Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 12, December 2014, pg.90 – 96

RESEARCH ARTICLE

Knowledge Discovery in Social Networking for Future Filtering Purposes

Srikanth Raju T[1], V.Santosh Kumar[2], Ch.Ravindranath Yadav[3]
Student, Department of CSE, Sreyas Institute of Engineering and Technology, Hyderabad, India [1]
Associate Professor, Department of CSE, Sreyas Institute of Engineering and Technology, Hyderabad, India [2]
Asst.Professor, Dept. of CSE, Keshav Memorial Institute of Technology, Hyderabad, India[3]
Tsrikanth91@live.in [1], vennu.santosh@gmail.com [2], ravi_chintala@hotmail.com [3]

Abstract— Social media has become very popular and is able to generate oceans of data which is a goldmine to prospective researches in the real world. Discovering hidden patterns from such data can provide business intelligence. Mining such social media data and learning collective behavior can have very useful utilities in the real world. Many real world applications like advertising, marketing, sales do have their focus on social networking. Scalable learning of collective behavior is very important research area where as the social media data is very vast. Sustainable solution should promote scalability. Object heterogeneity is one of the challenging aspects that can be incorporated in mining tasks so as to gain actionable knowledge. Recently Tang et al. proposed a solution for this using edge-centric approach. In this paper we are influenced by that work and propose a scheme that can solve the problem of object heterogeneity based on the multi-mode network. With this solution the prediction capabilities of the proposed application will be more. We build an application to demonstrate the proof of concept. The application can predict collective behavior with scalable feature and will be useful in real world applications.

Index Terms – Social networking, data mining, social dimensions, collective behavior, clustering

I. INTRODUCTION

Due to the advancements in technologies, virtual communities have been realized as a new phenomenon that empowers people get together online. Plenty of possibilities have been explored with social media including knowledge sharing, business collaborations, obtaining collective intelligence with unprecedented scope and scale and no time and geographical restrictions. Social media provides plethora of opportunities to gain business intelligence by studying human interactions and obtaining collective behavior of people of various walks of life who participate in virtual computing. Social network analysis [1] has become important in many fields such as targeted marketing, intelligent analysis [2],epidemiology [3], and sociology. An important study which has given much importance in social media is to predict collective behavior of some individuals of a group provided the knowledge of some people of the same group [4]. The connections in social media are heterogeneous in nature. For instance an individual can have connections with classmates, colleagues, friends, family members and others. Generally only connectivity information is available. But inferring collective behavior is a challenging job due to heterogeneity of connections. A framework was proposed in [5] for addressing the heterogeneity by exploring novel way of classifying social networks to capture real time affiliations across actors. In [6] modularity maximization was explored in order to obtain dimensions of social media. This technique also employed relational learning methods which were described in [5]. However, this framework is not scalable when the networks are very vast in scale.

Later on it is understood that scarifying social dimensions can eliminate the bottleneck of scalability. Recently Tang et al. [7] proposed a framework for scalable learning of collective behavior. Their approach exploits sparse social dimensions in order to make the solution scalable to millions of networks in social media without compromising prediction performance. They employed edge-cluster model to achieve this. However, this model is sensitive to number of social dimensions. In this paper, we enhance the object heterogeneity expressed by edge-centric clustering further by exploiting multiple modes of actors over social media resulting a multi-mode network. Multi-mode networks exhibit relationship among various users can provide very useful information for target marketing besides other utilities[8]. Our main contribution in this paper is the enhancement of edge-centric clustering scheme in order to handle object heterogeneity in multimode networks more effectively. The remainder of the paper is structured as follows. Section II reviews social networking, learning collective behavior and other related content in the literature. Section III provides insights into preliminaries required to understand our problem. Section IV throws light into the proposed scheme and implementation. Section V presents experimental results while section VI concludes the paper.

II. RELATED WORKS

Mining social media content has been around for some years. However, this kind of research started long back. For instance network instances classification was explored in [9]. In similar fashion relational learning [9] was focused by Getoor and Taskar. Conventional data mining is different from that of social data mining. The datasets used for network instances is not uniformly distributed. In such datasets objects have relationships and correlating with neighboring data objects is done with an assumption known as “Markov dependency assumption”. It does mean that label of one network node relies on one of more labels of neighboring nodes. Classification is one of the data mining techniques for supervised learning which is used to classify network objects. For instance in [9] a weighted vote relational classifier [10] was built which showed good performance in classification against benchmark datasets.

A network contains heterogeneous relations and only capturing local dependency is possible with Markov assumption. For this reason in [11] and [12] class labels are used with latent groups. Similar kind of research was done in [5] to explore heterogeneous relationships and differentiate the same by extracting social affiliations and dimensions from social media data. Soft clustering scheme was suggested by them in order to explore community membership in social dimensions. Social dimensions extracted from social media data are known as features and data mining technique such as Support Vector Machine (SVM) can be used to classify such data. The social dimensions approach is better than other approaches to explore social media data based on collective inference. Soft clustering techniques can be used to achieve this which is based on modularity maximization [6], spectral clustering [13], and matrix factorization. To solve the same problem other methods such as probabilistic methods came into existence [14], [15], and [16]. A drawback of soft clustering is that the social dimensions are naturally dense and throw challenges pertaining to computational overhead. The research which is similar to our method in this paper is finding overlapping communities. In [17] Palla et al. proposed this method by name “clique percolation method” which is used to find dense communities which are overlapping. This method has two fundamental phases. They are finding all cliques in graph, and finding connections between cliques in order to discover various communities. Similar idea was explored in [18] where all maximal cliques are found in a network through hierarchical clustering. Overlapping communities are handled by the method proposed by Newman-Girvan [19] which is an extension to the method proposed by Gregory [20]. The Newman-Girvan method recursively removes edges in order to generated disconnected components. The method removes only edges with high betweenness among them. Finally it generates output consisting of non-overlapping communities. Node splitting is another feature added by Gregory besides removing edges. Their algorithm splits nodes recursively where multiple communicates reside and remove edges that are used to interconnect communities. These methods list out all possible cliques and choose the paths which are very short in the network where computational cost is very high in case of large scale networks.

For finding overlapping communities, graph partition algorithms were explored in [21] and [22] that work on line graphs. However, just construction of ling graph is not sufficient as it prohibits functional with large scale networks. Scalable approaches are required in order to deal with huge number of networked objects present in the data of social media. Recently in [7] K-means was used to achieve partitioning of edges to for disjoint sets. They also proposed a variant of K-means to hand scarcity of data in order to handle huge number of edges effectively. IN order to accelerate the process more advanced data structures can be exploited [23], [24]. When the data loaded into RAM is very high, other variants of K-means such as distributed k-means [25], scalable k-means [26] and online k-means [27] can be used.

III. PRELIMINARIES

This section familiarizes the reader about preliminaries required to understand the problem solved in this paper. They details provided here include social networking, social dimension, affiliations, communities, collective behavior, sparse social dimensions, edge clusters, and object heterogeneity. Social networking refers to the online or virtual community including friends, relatives, classmates, family members, researchers and so on who can have a platform to get together and exchanging views. Social dimension refers to the relationship an actor has with others. Affiliation refers to a group of nodes in social network to which an actor belongs. One actor may belong to multiple affiliations. Community refers to a set of edges in the network. Collective behavior refers to the result of a process where the behavior of some objects is known based on the other objects in the same affiliation. Sparse social dimension refers to the social dimension where density is very low. Edge cluster refers to a cluster of objects that is connected to another such cluster in the network. Sample edge clusters are presented in figure 1 with a toy example.

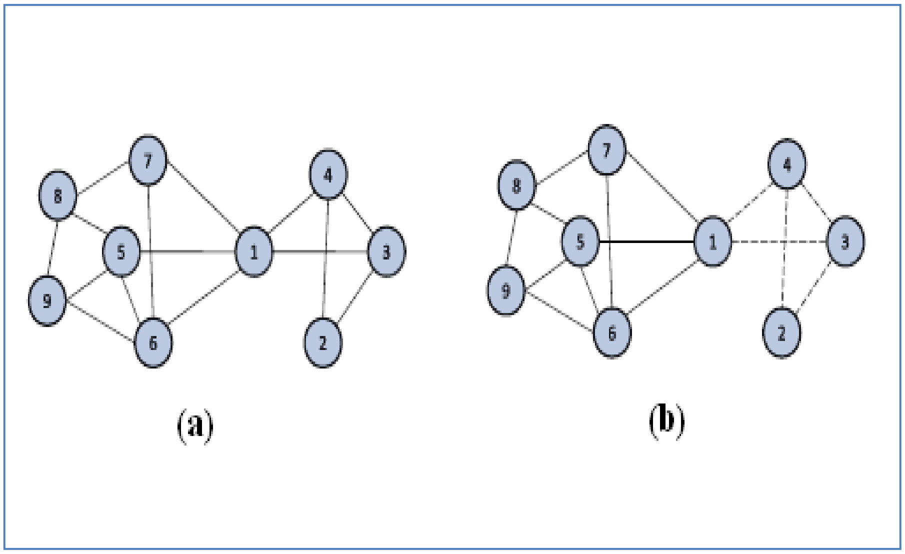
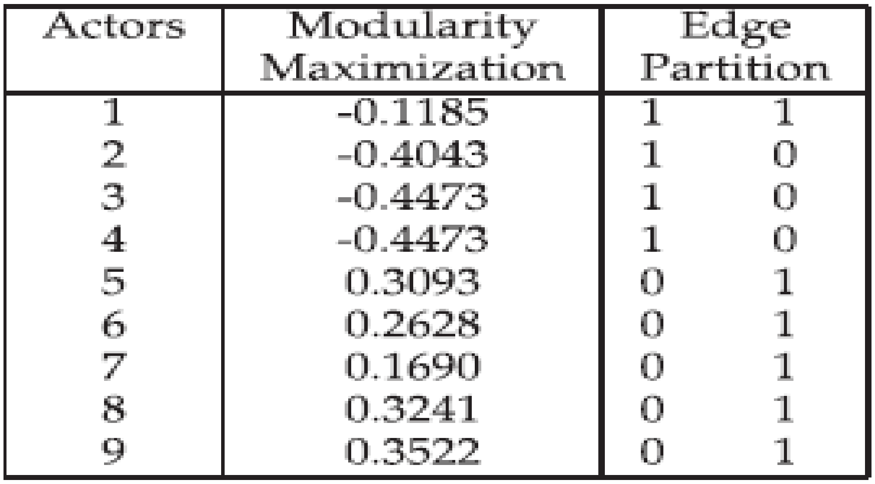


Fig. 1 (a) Toy example (b) Edge clusters (excerpt from [7])

Each object in the network is associated with other objects. One actor can have multiple affiliations. The affiliations can be represented with modularity maximization. The actors, modularity maximization and edge partition details are presented in table 1. It does mean that the table shows social dimensions in the Toy example.



Actors	Modularity Maximization	Edge Partition	
1	-0.1185	1	1
2	-0.4043	1	0
3	-0.4473	1	0
4	-0.4473	1	0
5	0.3093	0	1
6	0.2628	0	1
7	0.1690	0	1
8	0.3241	0	1
9	0.3522	0	1

Table 2 –Social dimensions of Toy example (excerpt from [7])

Multi-mode network is the network which essentially contains multiple and heterogeneous social actors. There are interactions among these actors through which communicates can be identified or evaluated over a period of time. Figure 2 shows a sample multi-mode network which is based on online marketing scenario. Such multi-mode networks exhibit object heterogeneity.

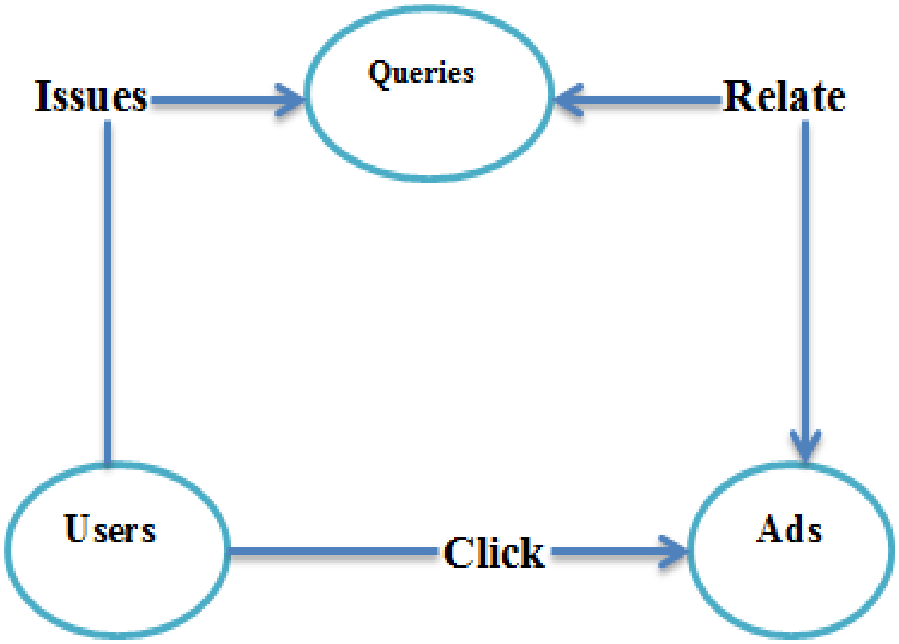
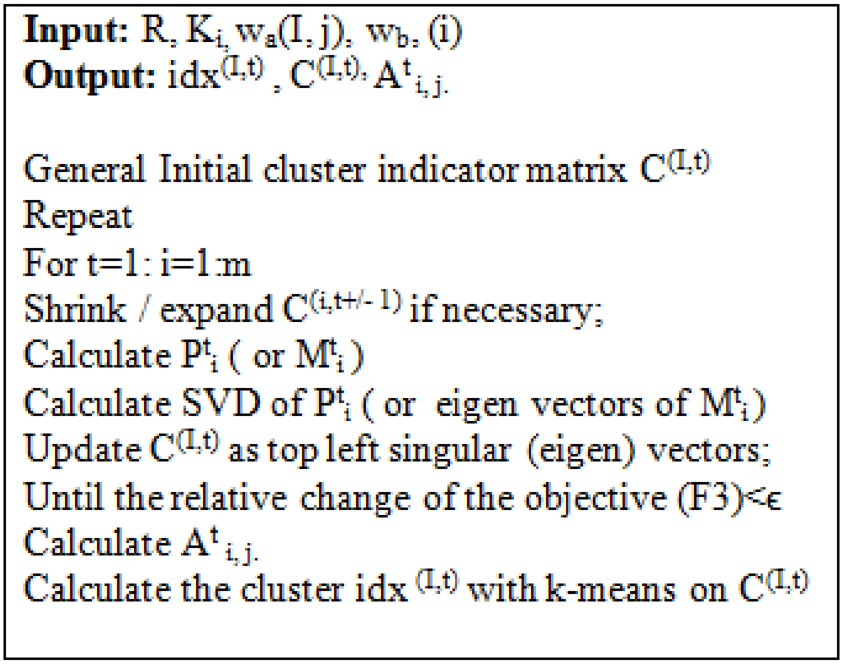


Fig. 2 – Illustrates a multi-mode network with three actors

As can be seen in figure 2, multiple modes are involved in the same network. The resultant network is known as multi-mode network [51]. The queries, users, and ads are intervened with seemingly perfect coexistence. This kind of network shows more object heterogeneity which has to be handled. Object heterogeneity is explored in [7] but the approach is sensitive to number of social dimensions. Handling object heterogeneity has many real world utilities social mining domain. In this paper we focus on addressing object heterogeneity through edge cluster in multi-mode networks.

IV. PROPOSED SCHEME TO ADDRESS OBJECT HETEROGENEITY

Scalable learning of collective behavior is explored in [7] in which a variant of K-means is sued for Edge Clustering. The generated clusters and used to mine the collective behavior. Given knowledge of some actors in a group predicting the behavior of other actors in the same group is known as learning collective behavior. Scalability of this approach is achieved in [7] by using sparse social dimensions. However their solution is sensitive to number of social dimensions. To overcome this drawback, in this paper, we proposed a new scheme that handles object heterogeneity more gracefully. The proposed scheme is presented as pseudo code in figure 3.

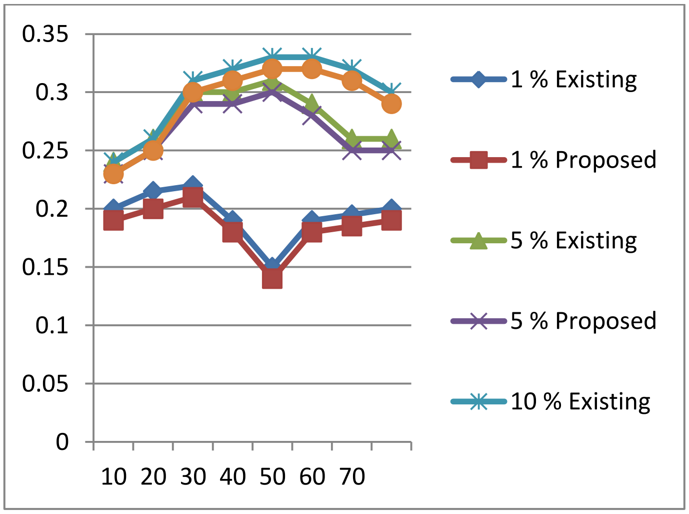


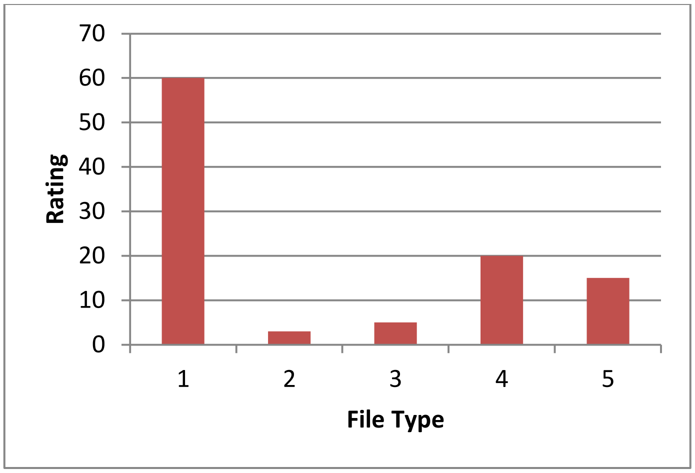
```
Input: $R, K_i, w_a(I, j), w_b, (i)$
Output: $idx^{(I,t)}, C^{(I,t)}, A^t_{i,j}$

General Initial cluster indicator matrix $C^{(I,t)}$
Repeat
For t=1: i=1:m
Shrink / expand $C^{(i,t+/-1)}$ if necessary;
Calculate $P^t_i$ ( or $M^t_i$ )
Calculate SVD of $P^t_i$ ( or eigen vectors of $M^t_i$)
Update $C^{(I,t)}$ as top left singular (eigen) vectors;
Until the relative change of the objective (F3)<ϵ
Calculate $A^t_{i,j}$
Calculate the cluster idx $^{(I,t)}$ with k-means on $C^{(I,t)}$
```

Fig. 3 –Multi-mode clustering algorithm

V. EXPERIMENTAL RESULTS





REFERENCES

[1] S. Wasserman and K. Faust. Social Netwok Analysis: Methods and Applications. Cambridge UniversityPress, 1994.

[2] J. Baumes, M. Goldberg, M. Magdon-Ismail, andW. Wallace. Discovering hidden groups in communication networks. In 2nd NSF/NIJ Symposiumon intelligence and Security Informatics, 2004.

[3] M. N. Lauren Ancel Meyers and B. Pourbohloul. Predicting epidemics on directed contact networks. InJournal of Theoretical Biology, volume 240, 2006.

[4] L. Tang and H. Liu, "Toward Predicting Collective Behavior via Social Dimension Extraction," IEEE Intelligent Systems, vol. 25,no. 4, pp. 19-25, July/Aug. 2010.

[5] L. Tang and H. Liu, "Relational Learning via Latent Social Dimensions," KDD '09: Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 817-826, 2009.

[6] M. Newman, "Finding Community Structure in Networks Using the Eigenvectors of Matrices," Physical Rev. E (Statistical, Nonlinear,and Soft Matter Physics), vol. 74, no. 3, p. 036104, http:// dx.doi.org/10.1103/PhysRevE.74.036104, 2006.

[7] Lei Tang, Xufei Wang and Huan Liu," Scalable Learning of Collective Behavior", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012.

[8] Lei Tang, Huan Liu, Jianping Zhang and Zohreh Nazeri, "Community Evolution in Dynamic Multi-Mode Networks".*KDD'08,* August 24–27, 2008.

[9] Introduction to Statistical Relational Learning, L. Getoor and B. Taskar, eds. The MIT Press, 2007.

[10] S.A. Macskassy and F. Provost, "A Simple Relational Classifier," Proc. Multi-Relational Data Mining Workshop (MRDM) at the NinthACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,2003.

[11] Z. Xu, V. Tresp, S. Yu, and K. Yu, "Nonparametric Relational Learning for Social Network Analysis," KDD '08: Proc. WorkshopSocial Network Mining and Analysis, 2008.

[12] J. Neville and D. Jensen, "Leveraging Relational Autocorrelationwith Latent Group Models," MRDM '05: Proc. Fourth Int'l Workshop Multi-Relational Mining, pp. 49-55, 2005.

[13] U. von Luxburg, "A Tutorial on Spectral Clustering," Statistics and Computing, vol. 17, no. 4, pp. 395-416, 2007.

[14] E. Airodi, D. Blei, S. Fienberg, and E.P. Xing, "Mixed Membership Stochastic Blockmodels," J. Machine Learning Research, vol. 9,pp. 1981-2014, 2008.

[15] K. Yu, S. Yu, and V. Tresp, "Soft Clustering on Graphs," Proc. Advances in Neural Information Processing Systems (NIPS), 2005.

[16] S. Fortunato, "Community Detection in Graphs," Physics Reports, vol. 486, nos. 3-5, pp. 75-174, 2010.

[17] G. Palla, I. Dere´nyi, I. Farkas, and T. Vicsek, "Uncovering the Overlapping Community Structure of Complex Networks inNature and Society," Nature, vol. 435, pp. 814-818, 2005.

[18] H. Shen, X. Cheng, K. Cai, and M. Hu, “Detect Overlapping and Hierarchical Community Structure in Networks,” Physica A:Statistical Mechanics and Its Applications, vol. 388, no. 8, pp. 1706- 1712, 2009.

[19] M. Newman and M. Girvan, “Finding and Evaluating Community Structure in Networks,” Physical Rev. E, vol. 69, p. 026113, http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0308217,2004. 1090 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012

[20] S. Gregory, “An Algorithm to Find Overlapping Community Structure in Networks,” Proc. European Conf. Principles and Practiceof Knowledge Discovery in Databases (PKDD), pp. 91-102, http://www.cs.bris.ac.uk/Publications/pub_master.jsp?id=2000712,2007.

[21] T. Evans and R. Lambiotte, “Line Graphs, Link Partitions, and Overlapping Communities,” Physical Rev. E, vol. 80, no. 1,p. 16105, 2009.

[22] Y.-Y. Ahn, J.P. Bagrow, and S. Lehmann, “Link Communities Reveal Multi-Scale Complexity in Networks,” http://www.citebase.org/abstract?id=oai:arXiv.org:0903.3178, 2009.

[23] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, and A.Y. Wu, “An Efficient k-Means Clustering Algorithm: Analysis and Implementation,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892, July 2002.

[24] J. Bentley, “Multidimensional Binary Search Trees Used for Associative Searching,” Comm. ACM, vol. 18, pp. 509-175, 1975.

[25] R. Jin, A. Goswami, and G. Agrawal, “Fast and Exact Out-of-Core and Distributed K-Means Clustering,” Knowledge and InformationSystems, vol. 10, no. 1, pp. 17-40, 2006.

[26] P. Bradley, U. Fayyad, and C. Reina, “Scaling Clustering Algorithms to Large Databases,” Proc. ACM Knowledge Discoveryand Data Mining (KDD) Conf., 1998.

[27] M. Sato and S. shii, “On-Line EM Algorithm for the Normalized Gaussian Network,” Neural Computation, vol. 12, pp. 407-432, 2000.

AUTHORS

Srikanth Raju T is currently working towards his M.Tech degree in Sreyas Institute of Engineering and Technology, Hyderabad, India. His research interests include networking and cloud computing.

Vennu Santosh Kumar received the Masters degree in Computer Science and Engineering in the year 2010. He is Microsoft Certified System Engineer & CISCO Certified Network Administrator, he worked as a System Engineer in WIPRO Technologies(INDIA). In 2011 he joined as an Associate Professor at Sreyas Institute of Engineering and Technology in Computer Science Department. He has been involved in several tutorials, workshops, technical paper presentations .His research interests are focused on Computer Networks, Network Security & Mobile Computing.

Chintala Ravindranath Yadav is a perceptive academician, with an M.B.A to his credit, and pursued his Post-masters in Business Administration from University of North Carolina at Greensboro, U.S.A. He worked in leading edge organizations for several years in U.S.