ISSN: 2321-7782 (Online)
Volume 2, Issue 1, January 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Paper
Available online at: www.ijarcsms.com

# Study of Effective Behavior Prediction by Scalable Learning Method

**Umesh B. Shingote**
Computer Science and Engineering
M.Tech Research Scholar
TIT, Bhopal - India

***Abstract: For the numerous individuals in world social networks taking part in major role day to day life. Reckoning on the user's behavior and interaction with one another, the social networking sites square measure reshaped. Growing interest and development of social network sites like Facebook, Twitter, Flicker, YouTube etc. imposing several analysis challenges. And hence this can be permitting analysisers to try and do several research studies victimization data processing ideas. The most important challenge of such on-line social networking websites is to search out the people behavior over social network. Understanding the user's behavior on social networking websites is termed as collective behavior. In previous, many methods are presented to spot the behavior of individuals. Such strategies of collective behavior permit to be told and predict the user's on-line behavior and supported it assign the acceptable label to actor in network. .however the main downside happens in such strategies is that the networks scalability because of that this systems becomes poor in performance and lots of not be work if the network size is simply too massive. To overcome this problem we need to have scalable learning of collective behavior to deal with any size of social networks. Recently one such methodology given, in this an edge-centric clustering technique is presented to extract social network dimensions. With sparse social dimensions, the projected approach can efficiently handle sparse social networks of any size. In this paper we are presenting the careful discussion on this methodology.***

***Keywords: Social networks, users behavior, learning, prediction of behavior, k-means variant, clustering.***

## I. INTRODUCTION

Social media generates large amount of data. With the increasing availability of large social network data, there is also an increasing interest in analyzing how those networks evolve over time .Social media can enable new mass collaborative behaviors that unlock the power of the collective and deliver new paths to enterprise results. Enterprises can employ these collective behaviors as the link between business value and social media technologies. Enterprises can use them to examine a target community and formulate new ways that people can interact to achieve enterprise value. In Present World Social media facilitate people of all walks of life to connect to each other. We study how networks in social media can help predict some sorts of human behavior and individual preference. In social media, millions of actors in a network are the norm. With this huge number of actors, the dimensions cannot even be held in memory, causing serious problem about the scalability. For instance, popular content- sharing sites like Del.icio.us, Flickr, and YouTube allow users to upload, tag and comment different types of contents (bookmarks, photos, videos).Users registered at these sites can also become friends, a fan or follower of others. The prolific and expanded use of social media has turn online interactions into a vital part of human experience.

We study how networks in social media can help predict some human behaviors and individual preferences. In particular, given the behavior of some individuals in a network, how can we infer the behavior of other individuals in the same social network [1]. This study can help better understand behavioral patterns of users in social media for applications like social advertising and recommendation? A social-dimension-based approach has been shown effective in addressing the heterogeneity of connections presented in social media. But the networks in social media are normally of colossal size, involving hundreds of thousands of actors. The scale of these networks entails scalable learning of models for collective behavior prediction.

To address the scalability issue, new method introduced an edge-centric clustering scheme to extract sparse social dimensions with sparse social dimensions; the proposed approach can efficiently handle networks of millions of actors while demonstrating a comparable prediction performance to other non-scalable methods. An incomparable advantage of our model is that it easily scales to handle networks with millions of actors while the earlier models fail. This scalable approach offers a viable solution to effective learning of online collective behavior on large scale. Social media facilitate people of all walks of life to connect to each other. In this paper we are presenting and discussing the all algorithms involved in this study with its practical evaluation.

## II. Related Work

There are many methods presented to learn and predict the collective behavior of users over online social networks. In this paper author solves the problem of scalability by using the new data mining methods. In this paper we are discussing the same methods with aim of our study [1].authors presented classification with networked instances are known as within-network classification, or a special case of relational learning [2, 3]. But network tends to present heterogeneous relations, and the Markov assumption can only capture the local dependency. Therefore authors presented approach to model network connections or class labels based on latent groups [4, 5]. In [6], similar idea is also adopted to differentiate heterogeneous relations in a network by extracting social dimensions to represent the potential affiliations of actors in a network.

Some researchers presented methods to conduct soft clustering for graphs. Some are based on matrix factorization, like spectral clustering [7] and modularity maximization [8]. Probabilistic methods are also developed [9, 10]. A disadvantage with soft clustering is that the resultant social dimensions are dense, posing thorny computational challenges.

Pallaetal propose a clique percolation method to discover overlapping dense communities. In [12] same idea is adopted as in [11], in which there is proposal to find out all the maximal cliques of a network and then perform hierarchical clustering.

In [13], Gregory extends the Newman- Girvan method [14] to handle overlapping communities. The original Newman-Girvan method recursively removes edges with highest between's until a network is separated into a pre-specified number of disconnected components. It outputs non- overlapping communities only.

a simple scheme proposed to detect overlapping communities is to construct a line graph and then apply graph partition algorithms[15,16].However, the construction of the line graph alone, as we discussed, is prohibitive for a network of a reasonable size. In order to detect overlapping communities, scalable approaches have to be developed.

## III. Methodology

A methodology of social dimension extraction and prediction is straightforward once a learned model is ready, since the social dimensions have been calculated for all actors. Applying the constructed model to the social dimensions of the actors without behavior information, we obtain the behavior predictions.

### A. *Social Dimension Extraction using K- means Variant Algorithm*

This algorithm based on the framework that styles the input datasets, and implements the algorithm of k-means variant in order to extract the social dimension extraction based edge centric approach k-means is adopted to extract social dimensions, it is easy to update social dimensions if a given network changes. As If a new member joins the network and a new connection emerges, we can simply assign the new edge to the corresponding clusters. The update of centroid with the new arrival of connections is additionally simple. This k-means scheme is especially applicable for dynamic large-scale networks.

International Journal of Advance Research in Computer Science and Management Studies
Volume 2, Issue 1, January 2014 pg. 600-605

Input: Social Network Dataset like flicker, YouTube.

Output: Social dimensions of this dataset.

For example, following figures showing Input and Output of this module:

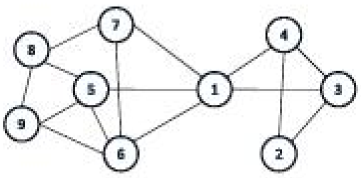


Fig.1 A toy example.

The Edge-centric view of the network data i.e. fig.1 is given below

Table I
Edge Centric view of input network data

| Edge | Features | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| $e(1,3)$ | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| $e(1,4)$ | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| $e(2,3)$ | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| ⋮ | ......... | | | | | | | | |

Further partition this information (edges) into the disjoint sets as per given in below.

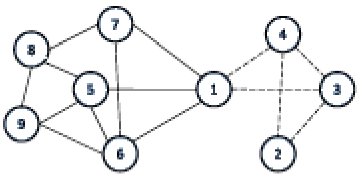


Fig.2 Edge Cluster

Based on Edge clustering above, further social dimensions can be constructed as per given in below Table II and this is the final outcome.

By observing the toy network with two communities in Fig.1 Its social dimensions following modularity maximization are shown in Table II. Clearly, none of the entries is zero. When a network expands into millions of actors, a reasonably large number of social dimensions need to be extracted. Hence, it is imperative to develop some other approach so that the extracted social dimensions are sparse.

Table II
Dimension(s) of the Toy Example

| Actor | 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 |

### *B. Discriminative Learning and Prediction*

The output of above algorithm A means the extracted social dimensions are used as input to this algorithm B

We have to use algorithm given in [1] for this section. The algorithm collective behavior is based on linear SVM.

Input: network data, labels of some nodes, number of social dimensions.

Output: labels of unlabeled nodes.

Apply regularization to social dimensions.

Construct classifier based on social dimensions of labeled nodes.

Use the classifier to predict labels of unlabeled ones based on their social dimensions.

## IV. COMPARATIVE RESULTS

Performance of Sparsity and Performance of Scalability is as shown in following tables.

Table III
Performance of Sparsity

| | Blog Catalog 10k nodes 333k links | Flickr 80k nodes 6M links | YouTube 1M nodes 3M links |
|---|---|---|---|
| ModMax | 194.4 sec | 40 minutes | N/A |
| EdgeCluster | 357.8 sec | 3.6hrs | 10mins |

Table IV
Performance of Scalability

| 500 social dimensions | Blog Catalog (10k) | Flickr (80k) | YouTube (1M) |
|---|---|---|---|
| ModMax | 41.2MB | 322.1MB | 4.6GB |
| EdgeCluster | 4.9MB | 44.8MB | 39.9MB |
| Reduction Rate | 88% | 86% | 99% |
| Density | 6% | 7% | 0.4% |

Comparative analysis shows that proposed methods showing best performances against the existing methods as well as improves scalability. For the experimental evaluation we have used following datasets of different social network.

Blog Catalog: 10K nodes 333K links

Flickr: 80K nodes 6M links

YouTube: 1.1Mnodes 3Mlinks

Based on above inputs we have measured the following performances for proposed methods.

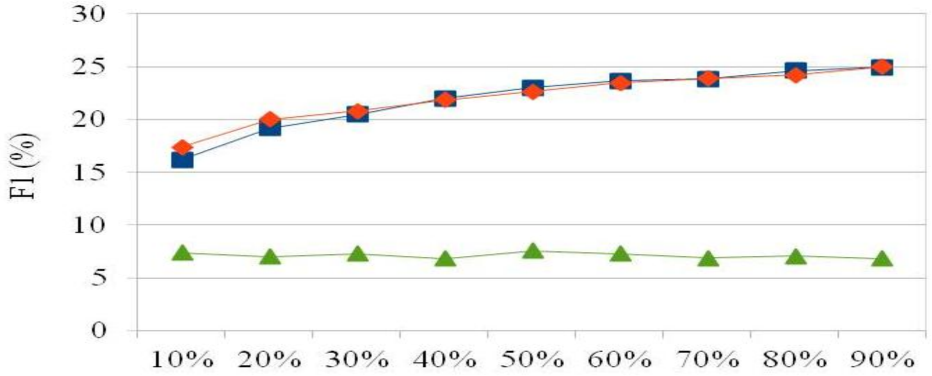


Fig.3 Percentage of label nodes (Blog Catalog)

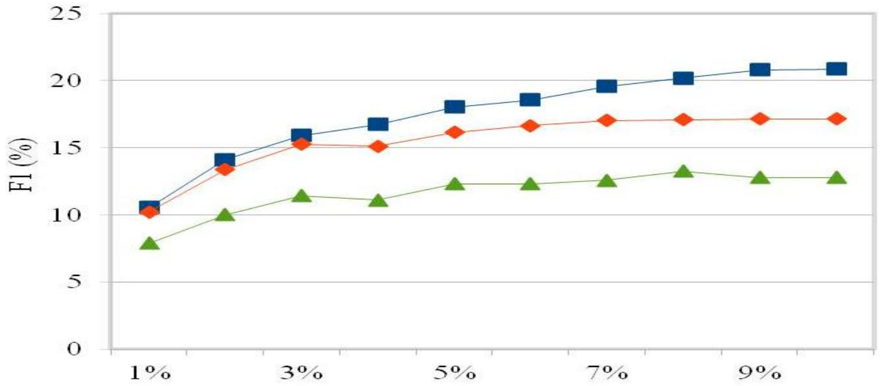


Fig.4 Percentage of label nodes (flickr)

Fig. 3 and 4 shows F1 Score against the Percentage of labeled Nodes for Blog Catalog Dataset and Flickr Dataset respectively.

Blue line: Edge Cluster

Cyan line: ModMax

Green line: NodeCluster

As per experimental results proposed social dimension extraction technique outperforms existing social dimension extraction techniques. This model is that it easily scales to handle networks with millions of actors while the earlier models fail. In all social networks, multiple nodes of actors are involving in similar network and hence this is resulting as multimode network [1].

## V. CONCLUSION

In this study, we have focused on efficient method such as algorithm for k-means variant and collective behavior. This method is presented to address the issues of scalability of all existing methods. The proposed method is given to address the issues of scalability using an edge-centric clustering scheme to extract social dimensions and a scalable k-means variant to handle edge clustering. Each edge can be associated with multiple affiliations while our current model assumes only one dominant affiliation. Given method is sensitive to the number of social dimensions. The advantage of this method is its Scalability, easily scales to handle networks with millions of actors while existing methods was unable to do so. This scalable approach offers an effective solution to effective learning of online collective behavior on a large scale. However this method further needs to improve in different directions. For example. Heterogeneity of social networks needs to be handled during the edge-centric clustering technique hence we can improve the prediction performance particularly in case of multimode networks. Further research is required to determine the dimensionality automatically.

**Acknowledgement**

The paper could not be in completed without the kind guidance and support of my project guide and colleagues. I express my deep gratitude to all the dignitaries who made the efforts in the completion of this paper.

**References**

1. Lei Tang, Xufei Wang, and Huan Liu, "Scalable Learning of Collective Behavior", IEEE 2012 Transactions on Knowledge and Data Engineering, Volume 24, Issue: 6.
2. S. A. Macskassy and F. Provost, "Classification in networked data: A toolkit and a univariate case study", J. Mach. Learn. Res., volume. 8, 2007.
3. L. Getoor and B. Taskar, Eds., "Introduction to Statistical Relational Learning". The MIT Press, 2007.
4. J. Neville and D. Jensen, "Leveraging relational autocorrelation with latent group models", in MRDM '05: Proceedings of the 4th international workshop on Multi- relational mining. New York, NY, USA: ACM, 2005, pp. 49–55
5. Z. Xu, V. Tresp, S. Yu, and K. Yu, Nonparametric relational learning for social network analysis, in KDD'2008 Workshop on Social Network Mining and Analysis, 2008.
6. Relational learning via latent social dimensions in KDD '09": Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2009, pp. 817–826.
7. U. von Luxburg, "A tutorial on spectral clustering", Statistics and Computing, vol. 17, no. 4, pp. 395–416, 2007.
8. 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.
9. K. Yu, S. Yu, and V. Tresp, "Soft clustering on graphs", in NIPS, 2005.
10. E. Airodi, D. Blei, S. Fienberg, and E. P. Xing, Mixed membership stochastic blockmodels,J. Mach. Learn. Res., vol. 9, pp. 1981–2014, 2008.
11. G. Palla, I. Der´enyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society", Nature, vol. 435, pp. 814–818, 2005.
12. H. Shen, X. Cheng, K. Cai, and M. Hu, "Detect overlapping and hierarchical community structure in networks", Physical A: Statistical Mechanics and its Applications, vol. 388, no. 8, pp. 1706–1712, 2009.
13. S. Gregory, "An algorithm to find overlapping community structure in networks", in PKDD, 2007, pp.91–102.
14. M. Newman and M. Girvan, "Finding and evaluating community structure in networks", Physical Review E, vol. 69, p. 026113, 2004.
15. T. Evans and R. Lambiotte, "Line graphs, link partitions, and overlapping communities", Physical Review E, vol. 80, no. 1, p. 16105, 2009.
16. Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multi-scale complexity in networks", 2009.
17. J. Hopcroft and R. Tarjan, "Algorithm 447: efficient algorithms for graph manipulation," Commun. ACM, vol. 16, no. 6, pp. 372–378, 1973.

**AUTHOR(S) PROFILE**

**Umesh B. Shingote** completed his bachelor degree in Information Technology from Govt.College of Engineering, SGBAU Amravati, India. And currently he is a scholar of Master of Technology (CTA) RGPV University Bhopal India. His areas of research are Data mining, Wireless Networks, Image processing.