusters determined

At each iteration of the k-means technique the cluster centroids are recomputed and all data points are reassigned to their nearest clusters. When there is no or little change in the cluster centroids the algorithm terminates its iterations. The second technique used in the proposed framework is bisecting k-Means which has a theme of combination of k-means and hierarchical clustering. Bisecting K-means is an improvement of the K-Means in both clustering quality and efficiency. Steinbach, Karypis et.al. introduced bisecting K-means [16] in 2000, with which a data set is first partitioned to two clusters using the Lloyd two-means(i.e. the K-means with K being two), resulting in two clusters. Before the total number of clusters reaches K, a cluster from the current pool of clusters is chosen to be partitioned into two clusters and total numbers of clusters is incremented by one. This bisecting process continues until the total number of clusters reaches K. Multiple runs of the two-means are performed on the same cluster that is chosen to bisect, and the best bisection result is chosen to have good clustering result in bisecting a cluster.

The algorithmic steps of bisecting K-means are as follows:

Step 1: (Initialization).
Randomly select a point, say $p_L \epsilon P$;
Compute the centroid $\bar{C}$ of P;
Compute $p_R \epsilon P$ as $p_R = \bar{C} - (p_L - \bar{C})$

Step 2: Divide P into two sub-clusters P_L and P_R, according to the following rule:

$$\begin{cases} p_i \in P_L\ if\ \|p_i - p_L\| \le \|p_i - p_R\| \\ p_i \in P_R\ if\ \|p_i - p_L\| > \|p_i - p_R\| \end{cases} \tag{6}$$

Step 3: Compute the centroids of P_L ($\bar{C}_L$) and P_R ($\bar{C}_R$)

Step 4: If $p_{L==\overline{C_L}}$ and $p_{R==\overline{C_R}}$,
Terminate iterations.
Else,
$p_L \leftarrow \overline{C_L}$
$p_R \leftarrow \overline{C_R}$
Go to Step 2.

Clustering techniques are applied individually on the collected datasets which are obtained after applying location based tendency measures. The values of sum of squared errors within clusters (WSS) are considered to compare the effectiveness of the clusters.

$$\sum_{i=1}^{K} \sum_{j \in P_i} \sum_{k=1}^{V} \left(p_{jk} - \overline{p_{ik}}\right)^2 \tag{7}$$

Where

K ← number of clusters formed after clustering
P_i ←Set of data points in the i[th] cluster
V ← Set of variables involved in the clustering process
$\overline{p_{ik}}$ ← Mean of the k[th] variable of all points in the i[th] cluster.

This WSS measure will give error value where the lesser value indicates better performance of the

clustering. But when the number of clusters increases and when the number of samples in the cluster decreases, the drop in WSS value is obvious.

4. Results and discussions

The data is collected from two datasets –gowalla and brightkite. Gowalla is a location-based social networking website and the users share their locations by checking-in. The friendship network here is undirected and was collected using their public API. It consists of 196,591 nodes and 950,327 edges. This dataset consists of 6,442,890 check-ins. This data is collected over the period of Feb. 2009 - Oct. 2010.

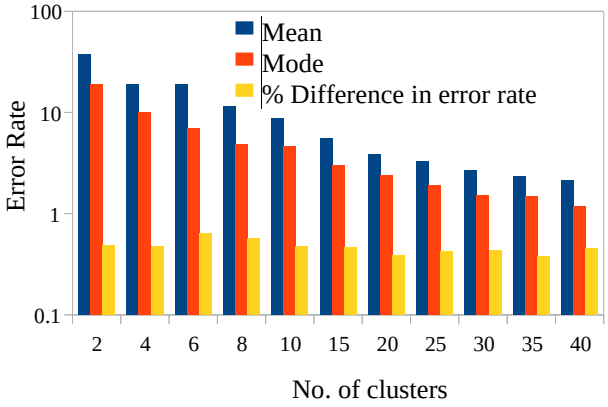
Brightkite was once a location-based social networking service provider where users shared their locations by checking-in. The friendship network was collected using their public API, and consists of 58,228 nodes and 214,078 edges. The network is originally directed but has constructed a network with undirected edges when there is a friendship in both ways. The datasets consists of a total of 4,491,143 check-ins of the users over the period of Apr. 2008 - Oct. 2010.

Since the dataset has no unwanted attributes, no attributes are removed as a part of pre-process step. The project focuses on comparing the user profiling methods based on the two different centroid measures [12]. As a part of mean based profiling method, the active locations of the users are found using mean measure and then clustered using pyspark which is a better framework than Hadoop. In [14] it is introduced Spark which runs faster than Hadoop and hence reduce the computation time. Similarly, mode based profiling method is applied to find the active location of the users and the resultant dataset is clustered using KMeans clustering technique. Sum of squared errors is obtained from both clusters in order to compare the effectiveness of the two profiling methods. During clustering, the number of clusters ranged from 2 to 40. Within sum of squared errors is obtained each time when clustered as shown in table 1.

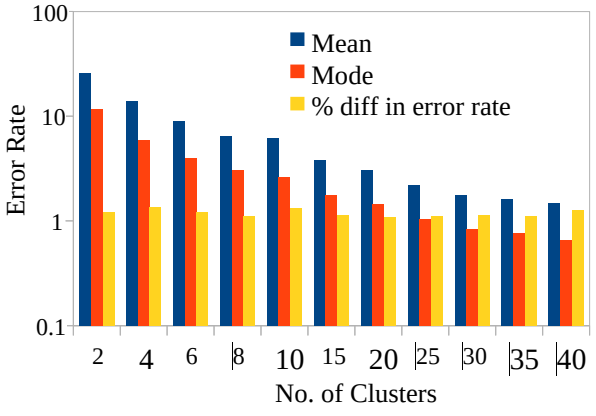
A graph is plotted to indicate the results as shown in fig 4. The graph has a scaling of no. of clusters on X-Axis and error rate on Y-Axis. Both the error rates of the clusters formed based on mean and mode based profiling techniques are plotted and are compared. When observed, the difference between the error rates of clusters which are formed based on mean based profiling method and mode based profiling method are almost always consistent. i.e., no matter what the number of clusters to be, but still there is a clear improvement in the quality of clusters which are formed based mode based clustering technique. Another clustering technique, KMeans bisecting clustering is applied and same results were obtained which showed user profiling based on mode method is better. By nature of WSS reduces when number of clusters is increasing and at the same time the performance hike in the proposed method is observed to be constant.

Table 1. WSS values occurred after clustering on gowalla and brightkite data sets

	KMeans clustering						**Bisecting KMeans clustering**					
	Gowalla dataset			**Brightkite dataset**			**Gowalla dataset**			**Brightkite dataset**		
#	**A**	**B**	**%**	**A**	**B**	**%**	**A**	**B**	**%**	**A**	**B**	**%**
2	26.62	11.62	1.20	37.51	19.12	0.49	835.06	367.74	0.56	363.98	183.08	0.50
4	13.87	5.92	1.34	18.85	9.93	0.47	214.12	96.60	0.55	90.25	47.62	0.47
6	8.87	4.01	1.21	18.85	6.90	0.63	135.38	64.28	0.53	56.81	29.86	0.47
8	6.43	3.04	1.12	11.39	4.84	0.58	53.02	25.02	0.53	22.68	11.62	0.49
10	6.09	2.61	1.33	8.69	4.60	0.47	40.61	19.68	0.52	18.43	9.45	0.49
15	3.80	1.78	1.14	5.58	3.01	0.46	17.60	9.11	0.48	7.91	4.24	0.46
20	3.03	1.46	1.09	3.87	2.37	0.39	10.77	4.99	0.54	4.64	2.50	0.46
25	2.18	1.03	1.11	3.32	1.91	0.43	7.48	3.82	0.49	3.31	1.83	0.45
30	1.76	0.83	1.13	2.65	1.50	0.43	4.32	2.09	0.52	1.97	1.13	0.43
35	1.61	0.76	1.11	2.37	1.48	0.37	3.01	1.39	0.54	1.36	0.78	0.42
40	1.48	0.65	1.26	2.14	1.17	0.45	2.58	1.24	0.52	1.19	0.71	0.41
#--Number of Clusters, A—Mean Measure, B—Mode Measure, %-- difference in error rate												

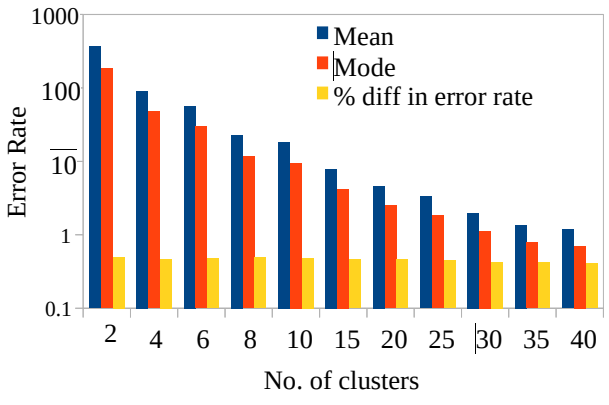


(a)

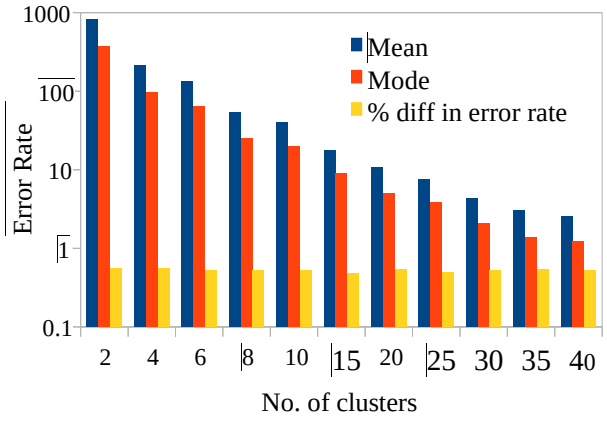


(b)

Figure.3 Comparison of error rate for mean and mode measures using K-means on: (a) brightkite dataset and (b) gowalla data set



(a)



(b)

Figure.4 Comparison of error rate for mean and mode measures using bisecting K-means on: (a) brightkite dataset and (b) gowalla data set

5. Conclusion and future work

The methods which use mean technique to calculate the location of the user may deviate from its actual active location. Since all the users don't provide the location due to various security reasons the dataset contain cold users too. Hence this method results in the decline in the quality of clusters thus formed. When the cold users are removed and mode technique is applied to find the active location of the user the clusters thus formed is having less sum of squared error rate. Users are clustered using pyspark which is faster than Hadoop and thereby improving the performance. The future work can be carried on by creating applications related to social networking such as recommendations systems, or applications which focus on marketing or disaster relief systems.

References

[1] Z. Zhang, X. Wang, and C. Zhao, "A Situational Analytic Method for User Behaviour Pattern in Multimedia Social Networks", *IEEE Transactions on Big Data*, pp.1-10, 2017.

[2] Y. Liua, H. Wang, G. Li, J. Gao, H. Hu, and W. Li "ELAN: An Efficient Location-Aware Analytics System", *Big Data Research, Elsevier*, Vol.5, pp. 16-21, 2016.

[3] X. Xiong and D. Jiang, "Empirical Analysis and Modelling of the Activity Dilemmas in Big Social Networks", *IEEE Journals & Magazines*, Vol.5, pp.967-97, 2017.

[4] Y. Zheng, "Methodologies for cross-domain data fusion: An overview", *IEEE Transactions on Big Data*, Vol. 1, No. 1, pp. 16–34, 2015.

[5] W. Fan, Michael D. Gordon. "The Power of Social Media Analytics", *Communications of the ACM*, Vol. 57, No. 6, pp. 74-81, 2014.

[6] W. Bi, M. Cai, M. Liu, and G. Li, "A big data clustering algorithm for mitigating the risk of customer churn," *IEEE Transactions on*

Industrial Informatics, Vol. 12, No. 3, pp. 1270–1281, 2016.

[7] H. Gao and H. Liu. "Data Analysis on Location-Based Social Networks", *Mobile Social Networking*, pp. 165-194, 2013.

[8] Y. Sun, "Recommendation in Location-Based Social Networks Based on Spatio-Temporal Data", PhD thesis, *Computing and Information Systems*, pp.9-24, 2017.

[9] A. Eldawy and M. Mokbel. "SpatialHadoop: towards flexible and scalable spatial processing using mapreduce", In: *Proc. of the 2014 SIGMOD*, pp. 46-50, 2014.

[10] H. Zhang, Z. Sun, Z. Liu, C. Xu, and L. Wang. "Dart: A geographic information system on Hadoop", In: *Proc. of the 2015 IEEE 8th International Conference on Cloud Computing*, pp. 90-97, 2015.

[11] A. Aji, F. Wang, H. Vo, R. Lee, Q. Liu, X. Zhang, and J. Saltz. "Hadoop-GIS: A high performance spatial data warehousing system over mapreduce", In: *Proc. of the VLDB Endowment*, Vol. 6, No. 11, pp. 1009-1020, 2013.

[12] Z. Sun, H. Zhang, Z. Liu, C. Xu, and L. Wang, "Migrating GIS Big Data Computing from Hadoop to Spark: An exemplary study using Twitter", In: *Proc. of the 2016 IEEE 9th International Conference on Cloud Computing*, pp. 351-358, 2016.

[13] J. Cao, Y. Zhou, and M. Wu, "Adaptive Grid-Based k-median Clustering of Streaming Data with Accuracy Guarantee", In: *Proc. of the International Conference on Database Systems for Advanced Applications*, pp. 75-91, 2015.

[14] M. Zaharia and M. Chowdhury, "Fast and Interactive Analytics over Hadoop Data with Spark", *Networked Systems*, Vol. 37, No. 4, pp. 45-51, 2012.

[15] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering", In: *Proc. of the Fifteenth International Conference on Machine Learning*, Vol. 98, pp. 91–99, 1998.

[16] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques", In: *Proc. of World Text Mining Conference, KDD Workshop on Text Mining*, pp.1-20, 2000.