Implementing CURE to Address Scalability Issue in Social Media

U.Vignesh[1,] S.Mohamed Yusuff[2,] M.Prabakaran[3]

[1,2]Mookambigai College of Engineering/Anna University,
[3]RVS College of Engineering and Technology/Anna University

ABSTRACT:

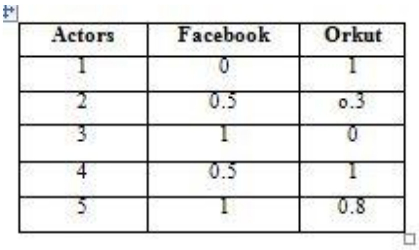
This paper presents a behavior of individuals from an collective behavior using a social-dimension based approach. Collective behavior, which indicates the group of data generated on a large scale, such as facebook, orkut, twitter etc. which generates a large amount of data on large scale and provides a way to act as a platform for targeting an actor's behavior involved on to them. Thus, we consider an place of social media network to choose an application for applying our work on to them. Social media network activities which includes millions of actors behaviors involved onto it in fraction of second. To predict these behaviors, we propose an edge centric clustering schema to extract social dimensions from a network. Although, there are many extracting schemes to extract social dimensions, their scalability paves the way to lack to guarantee in finding dimensions effectively and efficiently. Edge centric clustering schema with CURE algorithm solves the scalability issue that occurs in previous schemas and provides efficient and effective extraction of social dimensions. If the process implementation completed, the performance results using social media network shows that our cost included in CURE with edge centric scheme is being less than one third the cost required by existing schemas. The effectiveness of edge centric clustering schema is demonstrated by experiments conducted on various data and by comparison with existing clustering methods.

Keywords: *actors, collective behavior, CURE, scalability*

I. INTRODUCTION

The rapid advance of social media actors and the communication involved to them leads to a traffic control aspect and also a difficulty in tracing a particular actor involved in an social media and on background, the uninvolved actors to the target can also to be easily identified by using our latest techniques to be processed on once to it. Here we consider a facebook, orkut as shown in Table 1 as an example for social media networks and we works on over through to it to achieve the expected target to be traced on the network by calculating the affiliation of the actors with the weights that we have achieved on manual analysis through the way of analysis algorithm CURE to be implemented on the network. There are millions of actors and their behaviors are running and updating on a social media network (facebook, orkut) for a fraction of second. With the variation of this fraction of second, we have to trace a target behavior and their behavioral relationship with the unobserved individuals in the network.

Table 1. Affiliations identification



Actors	Facebook	Orkut
1	0	1
2	0.5	o.3
3	1	0
4	0.5	1
5	1	0.8

Table 1 shows the 5 types of actors that they have involved in a social media network such as facebook, orkut. The actor 1 has been only affiliated to orkut with the weighting factor of 1 whereas, the actor 2 affiliated to both facebook and orkut with the weighting factor of 0.5 and 0.3. the actor 3 has been only affiliated to facebook with 1. The actor 4 and actor 5 affiliated to both facebook and orkut with the weighting factor (0.5, 1) and (1, 0.8). Thus, through this analysis based upon affiliation give clear overview for the behavior of observed and unobserved individuals in a network.

In consideration with the social media network, the various parameters have to be taken for the process of analyzing. Clustering plays a major role in the phenomenon due to the actor analysis process based on their community involved in the network or its communication with its friends or colleagues on the running time in media and these updating have been simultaneously going through in the concerned media. As with millions of updating, we are going to search through the particular actors behavior in the network and their perceptions can also be overviewed, there we come across an most regarded features such as observed individuals and unobserved individuals. The term unobserved individuals which indicates the actors behavior concerned to them but they are not belong to same community or network and they may belong to some other network and these individuals are not being our target, we collect the behavior of individuals due to an target actors communication with these individuals on an notification process. Thus, the clustering concept which divides the common community or common network people into one cluster and the other network people, who does not belongs to target actor network are grouped into some other cluster. Thus, the classification does its work followed by clustering in the network.

Clustering Using Representatives (CURE), which involves its action with the constant number of representative points have been selected to represent or to form a cluster. The selected number of points is to play a role of parameters. The similarities are to be identified as a affiliation by the similarity of an neighbor pair of selected points that represents to various kinds of clusters. Fig. 1 shows the network of 9 actors to be included onto it. The targeted actor name considered to be V and how V performs affiliations and its behavior are to be traced based on the edge centric clustering schema. The edge partition deals with the status 1 and 0 for its behavioral connections simultaneously to his friends, family, school mates, college mates and office colleagues.

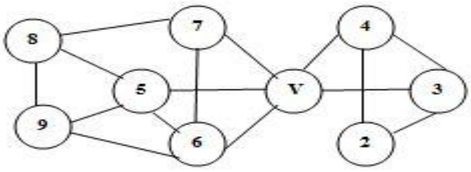
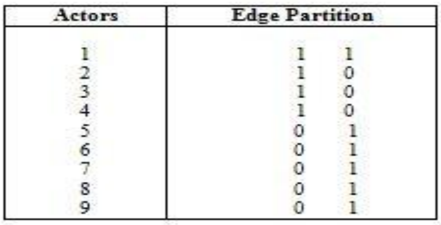


Figure 1. An actor V example

Table 2. Dimensional representation of the V example



Actors	Edge Partition	
1	1	1
2	1	0
3	1	0
4	1	0
5	0	1
6	0	1
7	0	1
8	0	1
9	0	1

Table 2 gives the clear overview of an edge partition pattern for the 9 actors involved in a media network. The target V is to be affiliated to both facebook and orkut as analysed by the edge partition with the status 1, whereas other actors plays its behavior to be either facebook or orkut with the status either (1,0) or (0,1) in the edge partition. If the actors are to be in count of millions and their related observations or communications with the other actors or unobserved individuals can easily be extracted by using the edge partition notification corresponding to the communication involved on to them on social media.

II. RELATED WORK

The work related to finding the behavior of individuals with the unobserved individuals behavior in social media of same network are to be followed.

J.Hopcraft and R.Tarjan [1] have introduced the biconnectivity components for a care of edges to be involved. Bicomponents are similar to edge cluster, it also at first separates the concerned edges into disjoint sets for extraction. Lei Tang, Xufei Wang and Huan Lie [2] have proposed the k-means clustering analysis to the edge centric view to find a similarity with the same network or with the different clusters. K-means which deals with the k class labels with the centroid of cluster, then computes $\mathrm{Sim}(i,c_j)$, then checks the condition

$$\mathrm{Sim}(i,c_j) > \mathrm{maxSim}_i \qquad (1)$$

If it possible, then update the action of centroid until the value of objective have been changed over the network. M.McPharson, L.Smith-Lovin and J.M.Cook [3] deals with the concept of homophily, which includes the motivation of link with one another rather than test results in correlations between connections involved. Homophily are very much interested to connect to other in the network with the similarity of an actor. Homophily the term is very popular in the online systems extraction.K.Yu, S.Yu and V.Tresp [6] deals with the probabilistic method because for an aspect of finding the actor behavior with relation to same or different network. R.-E. Fan and C.-J. Lin [9] considers a concept of threshold identification for the aspect of finding similarities that have been tied up. T.Evan and R.Lambiotte [7] proposed a graph related algorithms which constructs the line graph or the basic graph partitioning representation, then they have been followed by a graph partitioning algorithms instead of an edge partition in an clusters but it fails in an mark up scalable liability aspect.M.Newman [4], [5] proposed a various kinds of techniques for an aspect of best evaluation report. He describes a maximum likelihood estimation aspect with the power law degree distribution done on it. Another kind of technique he proposed was modularity maximizatioin, which extracts social dimension with the relational learning methods. Modmax tends to capture the affiliations of the actors on the social network connectivity with their regarding dimensions to be notifies and represented for extraction. S.Gregory [10] tries in extending Newman-Girvan methods for an aspect of solving the overlapping communities in annetwok. F.Harary et al. [13] noticed a line graph for an analysis aspect.X.Zhu et al. [11], [12] deals with the semi supervised learning and label propagation for a detection of individuals behavior in a social-media network with the relevant information. Semi supervised learning which has its differential aspect when compared it with the relational learning and proves its efficiency towards the relational aspect in the process of action involved in the extraction aspect. S.A.Macskassy and F.Provost proposes a collective inference method to solve the problem of identifying a actor in an network but achievement in the scalability are not considered to be up to level.

III. OUR PROPOSAL

3.1. Architecture

To achieve the scalability factor, we design a system with the example of social media on targeting some actor at first, with the relevant information to be given. Thus, the 2 media of facebook and orkut, which generates oceans of data in a second corresponding to the actor with the observed individuals, the unobserved individuals are to be traced out either in the same network or in the some other media network.

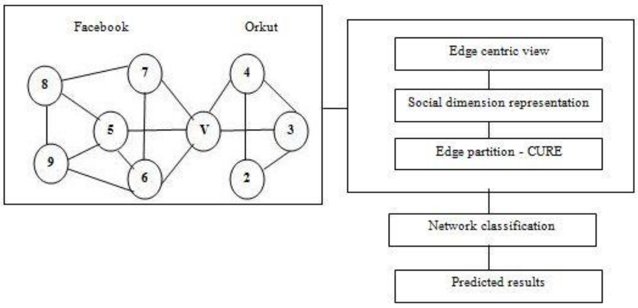


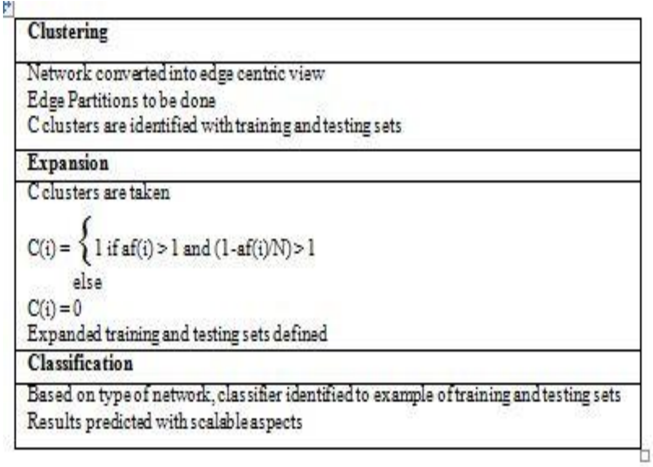
Figure 2. System architecture

Fig. 2 shows the system architecture to collect a behavior of an actor with the scalable manner. First the system starts with 9 actors involved in the facebook and orkut with 5 actors in facebook and 4 actors in orkut, whereas V actor target is communicating with the actors in the same network and also actors in the orkut. To find these coinciding factors, we undergo a network or convert a network into an edge centric view. The edge centric view which identifies edge of an network corresponding to target. Then, the social-dimension representations are to be noted with the selected number of points that we takes. As last, in the schematic factor edge partition are to be calculated with the analysis algorithm of CURE. It has been represented with the status 1 and 0, 1 indicates the communication to be positive, failure are represented as 0. Then the network classification to train a classifier with the expanding dataset has to be done based on selection of classifier corresponding to the defined network. As followed by classification predicted results has been noted in the scalable manner.

3.2. Edge Centric Clustering Scheme by CURE

CURE algorithm, which has its constant number of representative points to represent a cluster rather than a single centroid. With the edge centric clustering scheme it performs the action of extracting a target behavior with the scalability aspect applicable to a forming cluster. CURE algorithm forms a C cluster basis from the group then with C clusters it calculates the analysis factors for the extraction purpose of an target actor given as input to the extractor to extract the collective behavior from the network connectivity considered such as here we takes thefacebook and orkut as an example for the actor behavior to extract.

Table 3: The proposed schema



Clustering
Network converted into edge centric view Edge Partitions to be done C clusters are identified with training and testing sets
Expansion
C clusters are taken $C(i) = \begin{cases} 1 \text{ if } af(i) > 1 \text{ and } (1-af(i)/N) > 1 \\ \text{else} \end{cases}$ $C(i) = 0$ Expanded training and testing sets defined
Classification
Based on type of network, classifier identified to example of training and testing sets Results predicted with scalable aspects

We classify the process into 3 categories viz, clustering, expansion and classification. In the aspect of clustering concept, at first the network are to be converted into edge centric view. Then the edge partition has to be calculated with clustering algorithm called CURE. C clusters are identified with the training and testing sets. Whereas expansion, find the metric to describe the training sets and testing sets to an aspect of classification. In classification based on the consideration of network, classification identifies the example of training sets and testing sets. As last, results are predicted for an actor behavior in a social media.

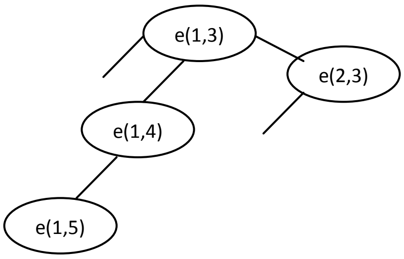
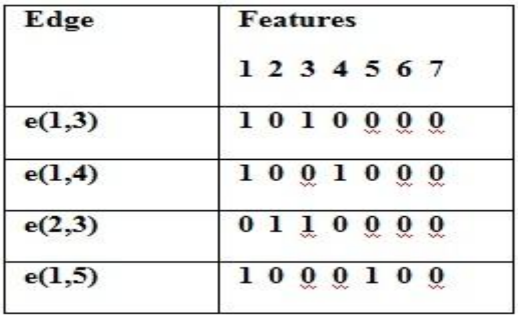


Figure 3. Example for line graph

Edge centric clustering scheme go through the blogcatalog and photosite for the updation of actors or communication of actors to the other actors or people in the social media. Fig. 3 shows how the line graph has been framed from network to find the edges for an aspect of partitioning the actor based on their behavior. Here, the graph paves the way for edge instances to be referred as given in Table 4, which is with the mentioning of status 1 and 0, as discussed earlier their status 1 refers to the positive attribute act and the status 0 refers to the negative aspect in the edges consideration to the social media network are to be useful for CURE algorithm for the purpose of analyzing factor, which analyses the both observed and unobserved individuals in the social medianetwork related to the given input as target to trace on the network.

Table 4: Edge instances



Edge	Features 1 2 3 4 5 6 7
e(1,3)	1 0 1 0 0 0 0
e(1,4)	1 0 0 1 0 0 0
e(2,3)	0 1 1 0 0 0 0
e(1,5)	1 0 0 0 1 0 0

IV. CURE AND CONTROL FLOW

CURE algorithm, which plays a major role in an extraction of individual behavior and in aspect of its collective behavior with scalability. It includes the representation of points that have been selected by using an analysis factor and its major description view are to be identified based on a CURE clustering algorithm. With the CURE clustering algorithm, the partitioned view of analysis are to be classified based on the factor of various view of aspect that we have included for the target consideration factor. The CURE algorithm continuously runs through its selected points in the network for finding out similarities. The advantage of considering CURE clustering is to be better scalability can be achieved compared to other clustering algorithm even when we compare it with ant based clustering.

Fig. 4 shows the control flow of whole system by how which the extraction of collective behavior of an individuals and unobserved individuals are to be done on a social media network.

At first, the extractor has been given with the input of target information with relational learning details, i.e to which social media network the target belongs to. Then with the given input, identified social media network view has been converted into an edge centric view. With the edge centric view, the relational edges are to be identified with the correspondence to the target. Then with this edge centric view, edge partition has been calculated based on the target information. Edge centric schematic representation, which deals with the whole system to be surrounded on towards it including the important parameters, techniques and algorithm to be build on through it for the purpose of extraction basis in the social media network to corner the target.

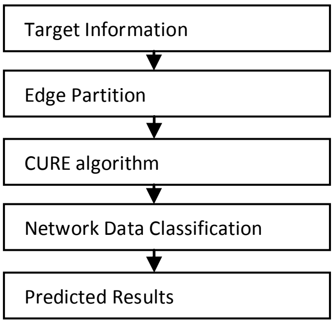


Figure 4. Control flow diagram

By representing the status of an edge partition, the various aspect of forming clusters are to be done based on an clustering algorithm called CURE. CURE analyses the similarities between communities, clusters or network and group them for better aspect of targeting an actor. Then the network data classification has to be done to the edge partitioned data with the selected classifier corresponding to the considered network, such as SVM, SCM etc. and finally, the results to the prediction of given target has been achieved with scalability.

Performance Analysis

The performance metric for the proposed CURE algorithm with edge centric clustering to achieve scalability has been done and compared with the existing clustering algorithms with the various techniques that they have already implemented in different applications for better efficiency and scalability achievement. Fig. 5 shows the scalability performance analysis with CURE clustering algorithm comparison to existing clustering algorithms such as k-means, modularity maximization, fuzzy C-means and K-modes. Whereas, the CURE has achieved the better scalability performance compared to other existing algorithms in the case of a clustering techniques to different types of applicable cases not limited to the application of an social media network connectivity alone.

Table 5 shows the performance of a facebook social media network with respect to the proportion of nodes and detection in micro and macro aspect. The four types of common existing techniques have been tested on through it for an aspect of comparison and to find the better clustering scheme applicable to the social media network to extract the target collective behavior. Edge cluster, which proves its efficiency and effectiveness over the network with high percentage consideration values for the target behavior to trace it over the considered network in social media. The proportion of nodes varies from 1% to 10%. With these proportions, the detection in micro and detection in macro has to be verified corresponding to edge cluster which includes the two sets to notify, the blog catalog and photo site to be viewed simultaneously. Modularity maximization follows the edge cluster actions with the accurate value parameters to be noted. Micro detections are to be compared with the detection in the case of macro. Detection of micro is considered to be better with the better propagation values.

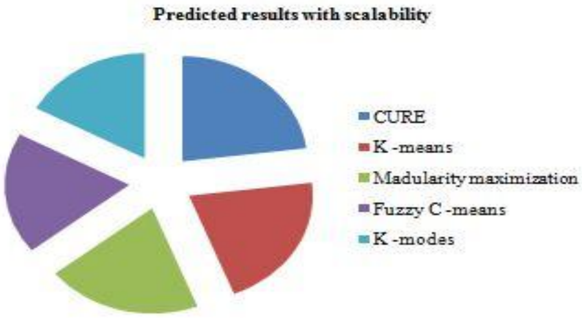
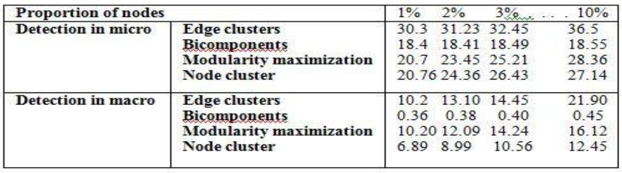


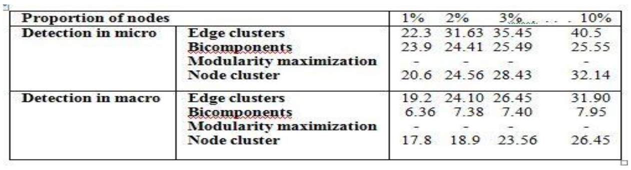
Figure 5. Scalability performance metric

Table 5. Performance on facebook network



Proportion of nodes		1%	2%	3%	10%
Detection in micro	Edge clusters	30.3	31.23	32.45	36.5
	Bicomponents	18.4	18.41	18.49	18.55
	Modularity maximization	20.7	23.45	25.21	28.36
	Node cluster	20.76	24.36	26.43	27.14
Detection in macro	Edge clusters	10.2	13.10	14.45	21.90
	Bicomponents	0.36	0.38	0.40	0.45
	Modularity maximization	10.20	12.09	14.24	16.12
	Node cluster	6.89	8.99	10.56	12.45

Table 6. Performance on orkut network



Proportion of nodes		1%	2%	3%	10%
Detection in micro	Edge clusters	22.3	31.63	35.45	40.5
	Bicomponents	23.9	24.41	25.49	25.55
	Modularity maximization	-	-	-	-
	Node cluster	20.6	24.56	28.43	32.14
Detection in macro	Edge clusters	19.2	24.10	26.45	31.90
	Bicomponents	6.36	7.38	7.40	7.95
	Modularity maximization	-	-	-	-
	Node cluster	17.8	18.9	23.56	26.45

Table 6 shows the clear overview of the performance of social media network orkut. Here as similar to the facebook the metrics are to be calculated with the comparison of existing clustering algorithms such as modmax etc. Here the modmax values are null and its fails to achieve the level to be considered for the formation of partitioning the data for the target behavior in a collective behavior along with the unobserved individuals also to be traced for the extraction.

V. CONCLUSION

This paper has introduced a new way to clustering with the classification that involved in an social media network for the aspect of achieving greater scalability in extraction of behavior of an individuals and also an collection of behavior of an individuals and also the behavioral relation with unobserved individuals with the edge centric clustering schema using CURE clustering algorithm to be applied to the input network. Then, with the edge partitions are to be formed with the affiliations of a target actor are to be identified with the social dimension representation in the concerned network that we are extracting the behavior of an actor. CURE algorithm deals with clustering for the target behavior and gives the result in a group by which the actor relation belongs to the same network or with some other network. Possible extensions and improvements of our model include meta features and different types of clustering algorithms on applications to improve better scalability for collective behavior.

VI. ACKNOWLEDGMENT

The authors would like to thank Dr. C. G. Ravichandran, Principal, RVS College of Engineering and Technology/Anna University, for his valuable advice, support and guidance during the preparation of this paper.

REFERENCES

[1] J. Hopcroft and R. Tarjan, "Algorithm 447: efficient algorithms for graph manipulation," Commun. ACM, vol. 16, no. 6, pp. 372–378, 1973.
[2] Lei Tang, Xuei Wang and Huan Liu, "Scalable Learning of Collective Behavior".
[3] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," Annual Review of Sociology, vol. 27, pp. 415–444, 2001.
[4] M. Newman, "Power laws, Pareto distributions and Zipf's law," Contemporary physics, vol. 46, no. 5, pp. 323–352, 2005.
[5] M. Newman, "Finding community structure in networks using the eigenvectors of matrices," Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), vol. 74, no. 3, 2006.[Online]. Available: http://dx.doi.org/10.1103/PhysRevE.74.036104
[6] K. Yu, S. Yu, and V. Tresp, "Soft clustering on graphs," in NIPS, 2005.
[7] T. Evans and R. Lambiotte, "Line graphs, link partitions, and communities," Physical Review E, vol. 80, no. 1, p.16105, 2009.
[8] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in SDM, 2005.
[9] R.-E. Fan and C.-J.Lin, "A study on threshold selection for-label classication," 2007.
[10] S. Gregory, "An algorithm to find overlapping community structure in networks," in PKDD, 2007, pp. 91–102. [Online]. Available: http://www.cs.bris.ac.uk/Publications/ pub master.jsp?id=2000712
[11] X. Zhu, "Semi-supervised learning literature survey," 2006. [Online]. Available: http://pages.cs.wisc.edu/_jerryzhu/ pub/ssl survey 12 9 2006.pdf
[12] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning usinggaussian fields and harmonic functions," in ICML, 2003.
[13] F. Harary and R. Norman, "Some properties of line digraphs," RendicontidelCircoloMatematico di Palermo, vol. 9, no. 2, pp. 161–168, 1960.

U.Vignesh completed his Master of Technology in Information Technology from VeltechMultitechDr.RangarajanDr.Sakunthala Engineering College/Anna University – Chennai in 2012. His research interests lie in the field of data mining, cloud computing and networking. He is working as an AP/IT in Mookambigai College of Engineering/Anna University – Pudukkottai, Tamilnadu.

S.MohamedYusuff completed his Master of Technology in Computer Science and Engineering from Prist University – Thanjavur in 2012. His research interests lay in the field of data mining, networking and security. He is working as an AP/CSE in Mookambigai College of Engineering/Anna University – Pudukkottai, Tamilnadu.

M.Prabakaran completed his Master of Technology in Information Technology from SNS College of Technology/Anna University – Coimbatore in 2011. His research interests lay in the field of data mining, cloud computing and image processsing. He is working as an AP/CSE in RVS College of Engineering and Technology/Anna University – Dindigul, Tamilnadu.