International Journal of Latest Engineering Research and Applications (IJLERA)
www.ijlera.com
Volume 1 – Issue 1 pp: 01-08

A Specialized Framework on Social Learning for Aggregate Performance

S. Prasanthi
Computer Science & Engineering
KMM Institute of Technology and Science, India

CH. Changamma,
Assist. Prof of Computer Science,
KMM Institute of Technology and Science, India

S. SivaRamaKrishna,
Head of the Department of Computer Science,
KMM Institute of Technology and Science, India

Abstract: Aggregate Performance is study of behavior of a person who's registered on a social network. More number of individuals are connected to each other through this networks. Currently media is facing a major problem for finding the individual behavior prediction. Because of so many people having on the network, the study of their behavior is so called Social Learning. Problems rising during social learning is because of more scalability. We introduced a new framework which simplify the study of social learning and handles huge amount of data over the social media. During this process of resolution, we used efficient classification methodologies and new features of distributing the social media such as centric based clustering.
Keywords: Social network, Collective Behavior, Clustering and Social dimensions.

I. INTRODUCTION

Social networking platforms like Twitter, Facebook and YouTube may allow several organizations to advance communication and productivity by publishing information among various groups of users in a more efficient mode. Social networking is the practice of expanding the number of one's business and/or social contacts by making connections through individuals. Traditional social networks have expanded from a few dozen acquaintances to hundreds of friends, friends of friends, connections, followers and public users. This results in huge rise in traffic over the social media. So the question here is: **How do we come up with a solution to allow this traffic over social media?**

When people are exposed in a social network environment, their behaviors can be influenced by the behaviors of their friends. People are more likely to connect to others sharing certain similarities with them. This indeed leads to behavior correlation between connected users [5]. ForExample: If any new friends come into networks, weareengrossed to know about what's the friend's media and so on.

When most people hear the term social network, they robotically think of online social networks. Sites like Myspace, Facebook, LinkedInaccounts for top 20 most visited Websites in the World. For many users,it's like a fully wired net generation. Online social networks are not only a way to keep in touch, but a way of life.

For example: Social networks are as follows:

International Journal of Latest Engineering Research and Applications (IJLERA)
www.ijlera.com
Volume 1 – Issue 1 pp: 01-08

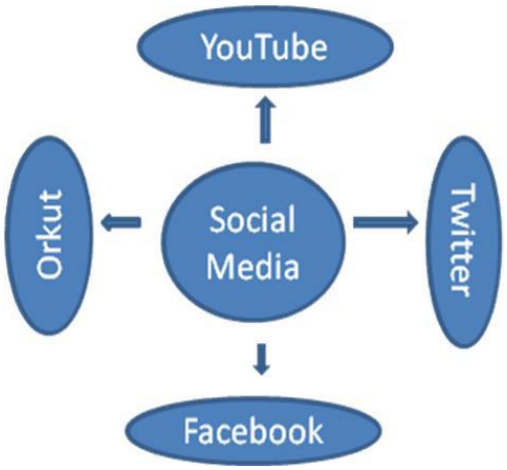


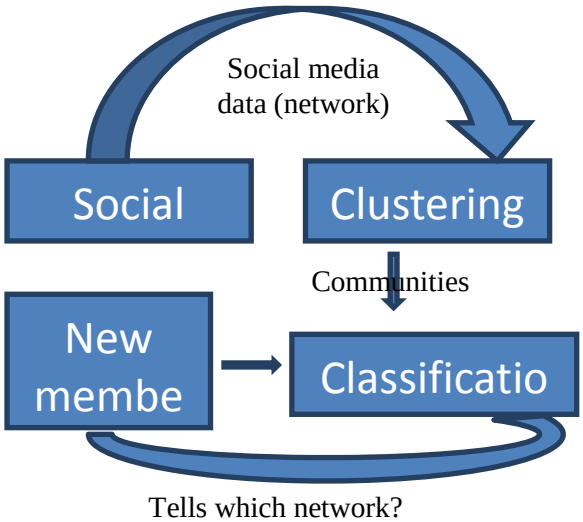
Fig1:Social Media

Social networks are a great way to meet with people and keep in touch with friends that's because online socialnetworks are also known as social networking sitesthat have exploded recently in popularity.
Based on the Six degrees of separation concept (the idea that any two people on the planet could make contact through a chain of no more than five intermediaries), social networking establishes interconnected online communities (sometimes known as social graphs) that helps people make contacts that would be good for them to know, but that they would be unlikely to have met otherwise.

II. FRAMEWORK

Here collecting the behavior of a number of users gives the aggregate performance. The aggregate performance is study of person's behavior who's registered on a social network. More number of individuals are connected to each other on this network. So here clusteringcomes into picture where individuals have same behavior. Predict the individuals of a network according to class or category. Predicting the type of an object based on itsattributes, links and attributes of linked objects. For Example: Predict the Venue type of a paper publication i.e., Conference,journal, workshop based on paper properties.

Architecture:



Given a network, graph partition algorithms can be applied to its corresponding line graph. The set of communities in the line graph corresponds to a disjoint edge partition in the original graph. Recently, such a scheme has been used to detect overlapping communities [16], [17]. It is, however, prohibitive to construct a line graph for a mega-scale network. We notice that edges connecting to the same node in the original network form a clique in the corresponding line graph. For example, edges e (1, 3), e (2, 3), and e (3, 4) are all neighboring edges of node 3 in Figure 1. Hence, they are adjacent to each other, forming a clique. This property leads to many more edges in a line graph than in the original network. In our framework the given network is scanned and divides into disjoints sets. Then it is converted into edge centric view shown in table 1.

Edge	1	2	3	4	5	6	7	8	9
E(1,4)	1	0	0	1	0	0	0	0	0
E(2,3)	0	1	1	0	0	0	0	0	0
:	:	:	:	:	:	:	:	:	:

Table1

By dividing into edge centric view it easy to identify whichnodes are connected each other for further process of ourframework. Then we apply edge clustering methods forfinding the similarity between the edges that means theindividuals. For this we used incremental clustering. Themain purpose of clustering the edges has two reasons.

Those are connections between the users and scalability ofthe connected users. If an edge connects two nodesdefinitely the corresponding edge features are non-zero numerical. In this mainly we have sparsity that means scattering of the edges in the network, so we have to gather similar properties of the nodes. So that we can easilyclassify the testing edges of the network. So we use incremental clustering algorithm to cluster the edges in the network. The clustering methodology is shown below:

Algorithm 1:

Input: edges in the edge centric view

Output: Clusters having the similar features of edges.

1. Select centroids based on number of affiliations.
2. For all centroids find distance between the centroid and the edge in the network.

dist =sqrt[(x2-x1)2 – (y2-y1)2]

Minimum distanced edge put on the cluster of corresponding centroid.

3. Save centroids. If we add affiliation, repeat step 2.
4. Don’t repeat step2.

We keep only a vector of MaxSim to represent the maximum similarity between one data instance and a centroid. In each iteration, we first identify the instances relevant to a centroid, and then compute similarities of these instances with the centroid. This avoids the iteration over each instance and each centroid, which will cost O(mk) otherwise. Note that the centroid contains one feature (node), if and only if any edge of that node is assigned to the cluster.

After clustering of the edges we construct classifier based on the social dimensions. We designed a new classifier. This classification is based on the cluster probability and testing edge. If any unlabeled edge connected to a network, we have classify that edge belongs to which cluster. So we used most efficient classification that is Bayesian classification to find the new belongs to which cluster. A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes’ theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model”. The discussion so far has derived the independent feature model, that is, the naive Bayes probability model.

The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule. The corresponding classifier, a Bayes classifier, is thefunction Classify defined as follows:

Classify (f1...,fn)= argmax p(C=c) $\pi\ i = 1\ To\ n$ P(Fi=fi |C=c)
In the classification the unlabeled edge is classified to labelled edge. In this mean, variance calculations are more dependent to calculate the probability of the unlabeled edge.

III. EXISTING SYSTEM

If we want to know the details of an individual, we have to know all the neighborsof an individual. We have to search the entire network of social media. In the existing method, individual search is mainly based on the similarities of an individual present in the network. And the classification of the individuals also depends upon the similarities. It's really a difficult task in finding the networks present in the social media. It shows some efficiency problems when working with large amount of data. During this process, clustering or grouping the data according to community density plays a vital role.

IV.PROPOSED SYSTEM

In the proposed approach, introduced an incrementalclustering algorithm for grouping or clustering accurate. In our framework, we reduced time complexity of processing by dividing into groups. By using proper clustering methodologies we reduce processing time, to gain best results. For classification some constraints should beconsidered for predicting the individuals to find to which network they belongs to.
Social media consists of several networks. These networks are connected to each other.
The information about connected networks is maintained by social media. This information is clustered according to the edge clustering mechanism. After clustering, it divided into communities. If any new user come into network, the Bayesianclassifier tells it belongs to which network.
This proposed system is divided into 3 modules:

MODULES:

1. Edge Centric View

2. Edge Clustering

3. Discriminative Learning and Prediction

Edge Centric View:
In this module we input the network that means the nodes and their connections. Then convert them into edge centric view. Then partition the edges into disjoint sets. We treat edges as data instances with their terminal nodes as features. One concern with this scheme is that the total number of edges might be too huge. Owing 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.

Edge Clustering:
After defining the edges we apply incremental algorithm to cluster the edges. We can keep maximum similarity of one data instance and centroid. In each iteration, we first identify the instances relevant to a centroid, and then compute similarities of these instances with the centroid.

This avoids the iteration over each instance and each centroid, which will cost O(mk) otherwise. Note that the centroid contains one feature (node), if and only if any edge of that node is assigned to the cluster. In effect, most data instances (edges) are associated with few (much less than k) centroids.

Discriminative Learning and Prediction:
After clustering of the edges, we will get fine clusters with labels. If any unlabeled edge connected to a network, we have classify that edge belongs to which cluster. So we used most efficient classification that is Bayesian classification to find the new belongs to which cluster.

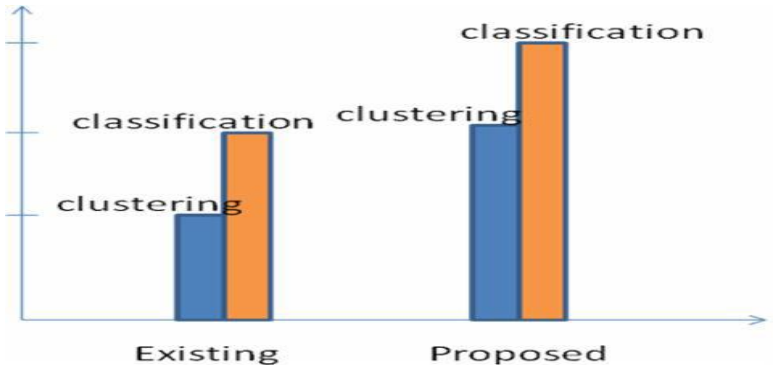
A naive Bayes classifier is a simple probabilistic classifierbased on applying Bayes Theorem with strong(naive) independence assumptions. A more descriptiveterm for the underlying probability model would be"**independent feature model**".

V. IMPLEMENTATION

The Social media data i.e., called network give as an input to the incremental clustering algorithm then it clustering or grouping the nodes according to the communities. Then it classifies the data .The Classification describes –if any node come into network then that classification method i.e., Naïve Bayesian classification tells that node belongs to which network.

VI. ANALYSIS AND RESULT

In this we provide analysis about the characteristics ofproposed system.



Effectiveness: By using the clustering method we get the effective grouping according to communities.

Accurate: By using these methods we get an accurate performance i.e. fast.

VII. RELATED WORK

Connections in social media are nothomogeneous. People can connect to their family,colleagues, college classmates, or buddies met online. Some relations are helpful in determining a targetedbehavior (category) while others are not. This relation-typeinformation, however, is often not readily available insocial media. A direct application of collective inference[9] or label propagation [12] would treat connections in asocial network as if they were homogeneous. To addressthe heterogeneity present in connections, a frame work(SocioDim) [2] has been proposed for collective behavior learning.

The framework SocioDim is composed of two steps:

1) SocialDimension Extraction
2) Discriminativelearning.

In the first step, latent social dimensions areextracted based on network topology to capture thepotential affiliations of actors. These extracted socialdimensions represent how each actor is involved in diverseaffiliations. One example of the social dimensionrepresentation is shown in Table 1. The entries in this tabledenote the degree of one user involving in an affiliation.

These social dimensions can be treated as features of actorsfor subsequent discriminative learning. Since a network isconverted into features, typical classifiers such as supportvector machine and logistic regression can be employed.

The discriminative learning procedure will determinewhich social dimension correlates with the targeted behavior and then assign proper weights.A key observation is that actors of the sameaffiliation tend to connect with each other. For instance, itis reasonable to expect people of the same department tointeract with each other more frequently. Hence, to inferactors' latent affiliations, we need to find out a group ofpeople who interact with each other more frequently than atrandom. This boils down to a classic community detection problem. Since each actor can get involved in more thanone affiliation, a soft clustering scheme is preferred. In theinitial instantiation of the framework SocioDim, a spectralvariant of modularity maximization [3] is adopted toextract social dimensions. The social dimensionscorrespond to the top eigenvectors of a modularity matrix.

It has been empirically shownthat this frameworkoutperforms other representative relational learningmethods on social media data. However, there are severalconcerns about the scalability of SocioDim with modularitymaximization:

• Social dimensions extracted according to soft clustering,such as modularity maximization and probabilisticmethods, are dense. Suppose there are 1 million actors in anetwork and 1, 000 dimensions are extracted. If standard double precision numbers are used, holding the full matrix alone requires 1M × 1K × 8 = 8G memory. This large-sizedense matrix poses thorny challenges for the extraction of social dimensions as well as subsequent discriminative learning.

• Modularity maximization requires us to compute the top eigenvectors of a modularity matrix, which is the same size as a given network. In order to extract k communities, typically k – 1 eigenvectors are computed. For a sparse or structured matrix, the eigenvector computation costs O(h(mk + nk2 + k3)) time [13], where h, m, and n are the number of iterations, the number of edges in the network, and the number of nodes, respectively. Though computing the top single eigenvector (i.e., k=1), such as PageRank scores, can be done very efficiently, computing thousands of eigenvectors or even more for a mega-scale network becomes a daunting task.

• Networks in social media tend to evolve, with newmembers joining and new Connections occurring betweenexisting members each day. This dynamic nature ofnetworks entails an efficient update of the model forcollective behavior prediction. Efficient online updates ofeigenvectors with expanding matrices remain a challenge.

SocioDim framework is proposed to address the relation heterogeneity presented in social networks. Thus, asensible method for social dimension extraction becomescritical to its success. Briey, existing methods to extractsocial dimensions can be categorized into node-view andedge-view.

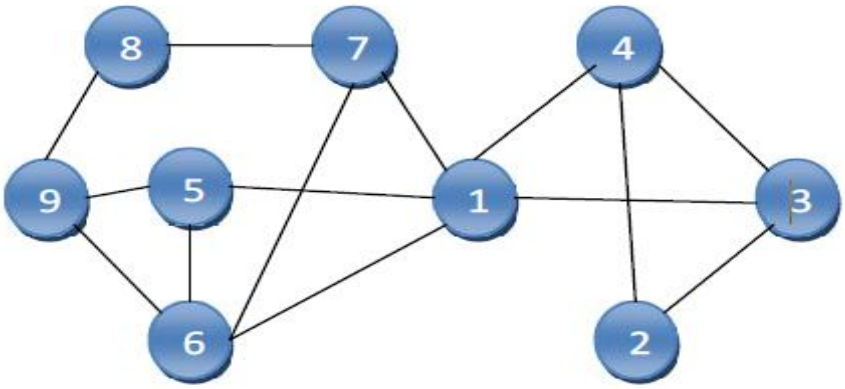


Fig: 1Sample network

➡ Node-view methods concentrate on clustering nodes ofa network into communities. As we have mentioned, theextraction of social dimensions boils down to a communitydetection task. The requirement is that one actor should beallowed to be assigned to multiple affiliations. Many existent community detection methods, with the

aim of partitioning the nodes of a network into disjoint sets, do not satisfy this requirement. Instead, a soft clustering scheme is preferred. Hence, variants of spectral clustering, modularity maximization, non-negative matrix factorization or block models can be applied. One representative example of node-view methods is modularity maximization [6]. The top eigenvectors of a modularity matrix are used as the social dimensions in [8].

Suppose we are given a toy network as in Figure 3, of which there are 9 actors, with each circle representing one affiliation. For k affiliations, typically at least k - 1 social dimensions are required. The top social dimension based on modularity maximization of the toy example is shown in Table 2. The actors of negative values belong to one affiliation, and actor 1 and those actors with positive values belonging to the other affiliation. Note that actor 1 is involved in both affiliations. Hence, actor 1's value is in between (close to 0). This social dimension does not state explicitly about the association, but presents degree of associations for all actors.

➡Edge-view methods concentrate on clustering edges of a network into communities. One representative edge-view method is proposed in [9]. The critical observation is that an edge resides in only one affiliation, though a node can be involved in multiple affiliations.

➡Reusable: SocioDim is composed of two parts:

- Community detection
- Supervised learning.

Both are well-studied. Many algorithms have been developed and numerous existing software packages can be plugged in instantaneously, enabling code reuse and saving many human efforts for practical deployment.

➡Efficient: A key difference of SocioDim framework from collective inference is that it is very efficient for prediction by trading more time in network pre-processing and training. Collective inference typically requires many scans of the whole network for prediction while SocioDim accomplishes the task in one shot.

VIII. CONCLUSION

In the previous sections, we have introduced the problem of Aggregate Performance prediction, covered a social learning framework based on social dimensions, in the present framework we introduced new edge centric based classification. Compared to existing algorithms it more advantageous and reduce time for unlabeled edge classification. We used incremental clustering for grouping the labelled edge, it is one of the best clustering algorithm. We tested theoretically and give best results.

REFERENCES

[1]. L. Tang and H. Liu, "Toward predicting collective behavior via social dimension extraction," *IEEE Intelligent Systems*, vol. 25, pp. 19–25, 2010.

[2]. ——, "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.

[3]. 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

[4]. L. Tang and H. Liu, "Scalable learning of collective behaviour based on sparse social dimensions," in *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2009, pp. 1107–1116.

[5]. P. Singla and M. Richardson, "Yes, there is a correlation: - from social networks to personal behavior on the web," in *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008, pp. 655–664.

[6]. 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.

[7]. 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*. New York, NY, USA: ACM, 2005, pp. 1371–1374.
[8]. H. W. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas, "Homophily in the digital world: A LiveJournal case study," *IEEE Internet Computing*, vol. 14, pp. 15–23, 2010.
[9]. S. A. Macskassy and F. Provost, "Classification in networked data: A toolkit and a univariate case study," *J. Mach. Learn. Res.*, vol. 8, pp. 935–983, 2007.
[10]. X. Zhu, "Semi-supervised learning literature survey," 2006. [Online]. Available: http://pages.cs.wisc.edu/∼jerryzhu/pub/ssl survey 12 9 2006.pdf
[11]. L. Getoor and B. Taskar, Eds., *Introduction to Statistical Relational Learning*. The MIT Press, 2007.
[12]. X. Zhu, Z. Ghahramani, and J. Lafferty, "Semisupervised learning using gaussian fields and harmonic functions," in ICML, 2003.
[13]. S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in SDM, 2005.
[14]. M. Newman, "Power laws, Pareto distributions and Zipf's law," Contemporary physics, vol. 46, no. 5, pp.323–352, 2005.
[15]. F. Harary and R. Norman, "Some properties of line digraphs," Rendiconti del Circolo Matematico di Palermo, vol. 9, no. 2, pp. 161–168, 1960.
[16]. T. Evans and R. Lambiotte, "Line graphs, link partitions, and overlapping communities," Physical Review E, vol. 80, no. 1, p. 16105, 2009.
[17]. Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multi-scale complexity in networks," 2009. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:0903.3178
[18]. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," Journal of Machine Learning Research, vol. 9, pp. 1871–1874, 2008.
[19]. J. Hopcroft and R. Tarjan, "Algorithm 447: efficient algorithms for graph manipulation," Commun. ACM, vol. 16, no. 6, pp. 372–378, 1973.