I J C T A, 9(40), 2016, pp. 587-596 ISSN: 0974-5572

Finding Community in Social Network Based on Profile Attributes

Amit Aylani* and Urjita Thakar*

ABSTRACT

Social Networking sites (SNS) such as Facebook, Twitter, and Linkedln etc. are very popular. A SNS allows exchange of different types of information among different users. People like to communicate with the people having common interests, relationship or background. Such kind of interaction between people leads to formation of community in a social network. Community in social network refers to a group of people having common interest. It provides a medium to a group of people linked in several aspects to get connected with each other. Connecting different people onto a same platform is difficult. Therefore, it is difficult to find social relationship between users. Hence, a new approach to find community in social network is proposed in this paper. Current techniques for finding community emphasize more on social activities rather than user attributes. The proposed approach focuses on three profile attributes which are Number of Tags, Current Affiliation and Mutual friends. String similarity function and K-means algorithm is applied along with these attribute values which results to a community for social network. These communities reflect closeness and considerably good interaction among community members.

Keywords: Social network, Social relationship, Community, Profile attributes

1. INTRODUCTION

A social networking site is an Internet based online platform where a user can create a profile and build a personal network that connects to other users. It offers a mechanism to connect, communicate and share information with other users. MySpace, Facebook, Twitter and LinkedIn are examples of widely popular social networks. Every user of social network has a profile which consists of personal information like name, details of education, college name, school name, date of birth, geographical location, home town etc. These are known as profile attributes. Such attributes can be used to match user's profiles and identify similar types of users in social networks i.e. this groups of users contains similar attributes. They may possibly have similar liking and share data often due to matching interests. It can be considered as a set of users where each user interacts more frequently within the group of users than outside the group [1]. These groups are called communities. It is structure which is focused on sharing of different types of information such as messages, photos, videos etc. which play important role in providing personalized services, marketing services and increased privacy from public users.

A various number of methods for find user community have been proposed by earlier researchers. Himel et al.[2] proposed method based on two observations, the degree of interaction between each pair of users can widely vary, termed as strength of ties and the interactions for each pair of users, with mutual friends, which we term the group behavior, play an important role to determine their belongs to the same community. Another approach for community of users based on tagging to individual users has been proposed by Hak-Lae et al. [3]. These methods are able to generate communities based on attributes such as tagging, bookmark, sharing content etc, but do not consider important user attributes mutual friends, Tagging of content which shows interest taken by the user and Current affiliation. Earlier methods of finding community focus more on social activities than social relationship and attributes.

* Department of Computer Engineering, Shri G.S. Institute of Technology and Science, Indore, India, *Emails: amit61177@gmail.com, urjita@rediffmail.com*

A new approach is proposed for finding community in which, profile attributes namely tagging; mutual friends and current affiliation are used. These attributes have been selected since users within the community have higher probability to share similar tags [4]. A seed user is selected, for whom community is being found and data related to this seed user is extracted which includes seed user's friend list and their respective profile attributes value. Using string similarity function and k-means clustering algorithm, communities to which a seed user can belong are suggested. Hence, it results with better communities for a user.

Rest of the paper is organized as follows, In the Section II; brief discussion of related work is given. In Section III, proposed approach is presented. Testing and results are discussed in section IV. The Conclusion and Future Scope of paper in section V.

2. RELATED WORK

In the last few years, researchers have made some important contributions towards detecting community in social networks. Some salient contributions are discussed in this section.

Newman et al. [5] was proposed an algorithm for community detection in social networks. It includes two definitive features: first, they involve iterative removal of edges from the network to split it into communities, the edges removed being identified using one of a number of possible "betweenness" measures, and second, these measures are recalculated after each removal. To compute this metric, the shortest path between each pair of vertices needs to be determined which leads to highly intensive computation. Another method proposed by Liaoruo et al. [6] is known as community kernel detection in which they discover that influential users pay closer attention to those who are more similar to them, which leads to a natural partition into different community kernels. They proposed two efficient algorithms GREEDY and WEBA for finding community kernels in large social networks. GREEDY is based on maximum cardinality search, while WEBA formalizes the problem in an optimization framework. This method only improves the performance over traditional cut-based and conductance-based algorithms, but consider the dynamic behavior of users.

Fadai Ganjaliyev et al.[7] proposed a method for community detection using clustering technique, which works for weighted networks. It maximizes total weight of selected clusters and minimizes similarity between these clusters and ensures, cluster being selected contains at least one object. This approach has a drawback for finding a better community, as data object is restricted to belong to a single cluster or community, to minimize similarity between the clusters. For a real world scenario, data object can belong to one or more clusters as individuals can have same data object belonging to different communities suggested for them accordingly.

Xutao et al.[8] proposed community discovery scheme in multi-dimensional network. A group of users is found based on attributes such as tag, photo, comment and stories. This algorithm calculates the probability of visiting contents or pages and compares their values to generate communities. Community detected using probability of interests calculated using contents and pages visited gives only a rough idea of any user's interest and willingness to join any community as interests change over a period of time and also according to current trends.

Feyza et al.[9] proposed an optimization algorithm which helps in analyzing community structure. It also discussed overlapping communities along with community structure. Node based and link based overlapping communities or social groups are defined with an optimized community structure. These overlapping communities are created without using any specific rule set defined for any user belonging to multiple communities which may lead to unnecessary overlaps among communities and decrease efficiency of finding community.

Vasavi et al.[10] proposed measurement of similarity between users on the basis of social, geographic, educational, professional attributes, shared interests, pages liked, common interested groups or communities

and Mutual friends. These attributes were manually assigned weights then string and semantic similarity metrics were used to predict the most similar profiles. Manual assignment of weights and consideration of too much profile attributes from which some are of least importance leads to a very less accurate community detection mechanism. There are various social network properties which are used to find members who interact with each other and have strong social relationship within a community [11]. These social networking properties can be mapped to graph properties as which are as follows.

Vasiliy et.al [12] proposed an algorithm for node grouping in social network by analysis of the influence of nodes in social network. This analysis of any social network is done using page rank analysis. This page rank analysis results in detecting communities similar to communities' detected using algorithm which were based on common interest of users or according to pages most visited but these interests may change in due course of time. Communities formed using such approach lack in user activeness and interest after some time as preferences changes.

Clustering Coefficient: The clustering coefficient quantifies how well connected are the neighbors of a vertex in a graph.

$$C_c = \frac{Average\ Degree\ of\ Node}{Number\ of\ Nodes - 1}$$

If a cluster has a value of clustering coefficient equal to 1, then it indicates an efficient cluster i.e. strong interconnection. Therefore, in a social network it represents degree of closeness in community member.

Density: The proportion of direct ties in a network relative to the total number possible.

$$Density = \frac{Actual\ Connection}{Potential\ Connection} \tag{2}$$

It signifies strength of bond between users as it shows one to one connectivity among members in a community of a social network.

Clique: Clique represents a group of people who interact with each other regularly and intensely than others in the same setting. When every node connects to every other node then it is a clique. It represents similar social characteristics between users in a community.

Many researchers have proposed various methods to detect communities which are based on user activities such as tagging, Mutual friends, pages liked, comments, seed items etc. but they have limitation in finding strong social relationship between users. Existing methods give more importance to attributes which emphasize more on social activities rather than their profile attributes. In the next section a new approach proposed for find community in social network based on profile attribute is discussed.

3. PROPOSED APPROACH

In a proposed approach, to match user profiles on online social network, a large and suitable dataset is required. Facebook social network has been considered in this work and data from the Facebook social network is extracted using Facebook API or graph API. For detecting community, a seed profile is considered for whom community is being detected. The data related to seed user is extracted which includes friend list and values for attributes Mutual Friends, Number of Tags and Current Affiliation. By using string similarity function attribute value of current affiliation is used to match their friend list users attribute and obtain some group of users from complete friend list. This similar user is required to more analyze. With the help of remaining two attribute mutual friend and Number of tag identified different community based on their interaction and social relation. In this approach focus is on three main profile attributes of user which are explained as follows.

Current Affiliation provides details of user's current professional affiliation (college name or company name).

Mutual Friends are the common friends between seed user and friends in the friend list.

Tagging is a (relevant) keyword or term associated with or assigned to a piece of information (like picture, article, or video clip) to the social users. Tagged value is count of the number of tags between two friends. The proposed approach is described in various steps which have given below.

3.1. Identify Users with Same Current Affiliation

In the first step, list of users who have same current affiliation value corresponding to seed user are identified. This attribute help in finding users with possible same work place or affiliation to same educational institute. A Seed profile attribute value (i.e. Current Affiliation) is matched with other user's profile attribute, who are in friend list of the seed profile. Attribute values are in string format, so to match this value string matching function is used. A similarity score is assigned to every profile after matching function is applied. If string matches then matched user score is assigned a value 1 else 0. Similarity score is assigned as follows.

$$S(x,y)=\begin{cases}1 \; if \; x.\,\varphi = y.\,\varphi & y \in Y \\ 0 \; if \; x.\,\varphi \neq y.\,\varphi & y \in Y\end{cases}$$

Where, S is the string similarity function.

Y = Set of users present in the friend list.

x = is the seed profile.

y = Friend list user profile

φ = is the current affiliation string attribute value

Using the similarity score, users whose score is equal to 1 are extracted from the friend list of the seed profile. This similarity score helps in finding profiles that have same work place or are affiliated to same educational institute.

3.2. Separation of Similar Users

The list of users obtained in previous step is those who have same affiliation but it does not confirm that they have strong social relationship and should belong to same community. To find social relationship among these users, more attributes are used based on which their social relationship or interaction can be identified. Two important attributes pertaining to identification of social relations are Number of Tags and Mutual Friends. These attribute values are extracted for every user in the friend list corresponding to the seed user.

3.3. Apply Attribute Normalization Technique

For all the users present in the friend list, the two attribute values are obtained. There is large variation in the values of these attributes for different users and hence these values need to be normalized i.e. need to be brought in a smaller range. For this normalization is done for the values of the two attributes as given below.

Let N: Number of users with same Current Affiliation

M_i: Set of attribute values of Mutual Friends

T_i: Set of attribute values of Number of Tags

$\bar{X}_m$: Mean of attribute value for Mutual Friends M_i.

$\bar{X}_t$: Mean of attribute value for Number of Tags T_i.

Step (i): Calculate Mean ($\bar{X}m$) for

$$\bar{X}m = \sum \frac{Mi}{N} \tag{3}$$

Step (ii): Calculate Mean (for

$$\bar{X}t = \sum \frac{Ti}{N} \tag{4}$$

Step (iii): if $\left(\bar{X}m < \frac{\bar{X}t}{2} \right)$ Goto Step (vi)

Step (iv): if $\left(\frac{\bar{x}m}{2} > \bar{X}t \right)$ Goto Step (vii)

Step (v): Goto Step (viii)

Step (vi): For i=1 to N

$$Mi = Mi + Xm \tag{5}$$

Step (vii): For i = 1 to N

$$Ti = Ti + Xt \tag{6}$$

Step (viii): Exit

The attribute values of Number of Tags and Mutual Friends are now in a smaller range for all the users with same Current Affiliation. For detection of community, a clustering method is applied as given next section.

3.4. Community Detection Algorithm

Considering the output of previous step, the two new attribute values in data set is consider as two variables on each individuals. This data set is to be grouped into three clusters by applying k-means clustering technique. It is used to minimize the average squared Euclidean distance of nodes from its cluster center. Hence, K-means clustering technique generates cluster with high intra-class similarity and a low inter-class similarity.

K-means clustering algorithm [13] is follows.

Input: $D = d_1, d_2 ..... d_n$

D is n data items set.

k = Number of desired clusters

Output: A set of k clusters.

Step: 1. Select k data-items from set D as initial centroid.

Step: 2. Repeat Assign each item i to the cluster which has the closest centroid.

Step: 3. Calculate new mean for each cluster; until convergence criteria is met.

By applying k-means cluster algorithms the data set values divided into three clusters. These clusters have nearest individuals as well as similar social relationships among these users, within a cluster are tightly coupled which is indicating similarity compare to others, this cluster is known as community. It is used to share information, photos and videos among community members. The obtained testing and results performed are discussed in the next section.

4. EXPERIMENT AND RESULT

The data is extracted from Facebook using graph API or FQL query. The data consist of friend list users corresponding of seed user, contain 131 users has various attribute like name, user id and current affiliation which is extracted from social network. Remaining two attribute value i.e. Number of Tags and Mutual Friends are extracted from the individual connection between friend list user and seed user. The seed user is randomly selected for the experimental purpose and their detail is as given below in Table-1.

Table 1
Seed User Detail

User Id	*Current Affiliation*	*Number of Friends*
201	SGSITS	131

Table-2 shows sample data corresponding to the seed user which includes User Id, Current Affiliation, Mutual Friends and Number of Tags.

Table 2
Users Data Corresponding to Seed User

User Id	*Current Affiliation*	*Mutual Friends*	*Number of Tags*
101	SGSITS	89	45
102	SVIT	38	35
103	LNCT	116	15
104	SGSITS	51	52
105	MANIPAL	4	1
106	SGSITS	128	2
107	SGSITS	33	35

In the next step, users having same affiliation have been identified by applying string similarity function. The result is as shown in Table-3 for the sample data.

Table 3
Users Data with Same Affiliation

User Id	*Current Affiliation*	*Mutual Friends*	*Number of Tags*	*Match Result*
101	SGSITS	89	45	1
104	SGSITS	51	52	1
106	SGSITS	128	2	1
107	SGSITS	33	35	1

In the next step, experimental data set contain 48 users remaining with same affiliation match with seed user. The remaining two attributes is used for next step. Table-4 shows number of Mutual friends and Number of Tags and the mean of both the attributes are calculated.

Table 4
Mean Values Calculated for the Two Attributes

User Id	*Current Affiliation*	*Mutual Friends*	*Number of Tags*
101	SGSITS	89	45
104	SGSITS	51	52
106	SGSITS	128	2
107	SGSITS	33	35
	Mean	$\bar{X}a$ =75.25	$\bar{X}b$ =33.5

As seen in the Table-4, the difference is more than double in the mean values of the two attributes, so normalization needs to be done. Result of Normalization is reduced variation in the attribute values which makes clustering more efficient. The result for sample data is shown in Table-5.

Table 5
Users with Updated Attribute Value

User Id	*Current Affiliation*	*Mutual Friends*	*Number of Tags*
101	SGSITS	89	78.5
104	SGSITS	51	85.5
106	SGSITS	128	35.5
107	SGSITS	33	68.5
	Mean	$\bar{X}a$ =75.25	$\bar{X}m$ =67

In the Table-5, the new updated values as well as their mean value on same data is given. In this proposed work, mean value of two attributes for actual data is $\bar{X}m$ = 8.45 and $\bar{X}t$ = 24.52. Ratio of these two values is more than 1:2, hence normalization must be applied and calculation of new values is done. The new values are updated to $\bar{X}m = 16.91$ and $\bar{X}t$ = 24.52, thus reducing the difference.

On applying the K-means clustering on the new updated attribute values considered as variable have 48 individuals and are divided into 3 clusters, which represent three communities. The result of k-means clustering algorithm is shows three community which is shown in the graph given below as Fig. 1.

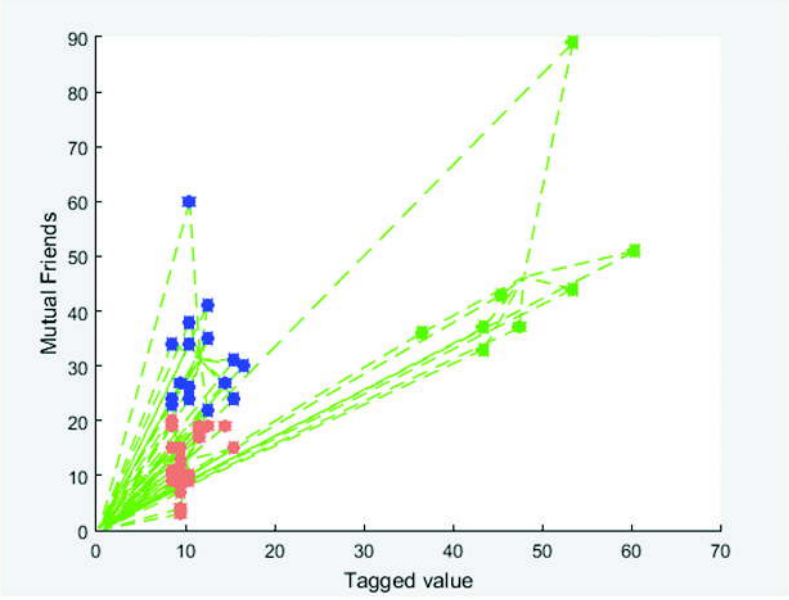

Figure 1: Result of K-means Clustering Algorithm

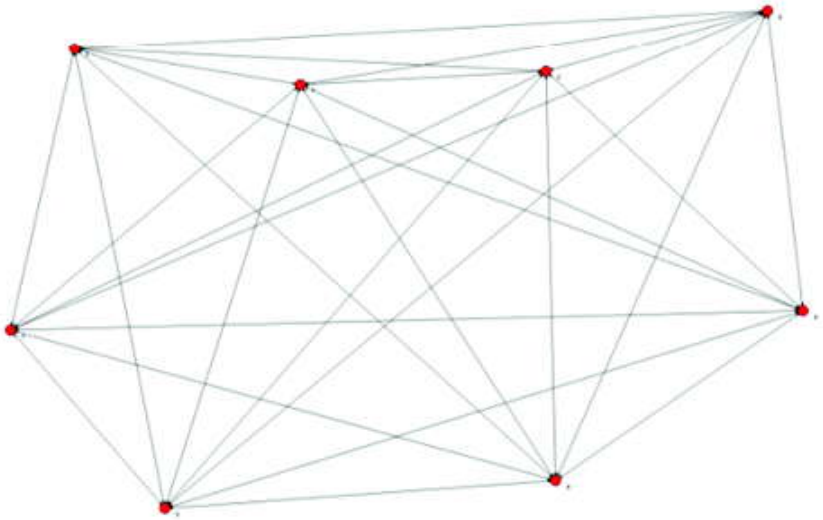

Figure 2: Social Connections among Community Members-1

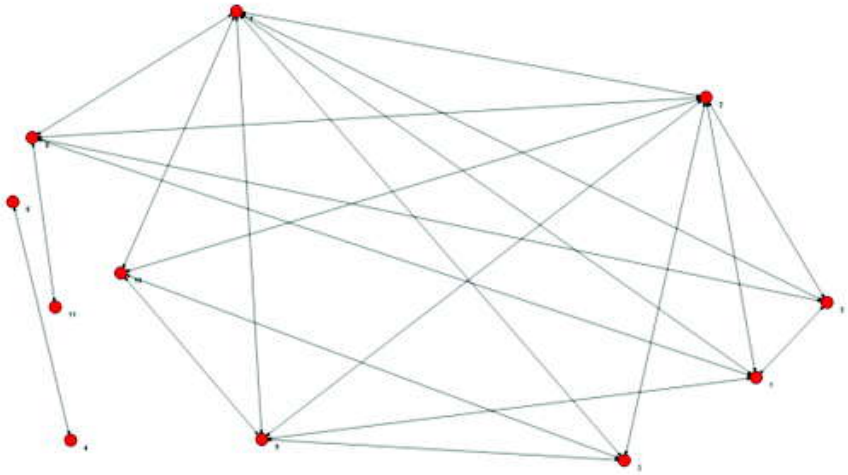

Figure 3: Social Connections among Community Members-2

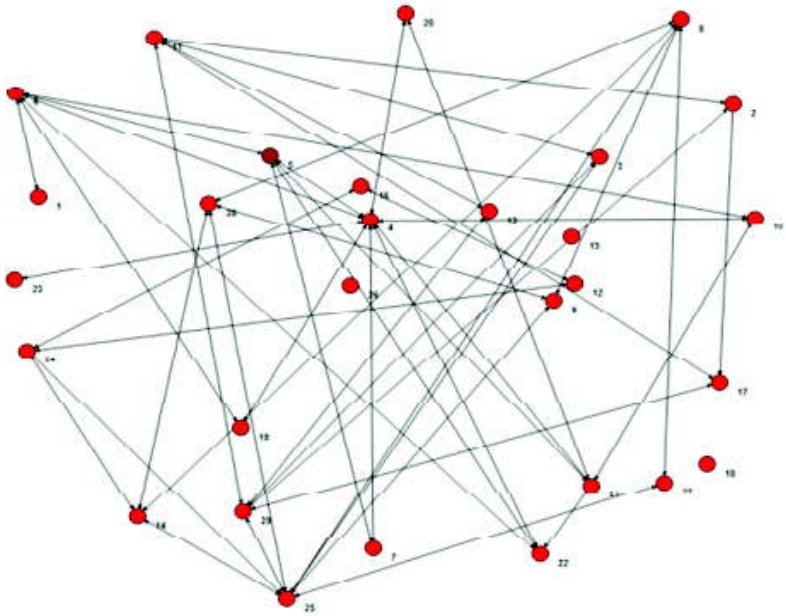

Figure 4: Social Connections among Community Members-3

In this cluster, if nodes are closer and the member closest or nearest, is the community member having strong social interaction with the seed user. In Fig. 1, it contains 8 users in one community, second community has 11 users and the last contains 29 users. These three communities result require more analysis on their interaction to identify their social connections. For this social connection, an adjacency matrix of community member is created, which helps to identify people who interact with others more regularly and actively. This matrix is imported in social network visualizer tool and following results are obtained.

Community connection graph is given in Fig. 2 for community 1 containing 8 users, it have highest density and is a clique. Graph for community 2 is shown in Fig. 3 which contains 11 users and have less density as compare to community 1. Fig.4 shows graph for community 3 which contains 29 users and it have least density. By using social network visualizer tool, various graphical result like density, clustering coefficient, clique, degree are determined which are used to perform comparison between these community results, which helps in identifying suitable community.

Table 6
Comparisons of Different Communities

	Community-1	*Community-2*	*Community-3*
Nodes	8	11	29
Edges	28	38	44
Average degree	7	3.72	3.83
Density	1	0.4	0.11
Average Clustering Coefficient	1	0.585	0.587

Table-6 shows comparison of various graph properties for each community. It is observed that all the users of Community-1 are connected to each other strongly, while the closeness is lesser in Community-2 and is least among the users of Community-3. It shows the interaction between community members is more as compared to other users. This results in reduced traffic on social network.

5. CONCLUSION

In this paper, an important issue of community detection in social networks has been discussed. A profile attribute based method has been proposed to find communities more accurately. Three user attributes have been considered, that may help in more precisely determining the interests of the user thereby indicating the people who may belong to same community. It has been observed that the method is able to find community of users who have stronger social relationship between community members and share data more actively.

In future some semantic matching on attribute value can be performed which may provide their area of interest more precisely.

REFERENCES

[1] Jaewon Yang, Julian McAuley, Jure Leskovec, "Community Detection in Networks with Node Attributes," in IEEE 13th International Conference on Data Mining, pp.1151-1156, Dec. 2013.

[2] Himel Dev, Mohammed Eunus Ali, and Tanzima Hashem, "User Interaction Based Community Detection in Online Social Networks", in Springer International Publication on Database Systems for Advanced Application, vol. 8422, pp. 296-310, April 2014.

[3] Hak-Lae Kim, John G. Breslin, Stefan Decker, and Hong-Gee Kim, "Mining and Representing User Interests: The Case of Tagging Practices", in IEEE transactions on systems, man and cybernetics-part A: systems and humans, vol. 41, no. 4, pp.683-692, July 2011.

[4] X.Wang, L.Tang, H.Gao and H.Liu, "Discovering Overlapping Groups in Social Media" in IEEE International Conference on Data Mining, pp. 569-578, Dec. 2010.

[5] M. E. J. Newman and M. Girvan, "Finding and Evaluating Community Structure in Networks", in Physical Review E, 69:026113, vol. 69, Feb. 2004.

[6] Liaoruo Wang, Tiancheng Lou, Jie Tang and John E. Hopcrof, "Detecting Community Kernels in Large Social Networks" in IEEE 11th International Conference on Data Mining, pp. 784-793, Dec. 2011.

[7] Xutao Li, Michael K. Ng, and Yunming Ye, "MultiComm: Finding Community Structure in Multi-Dimensional Networks" in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 4, pp. 929-941, April 2014.

[8] Ganjaliyev. F, "New Method for Community Detection in Social Networks Extracted from the Web", in Problems of Cybernetics and Informatics (PCI), IV International Conference IEEE, Sep. 2012.

[9] Feyza Altunbey and Bilal Alatas, "Overlapping Community Detection in Social Networks Using Parliamentary Optimization Algorithm" in International Journal of Computer Networks and Applications, vol. 2, Feb. 2015.

[10] Vasavi Akhila Dabeeru, "User Profile Relationships Using String Similarity Metrics in Social Networks" in Social and Information Networks (cs.SI), Aug. 2014.

[11] Elie Raad, Richard Chbeir, Richard Chbeir "User Profile Matching in Social Networks" in 13th International Conference on Network-Based Information Systems, pp. 297-304, Sep. 2010.

[12] Vasiliy A. Perepelitsyn, Alla G. Kravets "The social networks' nodes grouping algorithm for the analysis of implicit communities" in 7th International Conference on Information, Intelligence, Systems & Applications (IISA) IEEE, July. 2016.

[13] Mr. Pushpendra Bhatt, Prof. Tidake Bharat, "A review paper on improved k-means technique for outliear detection in high dimensional dataset" in International journal of engineering science and research technology, pp. 235-239, Jan. 2015.