ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

Classification of Activities in Social Media

[1]NareshBabu.K, [2]B.Prashant
[1]Final M Tech Student, [2]Associate professor
Dept of Computer Science and Engineering[1,2],
Eluru College of Engineeering and Technology, Eluru,W.G Dist,A.P[1,2].

Abstract: *S*ocial Media has been gaining tremendous momentum. Blogs have rapidly evolved into a plethora of social media platforms that enabled people to connect, share and discuss anything and everything, from personal experiences, to ideas, facts, events, music, videos, movies, and the list goes on forevermore. In this work, we propose an effective edge-centric approach to extract sparse social dimensions. The study of collective behavior is to understand how individuals behave in a social network environment. Oceans of data generated by social media like Facebook, Twitter, Flicker and YouTube present opportunities and challenges to studying collective behavior in a large scale. In this work, we aim to learn to predict collective behavior in social media. In particular, given information about some individuals, how can we infer the behavior of unobserved individuals in the same network? A social-dimension based approach is adopted to address the heterogeneity of connections presented in social media. However, the networks in social media are normally of colossal size, involving hundreds of thousands or even millions of actors. The scale of networks entails scalable learning of models for collective behavior prediction. To address the scalability issue, we propose an edge-centric clustering scheme to extract sparse social dimensions. With sparse social dimensions, the social dimension based approach can efficiently handle networks of millions of actors while demonstrating comparable prediction performance as other non-scalable methods.

Keywords: Social Dimensions, Behavior Prediction, Social Media, Relational Learning, Edge-Centric Clustering

I. INTRODUCTION

Social Media :

Social media refers to the means of interactions among people in which they create share, exchange and comment contents among themselves in virtual communities and networks. Andreas Kaplan and Michael Haenlein define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content. Furthermore,social media employ mobile and web-based technologies to create highly interactive platforms via which individuals and communities share, co create, discuss, and modify user-generated content. It introduces substantial and pervasive changes to communication between organizations, communities and individuals.

Different types of social media include collaborative projects such as Wikipedia, blogs such as Blogger, social networking sites like Facebook, content communities like Youtube and so on.

Collective Behavior

Social media such as Facebook, MySpace, Twitter, Blog Catalog, Digg, YouTube and Flickr, facilitate people of all walks of life to express their thoughts, voice their opinions, and connect to each other anytime and anywhere. For instance, popular content-sharing sites like Del.icio.us, Flickr, and YouTube allow users to upload, tag and comment different types of contents (e.g., bookmarks, photos, videos). Users registered at these sites can also become friends, a fan or follower of others. The police and expanded use of social media has turned online interactions into a vital part of human experience. The election of Barack

Obama as the President of United States was partially attributed to his smart Internet strategy and access to millions of younger voters through the new social media, such as Facebook, a popular social networking site claiming to attract 400 million active users up to date1. The large population actively involved in social media also provides great opportunities for business.

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

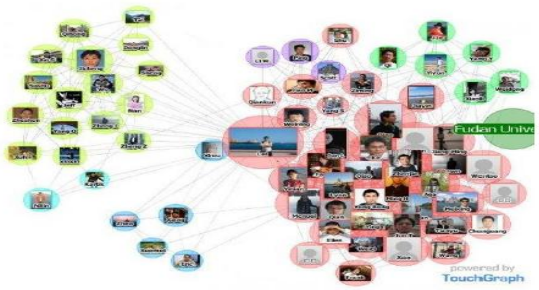


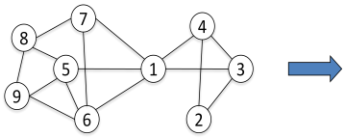
Figure 1: Contacts of One User in Facebook

One of the top Collective behaviors is not simply the aggregation of individuals' behavior. In a connected environment, behaviors of individuals tend to be interdependent. That is, one's behavior can be incensed by the behavior of his/her friends. This naturally leads to behavior correlation between connected users. Such collective behavior correlation can also be explained by homophile is a term coined in 1950s to explain our tendency to link up with one another in ways that comer rather than test our core beliefs. Essentially, we are more likely to connect to others sharing certain similarity with us. This phenomenon has been observed not only in the real world, but also in online environments. In other words, similar people tend to become friends, leading to similar behavior between connected egos in a social network. Take marketing as an example. If our friends buy something, there is a better-than-average chance that we'll buy it too.

Since a social network provides valuable information concerning actor behaviors, one natural question is how we can utilize the behavior correlation presented in a social network to predict collective behavior. In particular, the collective behavior prediction problem can be stated as follows:

1.1 Heterogeneous Relations in Social Networks:

It is often a luxury to have detailed relation information, though some sites like LinkedIn and Face book do ask people how they know each other when they become connected. Most of the time, people decline to share such detailed information, resulting in a social network between users without explicit information about pair wise relation type. Even if the pair wise relation information is available, it is not necessarily relevant or reined enough to help determine the behaviors of connected users. For example, knowing two actors are college classmates does not help much for the behavior prediction of voting for a presidential candidate.



Actor	Social Dimensions	
1	-0.1185	0.4746
2	-0.4043	-0.3135
3	-0.4473	-0.0594
4	-0.4473	-0.0594
5	0.3093	-0.1648
6	0.2628	0.3243
7	0.1690	0.5196
8	0.3241	-0.3027
9	0.3522	-0.4189

The above concerns pose the following two challenges to be addressed for collective behavior prediction:

- Without information of relation type, is it possible to different relations based on network connectivity?
- If relations are different, how can we determine whether a relation can help behavior prediction?

1.2 Social Dimensions:

1. To address the heterogeneity presented in connections, we have proposed a framework (SocDim) for collective behavior learning.
2. Framework SocDim is composed of two steps:
    - Social dimension extraction
    - discriminative learning

Table 1: Social Dimension Representation

Actors	Affiliation-1	Affiliation-2	⋯	Affiliation-k
1	0	1	⋯	0.8
2	0.5	0.3	⋯	0
⋮	⋮	⋮	⋱	⋮

These social dimensions can be treated as features of actors.

1. Since network is converted into features, typical classifier such as support vector machine can be employed.
2. Concerns about the scalability of SocDim with modularity maximization:

    a) The social dimensions extracted according to modularity maximization are dense.

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

b) Requires the computation of the top eigenvectors of a modularity matrix which is of size n*n.
c) The dynamic nature of networks entails efficient update of the model for collective behavior prediction.

Social dimensions are introduced to represent the relations associated with actors, with each dimension denoting one relation. Suppose two actors' ai and aj are connected because of relationship R, both ai and aj should have a non-zero entry in the social dimension which represents R. Let us revisit the example in Figure 1. The relations between the user and his friends can be characterized by three ablations Arizona State University (ASU), Fudan University (Fudan), and a high school (Sanzhong). The corresponding social dimensions of actors in Figure 1 are shown in Table 1. In the table, if one actor belongs to one allegation, then the corresponding entry is non-zero. Since Lei is a student ASU, his social dimension includes a non-zero entry for the ASU dimension to capture the relationship of his ASU friends and him. Social dimensions capture prominent interaction patterns presented in a network. Note that one actor is very likely to be involved in multiple deferent social dimensions (e.g., Lei participates in 3 deterrent relations in the table). This is consistent with multi-facet nature of human social life that one is likely to be involved in distinctive relations with deferent people.

1.3 Social-Dim Framework:

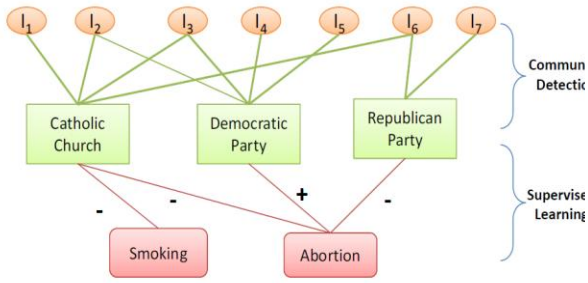


Figure 2: Underlying Collective Behavior Model for SocioDim framew

The social dimensions shown in Table 1 are constructed based on the explicit information of relations. In reality, without knowing true relationship, how can we extract latent social dimensions? One key observation is that actors of the same relation tend to connect to each other as well. For instance, as shown in Figure 1, the friends of Lei at ASU tend to interact with each other as well. Hence, to infer a latent social dimension, we need out a group of people who interact with each other more frequently than random. This boils down to a classical community detection problem. A requirement is that one actor is allowed to be assigned to multiple communities. After we extract the social dimensions, we consider them as normal features and combine them with the behavioral information to conduct supervised learning. Deferent tasks might represent the user behavior in divergent ways. In certain cases, we can represent the behavior output by labels. For instance, In summary, a social-dimension based learning framework SocioDim [8] can be applied to handle the network heterogeneity.

II. ALGORITHM—EDGECLUSTER

In this section, we first show one toy example to illustrate the intuition of our proposed edge-centric clustering scheme *Edge Cluster*, and then present one feasible solution to handle large-scale networks.

2.1 Edges-Centric View:

The social dimensions according to modularity maximization or other soft clustering scheme tend to assign a non-zero score for each actor with respect to each affiliation. However, it seems reasonable that the number of affiliations one user can participate in is upper bounded by the number of connections. Consider one extreme case that an actor has only one connection. It is expected that he is probably active in only one affiliation. It is not necessary to assign a nonzero score for each affiliation. Assuming each connection represents one dominant affiliation, we expect the number of affiliations of one actor is no more than his connections. Instead of directly clustering the nodes of a network into some communities, we can take an edge-centric view, i.e., partitioning the edges into disjoint sets such that each set represents one latent affiliation. For instance, we can treat each edge in the toy network in Figure 2 as one instance, and the nodes that define edges as features. This results in a typical feature-based data format as in Figure 3.

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

Based on the features (connected nodes) of each edge, we can cluster the edges into two sets as in Figure 4, where the dashed edges represent one affiliation, and the remaining edges denote another affiliation. One actor is considered associated with one affiliation as long as any of his connections is assigned to that affiliation. Hence, the disjoint edge clusters in can be converted into the social dimensions as the last two columns for edge-centric clustering in Table 2. Actor 1 is involved in both affiliations under this *Edge Cluster* scheme.

In summary, to extract social dimensions, we cluster edges rather than nodes in a network into disjoint sets. To achieve this, k-means clustering algorithm can be applied. The edges of those actors involving in multiple affiliations (e.g., actor 1 in the toy network) are likely to be separated into different

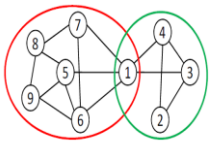


Figure 3: A Toy Example

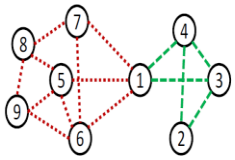


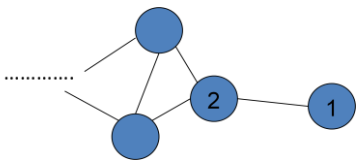
Figure 4: Edge Clusters

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

2.2 Bounded Number of Affiliations:

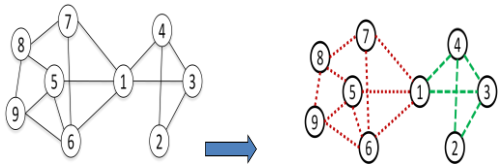
1. One actor is likely to be involved in multiple affiliations
2. Number of affiliations should be bounded by the connections one actor has.

- $Actor_1$: 1 connection, at most 1 affiliation
- $Actor_2$: 3 connections, at most 3 affiliations



2.3 Edge Partition:

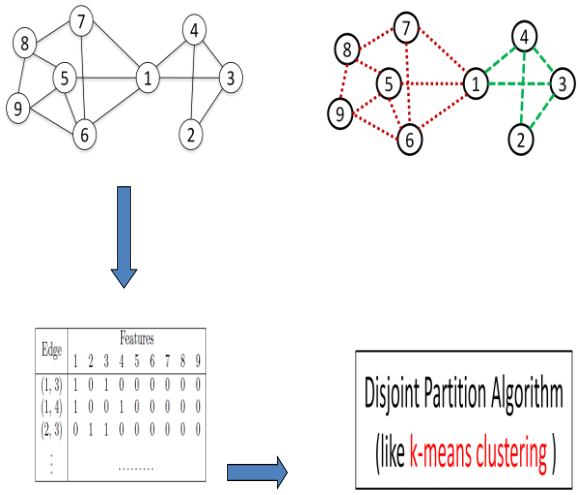
- Each edge is involved in only one relation
- Partition edges into disjoint sets



2.4 Sparsely of Social Dimensions:

1) Power law distribution in large-scale social networks
   i. $p(x) = Cx^{-\alpha}, \quad x \geq 1$
2) Density Upper bound (More details in the paper)
   i. $\frac{\alpha-1}{\alpha-2}\frac{1}{k} - \left(\frac{\alpha-1}{\alpha-2} - 1\right) k^{-\alpha+1}$

2.5 Edge Cluster Algorithm



2.6 K-means exploiting sparsity:

1) Apply k-means algorithm to partition edges Millions of edges are the norm Need a scalable and efficient k-means implementation
2) Exploit the sparsity of edge-centric data Build feature-instance mapping (like inverse-index table in IR) Only compute the distance between a centroid to those relevant instances with sharing features please refer to paper for details

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

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

2.7 Overview of Edge Cluster Algorithm

- Apply k-means algorithm to partition edges into disjoint sets
  1. One actor can be assigned to multiple affiliations
  2. Sparse (Theoretically Guaranteed)
  3. Scalable via k-means variant
     Space: O(n+m)
     Time: O(m)
  4. Easy to update with new edges and nodes
     Simply update the centroids

In addition, the social dimensions based on edge-centric clustering are *guaranteed to be sparse*. This is because the affiliations of one actor are no more than the connections he has. Suppose we have a network with m edges, n nodes and k social dimensions are extracted. Then each node vi has no more than min (di, k) non-zero entries in its social dimensions, where di is the degree of node vi. We have the following theorem.

Theorem 1. *Suppose* k *social dimensions are extracted from a network with* m *edges and* n *nodes. The density (proportion of nonzero entries) of the social dimensions extracted Based on edge-centric clustering is bounded by the Following formula:*

$$density \leq \frac{\sum_{i=1}^{n} \min(d_i, k)}{nk} = \frac{\sum_{\{i|d_i \leq k\}} d_i + \sum_{\{i|d_i > k\}} k}{nk} \quad (1)$$

Moreover, for networks in social media where the node degree follows a power law distribution, the upper bound in Eq. (1) can be approximated as follows:

$$\frac{\alpha-1}{\alpha-2}\frac{1}{k} - \left(\frac{\alpha-1}{\alpha-2} - 1\right) k^{-\alpha+1} \quad (2)$$

Where $\alpha < 2$ is the exponent of the power law distribution. Please refer to the appendix for the detailed proof. To give a concrete example, we examine a YouTube network2 with More than 1 million actors and verify the upper bound of the Density.

III. ALGORITHM

One concern with this scheme is that the total number of edges might be too huge.

1.1 K-means Variant

Owning to the power law distribution of node degrees presented in social networks, the total number of edges is normally linear, rather than square, with respect to the number of nodes in the network. That is, m = O(n). This can be verified via the properties of power law distribution. Suppose a network with n nodes follows a power law distribution as

Input: data instances $\{x_i | 1 \leq i \leq m\}$
number of clusters k
Output: $\{idx_i\}$

construct a mapping from features to instances
initialize the centroid of cluster $\{C_j | 1 \leq j \leq k\}$
repeat
  Reset $\{MaxSim_i\}$, $\{idx_i\}$
  for j=1:k
    identify relevant instances S_j to centroid C_j
    for i in S_j
      compute $sim(i, C_j)$ of instance i and C_j
      if $sim(i, C_j) > MaxSim_i$
        $MaxSim_i = sim(i, C_j)$
        $idx_i = j$;
  for i=1:m
    update centroid C_{idx_i}
until no change in idx or change of objective $< \epsilon$

Figure 6: Algorithm for Scalable K-means Variant

Then the expected number of degree for each node is

$$p(x) = Cx^{-\alpha}, \quad x \geq x_{min} > 0$$

where xmin is the minimum nodal degree in a network. In Reality, we normally deal with nodes with at least one connection, so xmin _ 1. The of a real-world network following power law is normally between 2 and 3 as mentioned in [14]. Consider a network in which all the nodes have non-zero degrees, the expected number of edges is

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

Input: network data, labels of some nodes
Output: labels of unlabeled nodes

1. convert network into edge-centric view as in Figure 3
2. perform clustering on edges via algorithm in Figure 6
3. construct social dimensions based on edge clustering
4. build classifier based on labeled nodes' social dimensions
5. use the classifier to predict the labels of unlabeled ones based on their social dimensions

Figure 7: Scalable Learning of Collective Behavior

Unless α is very close to 2, in which case the expectation diverges, the expected number of edges in a network is linear to the total number of nodes in the network. Still, millions of edges are the norm in a large-scale social Direct application of some existing k-means implementation cannot handle the problem. E.g., the k-means code provided in Matlab package requires the computation of the similarity matrix between all pairs of data instances, which would exhaust the memory of normal PCs in seconds. Therefore, implementation with an online fashion is preferred.

As a simple k-means is adopted to extract social dimensions, it is easy to update the social dimensions if the network changes. If a new member joins a network and a new connection emerges, we can simply assign the new edge to the corresponding clusters. The update of centroids with new arrival of connections is also straightforward. This k-means scheme is especially applicable for dynamic large scale networks.

II. Performance on YouTube:

Prediction performance on all the studied social media data is around 20-30% for F1 measure. This is partly due to:

1. large number of labels in the data
2. only employ the network information

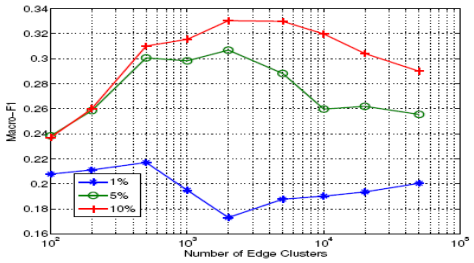


Figure 10: Sensitivity to Dimensionalit

III. Conclusions and Future Work:

1. To address the scalability issue, we propose an edge-centric clustering scheme to extract social dimensions and a scalable k-means variant to handle edge clustering.
2. The model based on the sparse social dimensions shows comparable prediction performance as earlier proposed approaches to extract social dimensions.
3. In reality, each edge can be associated with multiple affiliations while our current model assumes only one dominant affiliation.
4. The proposed Edge Cluster model is sensitive to the number of social dimensions.

Contributions:

- Propose a novel Edge Cluster algorithm to extract sparse social dimensions for classification
- Develop a k-means algorithm via exploiting the sparsity

***Core Idea*:** Partition edges into disjoint sets

- Actors are allowed to participate in multiple affiliations
- Representation becomes sparse with theoretical justification
- Time and space complexity is linear
- Performance is comparable to dense social dimensions

REFERENCES

[1] J. Bentley. Multidimensional binary search trees used for associative searching. Comm. ACM, 1975.

[2] P. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In ACM KDD Conference, 1998.

[3] R.-E. Fan and C.-J. Lin. A study on threshold selection for multi-label classification. 2007.

[4] A. T. Fiore and J. S. Donath. Homophily in online dating: when do you like someone like yourself? In CHI '05: CHI '05 extended abstracts on Human factors in computing systems, pages 1371–1374, 2005.

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

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

[6] M. Hechter. Principles of Group Solidarity. University of California Press, 1988.

[7] R. Jin, A. Goswami, and G. Agrawal. Fast and exact out-of-core and distributed k-means clustering. Knowl.Inf. Syst., 10(1):17–40, 2006.

[8] 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 Transactions on Pattern Analysis and Machine Intelligence, 24:881–892, 2002.

[9] Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In AAAI, 2006.

[10] S. A. Macskassy and F. Provost. A simple relational classifier. In Proceedings of the Multi-Relational Data Mining Workshop (MRDM) at the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[11] S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. J. Mach. Learn. Res., 8:935–983, 2007.

[12] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 27:415–444, 2001.

[13] 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, pages 49–55, 2005.

[14] M. Newman. Power laws, Pareto distributions and Zipf's law. Contemporary physics, 46(5):323–352, 2005.

[15] M. Newman. Finding community structure in networks using the eigenvectors of matrices. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), 74(3), 2006.

[16] C. Ordonez. Clustering binary data streams with k-means. In DMKD '03: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, pages 12–19, 2003.

[17] M. Sato and S. Ishii. On-line em algorithm for the normalized gaussian network. Neural Computation, 1999.

[18] L. Tang and H. Liu. Relational learning via latent social dimensions. In KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 817–826, 2009.

[19] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. Community evolution in dynamic multi-mode networks. In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and

[20] Lei Tang and Huan Liu. Scalable Learning of Collective Behavior based on Sparse Social Dimensions. In CIKM'09, 2009.

[21] Macskassy, S. A. and Provost, F. Classification in Networked Data: A Toolkit and a Univariate Case Study. J. Mach. Learn. Res. 8 (Dec. 2007), 935-983. 2007

[22] Neville, J. and Jensen, D. 2005. Leveraging relational autocorrelation with latent group models. In Proceedings of the 4th international Workshop on Multi-Relational Mining, 2005.

About Author

NareshBabu.K received his Mastres Degree in computer Applications (MCA) From LakiReddy BaliReddy College of Engineering, mylavaram, Krishna Dist, in 2011, the M.Tech. Degree in CSE from Eluru College of Engineering and Technology Eluru in 2013. At present, He is engaged in "Classification of Activities in Social Media".

B.Prashant received his B.Tech Degree In EEE From Bapatla Engineering College, Bapatla, Guntur(Dt), In 2002, M.Tech. Degree In Cse From Nova College Of Engineering And Technology, Jangareddygudem In 2011. He has 8 years of experience in teaching. Currently he is working as associate professor in Eluru College Of Engineering And Technology, Eluru Area of interests artificial intelligence and neural networks, memory management, data mining.