Point-of-Interest Based Classification of Similar Users by Using Support Vector Machine and Status Homophily

K. Mohan Kumar and B. Srinivasan

***Abstract*—Online social networks (OSN) are becoming an indispensable part of everyday life. Apart from allowing various users to perform a vast amount of conventional sharing per second, the networks also provide location-based services, like sharing of location, traveling activities, and performing check-ins. These new dimensions changed the traditional OSN into location-based social networks (LBSN). This paper proposes a novel approach for finding similar users, based on check-ins or points-of-interest (POI) as the key feature and by building a model using a support vector machine (SVM) and status homophily. Furthermore, the performance of the approach and the accuracy of grouping are analyzed in the result analysis section.**

***Index Terms*—Similar users, LBSN, check-ins, POI, support vector machine, status homophily.**

I. Introduction

In the beginning stages of online social networking (OSN), all social networks allowed their users to do conventional social tie-up processing through sharing of posts, photos, videos, and publishing or replying to comments, etc. After the emergence of smartphones, social networks began to move into the field of the mobile phone industry. The moving nature of mobile phones caused social networks to bring newer dimensionalities among users. The newer functionalities include sharing of locations by means of check-ins, posting of traveling activities, action activities like watching movies, drinking coffee, etc., with location information by including the location information with the users' posts. These newer activities changed the traditional OSN into location-based social networks (LBSN). Social networks like Gowalla and Foursquare are the famous LBSN, which allow users to perform location-based activities like checking in to a location, sharing location information, which would allow a greater chance for the users of such networks to find nearer places and friends across these networks. Leading social networks like Facebook and Twitter are also allowing location-based activities so that the users of such networks may produce location-aware information through location-based services. Introduction of such services not only produced a new shape among the users of social networks but also yielded tremendous opportunities for researchers and business organizations. Data produced by the users of LBSN contain huge amounts of hidden information about the location and the relationship between the user and the location of other users. By using the rough information, researchers may research among recommender systems to find similar users. Business organizations will find a way of marketing varieties of products among different users who have visited the business point on a location.

In the beginning stages of location-based services, users just shared their current location without any aim. However, recent location-based services are making locations with a specialized meaning and made them points-of-interest (POI). A POI is a place of attraction and affection arises from the mind of users from different parts of the world. So, now the check-ins activity of users over POI may lead to a grouping of similar users who are from the same place or belong to the same institution, company, or organization.

This paper tries to find the nearer and similar users by using the POI location check-ins as the key feature and constructing a classification model by using the support vector machine (SVM) along with the support of status homophily of each user across a social network.

Manuscript received January 26, 2019; revised June 15, 2019.

The authors are with PG & Research Department of Computer Science, Rajahs Serfoji Govt. College, Thanjavur, India (e-mail: tnjmohankumar@gmail.com, prof.b.srini@gmail.com).

II. SVM

SVM or support vector network (SVN) is a binary classifier model that discriminatively produces new predictions from the given set of training data. The model produces an optimal hyperplane with new predictions existing on each side of the hyperplane.

The main advantage of the SVM is that the model works on a set of points, which are both linearly and non-linearly separable, and produces optimal outputs.

The SVM classifier uses the following formula for constructing classification models:

$$(x) = \sum^{N}_{n=1}(\alpha_n y_n K + b) \quad (1)$$

$$K=K(x_n, x_i) \quad (2)$$

α_n is the coefficient associated with the i^{th} training sample and b is the scalar value used to minimize the distance of hyperplane as well as maximize the margin boundary of the predicted label y_n. N is the total number of training samples. K is the Kernel function with the set of feature domain, which is defined under the following mode:

A. Linear

Linear mode is used when the points are uniformly distributed.

$$K(x_n, x_i)=x_n x_T \quad (3)$$

In linear separation, only the hyperplane is constructed.

doi: 10.18178/ijmlc.2019.9.5.848

B. Non-Linear

This mode is used when the points are randomly distributed. The Non-Linear mode has the following variations:

C. Polynomial

$$K(x_n, x_i)=(\gamma.\ x_n x^T+r)^d \quad (4)$$

where $\gamma>0$

D. Radial Basis Function (RBF)

$$K(x_n, x_i)=exp(-\gamma.||X_n-X^T||^2 \quad (5)$$

where $\gamma>0$

E. Sigmoid

$$K(x_n, x_i)=tanh(\gamma.\ x_n x^T+r) \quad (6)$$

where γ, r, and d are kernel parameters.

This paper uses both linear and non-linear separation approaches with the above-mentioned kernel parameters for classifying users and constructs predictions about a user's chance of becoming a friend of a single user or a group of other users. The predictions are plotted in the result analysis section.

III. Homophily

Homophily is the ancient tendency of trying to tie similar traits across society. In the world of social networks, homophily is used as a metric for assessing similar ties between groups of users. There are two types of homophily: status and value.

i) Status homophily

Status homophily defines the attributed social characteristics of users, such as age, gender, culture, city, country, education, and workplace.

ii) Value homophily

Value homophily considers thoughts, aesthetics, and feelings of the users.

In social networks, the status homophily exists in the form of a user profile and the value homophily exists in terms of the users' likes, posts, and shares. This study uses status homophily for finding the favor of similar ties between two users to become friends after classifying all the users using the SVM classifier. Similar users are the ones who have similar traits or who belong to the same country, institution, and/or organization.

IV. Existing Works

LBSN involves performing a check-in, inserting the present location while publishing a post or publishing an activity such as traveling information. Normally, the interesting locations from the user's point of view are known as POI locations. By using the POI, users can find similar users with similar traits in terms of who carried out the same work or whether the users studied in the same college, school, and/or worked in the same company.

Finding such similar users with respect to a location over social networks is now a leading study performed by many researchers. Several studies have been performed to find and recommend similar users across social networks.

The widely accepted method for finding similar objects would be the proportion of common friends accepted by trending OSN. Filtering methods, such as Jaccard Correlation and Pearson Coefficient, are concepts accepted by these methods. The next most used technique is collaborative filtering, which analyses users' personalized data (status homophily alone) as the source. However, the approach would be useless when users have insufficient data.

Another popular approach used by most of the LBSN is the use of check-ins history. However, the technique only considers the geo distance of users between the users' check-ins.

Similarity measurement techniques like cosine similarity are also used for finding users with similar traits but would take a longer time when the size of the input is large.

Fei Yu and Zhiujun Li *et al*. [1] formulated a study through which the users' interests are analyzed over a pre-fixed geographical area. Manfang Wu and Zhanquan Wang *et al*. [2] devised an approach in which a random walk was performed over a limited amount of a user's node with the aid of Markov Chain and cosine similarity measurements. Sheng-Min Chiu and YI-Chung Chen *et al*. [3] proposed a neuro-fuzzy approach to find similar users, but the performance is low compared to other relevant works. Kunhui Lin and Yating Cheen, *et al*. [4] developed studies with the status homophily attributes of users from the user-centric point of view and degrades with a high number of users.

Chen Yu, Baiyun Xiao, *et al*. [5] proposed a related work to create partitions in terms of the locations of users based on the users' check-ins, but this work lags to produce correct locations because of poor partition tracking. Reshma M and Raji R Pillai [6] developed a study, which only considered content-sharing over a central page and not involved any location-oriented services.

V. Proposed Work

All the existing studies carried out the classification of similar users as the problem of a user-centric process at maximum, in which the binding is considered between the users.

Even though the studies produced results, the accuracy of the predicted results is low when the data used is larger in the count and also when the performance of the process tends to delay.

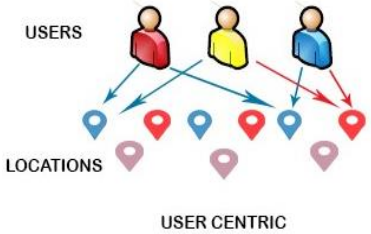

Fig. 1. User-centric binding of similar users.

The proposed work considered the locations or check-ins (POI) as the key feature through which there might be a maximum chance for classifying similar users who are same

in their status homophily, such as place, country, education, and workplace. The working principle of existing works and proposed works is shown in Fig. 1 and Fig. 2.

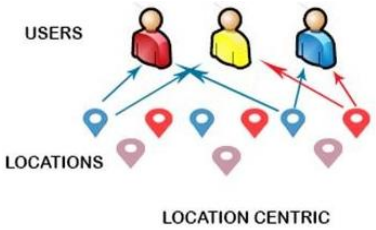

Fig. 2. Location centric binding of similar users.

The present work used the following algorithm, *POISV*, to find the maximum number of similar users by the means of a location with a high accuracy rate.

TABLE I: POISV ALGORITHM FOR GROUPING SIMILAR USERS BELONGING TO SAME POI

Algorithm *POISV(N): Support Vectors* (S_V)
Step 1: Initialize dataset Step 2: Pre-process the data Step 3: Fix the Feature and Label values Step 4: Calculate training and test data Step 5: Apply SVM fit with K, N samples, training, and test data, and get S_V Step 6: Classify the nearer points with the aid of S_V Step 7: If a point is remote, take the homophily of the concerned user and try the closer distance S_V Step 8: Measure the metrics Step 9: Group all the similar users and plot Step 10: Stop

The entire process is depicted in the following block diagram.

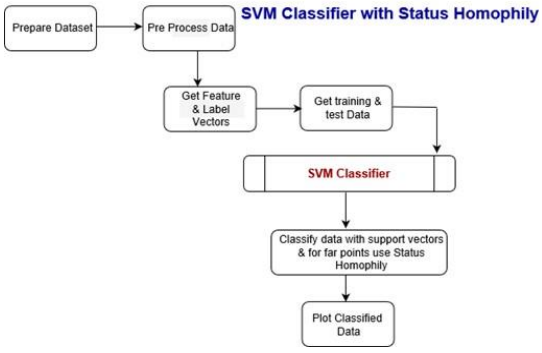

Fig. 3. Steps for classifying similar users with SVM & status homophily.

VI. RESULT ANALYSIS

The present work used the Gowalla check-ins dataset [7] containing around one million check-ins with user visited details like userid, placeid and the time of visiting. SVM classification is carried out with randomly selected samples from that dataset. The present work used placeid and the visited date and time as features, and userid as a label for the classification. The results and observations are plotted below.

Fig. 4–Fig. 7 show the plotting of similar users by check-ins with various samples into two-dimensional space points by the SVM classifier. Each check-in by all users is represented with a different color as a class. The two selected features placid and datetime are plotted on the x-axis and the userid is plotted on the y-axis as a label. The boundaries of the plotting in each the above figure also clearly depicts the SVM classifier decision function for grouping of similar users. Hyperplanes are drawn in between the selected samples in Fig. 4. The arbitrary shape is classifying the users with their respective check-ins in Fig. 5. A closed boundary was constructed with the nearest points in Fig. 6 by the Radial Base Function (RBF) kernel of SVM classifier. A curve boundary with logistic respect to a logistic value is plotted in the Fig. 7.

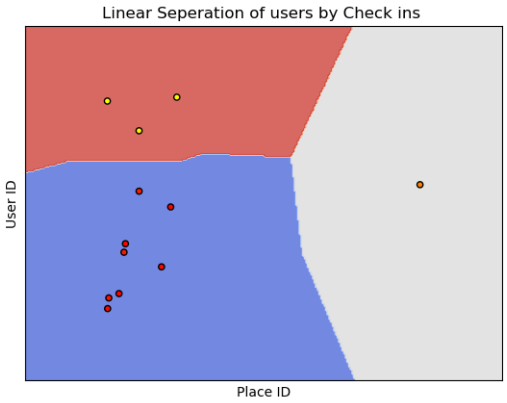

Fig. 4. Linear plotting of similar user group.

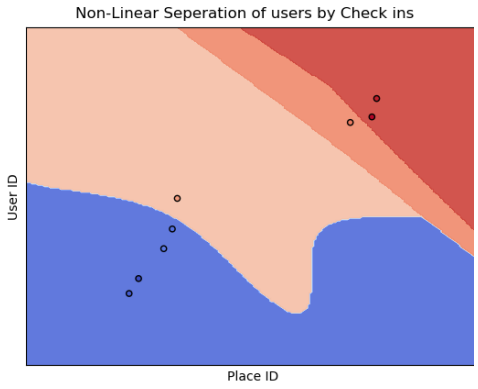

Fig. 5. Non-linear plotting with polynomial kernel for grouping similar user.

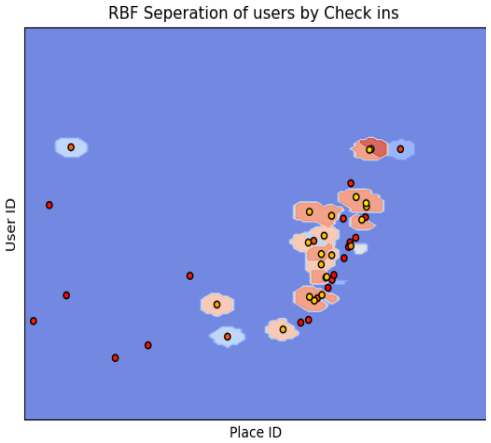

Fig. 6. Non-linear plotting with *RBF* kernel for grouping similar user.

The linear plot in Fig. 4 groups the samples by using the hyperplane in between the plotted points and plots the

similarity of users as linearly separable spaces, and the non-linear plot with polynomial plotting in Fig. 5 is not drawing any hyperplane and plots the same over the non-linear random spaces with the aid of constructed arbitrary shape. The RBF plotting in Fig. 6 also plots similar users as non-linear distribution, but the plot tries to group similar users within the closed boundary. The non-linear plotting with sigmoid kernel depicted in Fig. 7. The sigmoid kernel is coming from neural networks and the SVM classifier uses the kernel by grouping points nearer to the drawn curve boundary. Both linear and non-linear plotting with polynomial kernels are plotting similar users with variant possible distances, whereas the non-linear plotting with RBF plots similar users with a fixed distance.

There are fewer chances of having linearly separable data arrived from the regular activities of users on Social Networks and under such circumstances, the linear kernel is to be selected for faster classification. Otherwise, the non-linear kernels are selected for classifying the randomly distributed points. Both rbf and sigmoid kernels are applied frequently on such data. The sigmoid kernel uses a number of parameters for effective tuning and in case any parameter is lagging its performance, there are chances for the deviation of the expected results. The RBF kernel draws a boundary curve containing the points, which are having the maximum chance of grouping with the tight relationship. Moreover, the RBF is having a little parameter for better tuning of results.

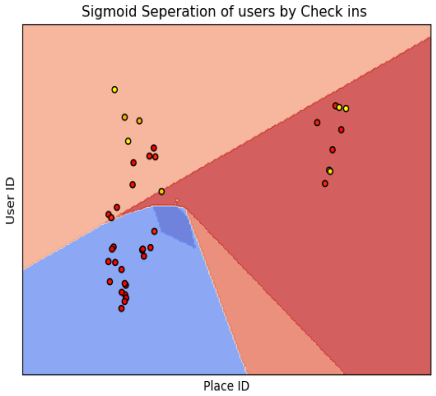

Fig. 7. Non-linear circular plotting for grouping similar users.

Samples from different observations and with all kernel modes are tabulated below.

TABLE II: SHOWING THE ACCURACY OF SVM CLASSIFICATION WITH DIFFERENT SAMPLES

S.No	Samples	Accuracy of Training Data	Accuracy of Test Data	Accuracy of user grouping	Error Rate
1.	9,00,000 Users, 25 Places, Linear *C*=1.00	1.00	1.00	99.99	.0001
2.	90,000 Users 50 Places, Linear *C*=1.00	0.96	0.97	96.72	.036
3.	90,000 Users 25 Places Non- Linear Poly *C*=1.00	1.00	1.00	99.97	.0001
4	90,000 Users 25 Places Non- Linear RBF *C*=10.00	1.00	1.00	99.99	.0001
5	50,000 Users 100 random states Non- Linear Curve *C*=10.00	1.00	1.00	99.98	.0001
6	50,000 Users 100 random states Non- Linear Circle *C*=10.00	0.89	0.89	89.04	0.11
7	50,000 Users 200 random states Non- Linear Circle *C*=10.00	0.99	0.99	99.37	0.006

TABLE III: SHOWING THE PRECISION, RECALL AND F1-SCORE OF PREDICTED 25 CLASSES OF PLACES WITH 90,000 USERS

Class	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	906
1	1.00	1.00	1.00	922
2	1.00	1.00	1.00	891
3	1.00	1.00	1.00	936
4	1.00	1.00	1.00	895
5	1.00	0.98	1.00	868
6	1.00	1.00	1.00	909
7	1.00	1.00	1.00	952
8	1.00	1.00	1.00	885
9	0.99	1.00	0.98	893
10	1.00	1.00	1.00	907
11	0.97	1.00	1.00	921
12	1.00	1.00	1.00	925
13	1.00	1.00	1.00	916
14	1.00	1.00	1.00	883
15	1.00	1.00	1.00	868
16	1.00	1.00	1.00	925
17	0.98	1.00	0.99	928
18	1.00	1.00	1.00	916
19	1.00	1.00	1.00	887
20	1.00	1.00	1.00	867
21	1.00	1.00	1.00	923
22	1.00	1.00	1.00	862
23	1.00	1.00	1.00	848
24	1.00	1.00	1.00	880
Avg/Cls	1.00	1.00	1.00	22,500

The following table shows the precision, recall, F1-Score, and support values of the SVM classifier with 25 classes.

The precision column in the above table shows the accuracy ratio of prediction on each class. The value 1.00 indicates that the SVM classifier correctly predicted the labels with their actual role. The recall is the ability of the classifier to find all positive instances. The value on the Recall column indicates that the classifier correctly found the maximum occurrence of positive instances. The F1-Score indicates the harmonic mean and the value 1.00 on the F1-Score column indicates that the SVM classifier is having the best classification score. The last column indicates the number of supporting samples. These observations indicate that the SVM classifier perfectly classified 99% of similar users by considering the check-ins.

From Table II, the SVM classifier effectively classifies the similar users of a place together with a high rate of accuracy. The C value is the confidence constant and has the range of values from 1 (Minimum) to 1000 (Maximum) which would ultimately increase the margin range of classification. In this work, the C value is taken as 1, 10, and 100. The minimum value of C increases the maximum margin which would also decrease the distance of the hyperplane and gamma value (used for adjusting the margin of hyperplane), outputs a high rate of grouping with minimum misclassification of data (error rate), and observed as 89% to 99.9%, with the respective error rate of 0.0001

and 0.006. The high rate of classification of users is determined from a non-linear fit than the linear fit with an average of 99.9%. Moreover, the RBF plotting fits similar users faster than the other classification modes (either linear or non-linear) with a high rate of accuracy and without any hyperplane from any originated point. The RBF plotting also tries to tightly bound the closer points and produces faster predictions with high rate.

The trade-off between the true positive and false positive rate of randomly chosen 50,000 users with 20 places is plotted as the Receiver Operating Characteristic (ROC) curve and is shown below.

The dotted line in the ROC curve shows that the SVM classification over the check-ins data classifies the users with 95% of accuracy.

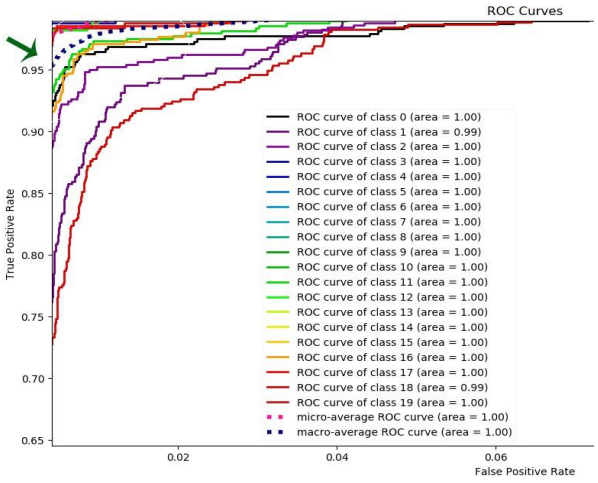

Fig. 9. ROC plotting for 50,000 users and 20 places classification.

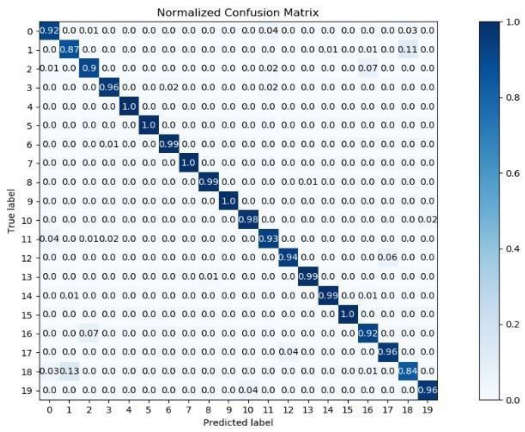

Fig. 10. Confusion matrix showing the true positive and false-positive rates for 50,000 users with 20 places.

The ROC curve plotted in Fig. 9 shows that the True Positive Rate (TP), also known as sensitivity of the prediction, is 95% at maximum, yielding the result that the SVM classifier exactly predicts and groups the users based on location. The Confusion Matrix in Fig. 10 for the same sample also shows that the TP rate is 95% at maximum.

The corresponding confusion matrix for the same sample is depicted Fig. 10 and describing the performance of the SVM over the 50,000 users with 20 places.

Each row in the confusion matrix is the instances of the respective predicted class (check-ins place) and each column represents the instances in the actual class.

The diagonal elements of the confusion matrix in Fig.10 show the occurrence of true positive elements as expected within the 20 places. The observed values are falling in between 0.84 to 1.0, indicating that the classifier predicted the result with a high rate of accuracy.

From all the observations, the SVM classifier used in this work effectively classified the given data into groups in both linear and non-linear modes. The field of social network has a huge amount of rough data so it can't be in a linearly distributed form. Under such circumstances, classifiers like SVM with non-linear mode kernel can be used for classification. In this work, the non-linear fitting of data with the RBF works faster than the linear distribution. The standard running time of SVM is O (n^3), which depends on the total number of training test data and the total number of support vectors to be formed out. After SVM classification, the status homophily is also checked out as a minor task for further classification, for the remotely placed points from the support vector. So, the overall work would be effective with SVM classification with optimized data and optimized support parameters.

VII. Conclusions

The present work classified similar users based on the user check-ins by using SVM classifier and status homophily. The SVM classifier itself classified the similar users into the respective group with a high rate of accuracy (ranging from 89% to 99%). The closely placed user points on SVM are grouped nearer to support vectors and the remotely placed points are grouped with the aid of status homophily. In this work, the SVM classifier predicted and grouped most of the similar users belonging to a place at maximum, and the status homophily is taken at a certain extent for remotely placed points. In case of adjusting the tuning parameters of kernel parameter, the best-selected values of C, gamma and alpha, avoid the usage of status homophily.

The time complexity of the present work is also minimal compared to other related works which would produce a low accuracy of similar user groups with high computational cost and tend to be infinite when the data becomes large.

VIII. Future Enhancements

The present work considered optimal data with minimum feature values before classification and produced better results with such optimized inputs. However, the world of social network has rough and huge data, and there are occasions to classify with multiple features and labels. Better optimization of data is performed as a process before applying classification over the data. Optimizing the given input, by using a supportive algorithm or any normalized or scaling method, and applying classification for effective social media mining is beyond the scope of this work. The proposed work only considered the check-ins activity of users as status homophily value for classification. There is a variety of status homophily exist in the form of user's attributes like age, dob and native place, etc. The used data in this work is of the type numeric and the SVM classifier

easily classified by directly applying the data. Most of the user's attributes on the social network are of the type categorical and hence the classification of similar users with such status homophily needs more steps. Combining both numerical and categorical data for better classification is also appreciated.

REFERENCES

[1] F. Yu, Z. J. Li, and S, X. Jian, and S. R. Lin, "Point-of-interest recommendation for location promotion in location based social networks," in *Proc. 2017 IEEE 18th International Conference on Mobile Data Management*, School of Computer Science and Technology.

[2] M. F. Wu, Z. Q. Wang, and H. R. Sun, "Friend recommendation algorithm for online social networks based on location preference," in *Proc. 2016 3rd International Conference on Information Science and Control Engineering*.

[3] S.-M. Chiu, Y.-C. Chen *et al*., "A fast way for finding similar friends in social networks by using neuro-fuzzy networks," in *Proc. the 2016 International Conference on Machine Learning and Cybernetics*, Jeju, South Korea, July 10-13, 2016.

[4] K. H. Lin, Y. T. Cheen, X. Li, Q. F. Wu, and Z. T. Xu, "Friend recommendation algorithm based on location-based social networks," in *Proc. 2016 7th IEEE International Conference on Software Engineering and Service Science*.

[5] C. Yu, B. Y. Xiao, D. Z. Yao, X. F. Ding, and H. Jin, "Using check-in features to partition location for individual users in location-based social network," *Information Fusion*, 2017.

[6] M. Reshma and R. R. Pillai, "Semantic based trust recommendation system for social networks using virtual groups," in *Proc. 2016 International Conference on Next Generation Intelligent Systems*.

[7] Gowalla Dataset. [Online]. Available: http://www.yongliu.org/datasets/

K. Mohan Kumar received his Ph.D. in computer science from Bharathidasan University. Presently, he works as the head and assistant professor in the PG and Research Department of Computer Science at Rajah Serfoji Government College, Thanjavur, Tamil Nadu, India. He published more than 50 research papers in reputed journals. He has 23 years' teaching experience and 18 years' research experience. His main research areas are machine learning, IOT, network security, and big data analytics.

B. Srinivasan is a part- time research scholar. He is doing his Ph.D in the PG and Research Department of Computer Science at Rajah Serfoji Government College, Thanjavur. He received his master and M.Phil degrees from Bharathidasan University and published more than 30 research articles in reputed journals. Presently, he works as an assistant professor in the PG and Research Department of Computer Science, Khadir Mohideen College, Adirampattinam, Tamil Nadu, India. He has 20 years' teaching experience and 13 years' research experience. His main research areas are social media mining, social network analytics, machine learning, data mining, and sentimental mining.