IOSR IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. II (Mar-Apr. 2014), PP 75-79
www.iosrjournals.org

Collective Behavior of social Networking Sites

Ashwini Vispute, Prerna Jadhav, Prof. P. V Kharat
Dept. Information Technology Jspm's BSIOTR(W), wagholi. Pune,India
Dept. Information Technology Jspm's BSIOTR(W), wagholi. Pune,India
Dept. Information Technology Jspm's BSIOTR(W), wagholi. Pune, India

Abstract: *Now a days a huge data is generated by social media like Facebook, Twitter, Flickr, and YouTube .This big data present opportunities and challenges to study collective behavior of data. In this work, we 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 has been shown effective in addressing the heterogeneity of connections presented in social media. However, 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, we propose 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.*

Index Terms: *Classification with Network Data, Collective Behavior, Community Detection, Social Dimensions.*

I. Introduction

The advancement in computing and communication technologies enables people to get together and share information in innovative ways. Social networking sites (a recent phenomenon) empower people of different ages and backgrounds with new forms of collaboration, Communication, and
collective intelligence. Prodigious numbers of online volunteers collaboratively write encyclopedia articles of unprecedented scope and scale; online marketplaces recommend products by investigating user shopping behavior and interactions; and political movements also exploit new forms of engagement and collective action. In the same process, social media provides ample opportunities to study human interactions and collective behavior on an unprecedented scale. In this work, 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.

The original framework, however, is not scalable to handle networks of colossal sizes because the extracted social dimensions are rather dense. In social media, a network of millions of actors is very common. With a huge number of actors, extracted dense social dimensions cannot even be held in memory, causing a serious computational problem. Specifying social dimensions can be effective in eliminating the scalability bottleneck. In this work, we propose an effective edge-centric approach to extract sparse social dimensions [4]. We prove that with our proposed approach, sparsity of social dimensions is guaranteed. Extensive experiments are then conducted with social media data. The framework based on sparse social dimensions, without sacrificing the prediction performance, is capable of efficiently handling real-world networks of millions of actors.

II. Collective Behavior

Collective behavior refers to the behaviors of individuals in a social networking environment, but it is not simply the aggregation of individual behaviors. In a Connected environment, individuals' behaviors tend to be interdependent, influenced by the behavior of friends. This naturally leads to behavior correlation between connected users [5]. Take marketing as an example: if our friends buy something, there is a better-than-average chance that we will buy it, too.

This behavior correlation can also be explained by homophily [6]. Homophily is a term coined in the 1950s to explain our tendency to link with one another in ways that confirm, rather than test, our core beliefs. Essentially, we are more likely to connect to others who share certain similarities with us. This phenomenon has been observed not only in the many processes of a physical world, but also in online systems [7], [8]. Homophily results in behavior cor-relations between connected friends. In other words, friends in a social network tend to behave similarly.

The recent boom of social media enables us to study collective behavior on a large scale. Here, behaviors include a broad range of actions: joining a group, connecting to a person, clicking on an ad, becoming interested in certain topics, dating people of a certain type, etc. In this work, we attempt to leverage the behavior correlation presented in a social network in order to predict collective behavior in social media. Given a network with the behavioral information of some actors, how can we infer the behavioral outcome of the remaining actors within the same network?

III. Social Dimensions

Connections in social media are not homogeneous. People can connect to their family, colleagues, college classmates, or buddies met online. Some relations are helpful in determining a targeted behavior (category) while others are not. This relation-type information, however, is often not readily available in social media. A direct application of collective inference [9] or label propagation [12] would treat connections in a social network as if they were homogeneous. To address the 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) social dimension extraction, and 2) discriminative learning. In the first step, latent social dimensions are extracted based on network topology to capture the potential affiliations of actors. These extracted social dimensions represent how each actor is involved in diverse affiliations. One example of the social dimensions representation is shown in Table 1. The entries in this table denote the degree of one user involving in an affiliation. These social dimensions can be treated as features of actors for subsequent discriminative learning. Since a network is converted into features, typical classifiers such as support vector machine and logistic regression can be employed. The discriminative learning procedure will determine which social dimension correlates with the targeted behavior and then assign proper weights.

A key observation is that actors of the same affilia-tion tend to connect with each other. For instance, it is reasonable to expect people of the same department to interact with each other more frequently. Hence, to infer actors' latent affiliations, we need to find out a group of people who interact with each other more frequently than at random. This boils down to a classic community detection problem. Since each actor can get involved in more than one affiliation, a soft clustering scheme is preferred.

In the initial instantiation of the framework So-cioDim, a spectral variant of modularity maximization [3] is adopted to extract social dimensions. The social dimensions correspond to the top eigenvectors of a modularity matrix. It has been empirically shown that this framework outperforms other representative relational learning methods on social media data. However, there are several concerns about the scald-ability of SocioDim with modularity maximization:

Social dimensions extracted according to soft clustering, such as modularity maximization and probabilistic methods, are dense. Suppose there are 1 million actors in a network 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-size dense matrix poses thorny challenges for the extraction of social dimensions as well as subsequent discriminative learning.

Networks in social media tend to evolve, with new members joining and new connections occurring between existing members each day. This dynamic nature of networks entails an efficient update of the model for collective behavior pre-diction. Efficient online updates of eigenvectors with expanding matrices remain a challenge.

IV. Algorithm

A] Edge-Centric Clustering Algorithm

1. function HCS (G (V, E))
2. (H1, H2, C) MINIMUMCUT (G)
3. if G is highly connected
4. then return (G)
5. else
6. HCS (H1)
7. HCS (H2)
8. end if
9. end

B] K-Means Variant Algorithm

1. Select k center in the problem space (it can be random)
2. Partition the data into k clusters by grouping points that are closest to that k centers.
3. Use the mean of these clusters to find new centers.
4. Repeat steps 2 and 3 until centers do not change.

C] Scalable Learning Of Collective Behavior Algorithm

1. Convert network into edge-centric view.
2. Perform edge clustering.
3. Construct social dimensions based on edge partition node belongs to one community as long as any of its neighboring edges is in that community.
4. Apply regularization to social dimensions.
5. Construct classifier based on social dimensions of labeled nodes.
6. Use the classifier to predict labels of unlabeled ones based on their social dimensions.

V. Sparse Social Dimensions

In this section, we first show one toy example to illustrate the intuition of communities in an "edge" view and then present potential solutions to extract sparse social dimensions.

5.1 Communities in an Edge-Centric View

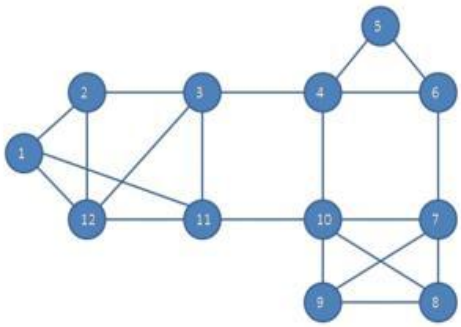


Fig 1: Clusters from Circuit

Though SocioDim with soft clustering for social dimension extraction demonstrated promising results, its scalability is limited. A network may be sparse (i.e., the density of connectivity is very low), whereas the extracted social dimensions are not sparse. Let's look at the toy network with two communities in Figure 1. Its social dimensions following modularity maximization are shown in Table 2. 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. The corresponding memory requirement hinders both the extraction of social dimensions and the subsequent discriminative learning. Hence, it is imperative to develop some other approach so that the extracted social dimensions are sparse.

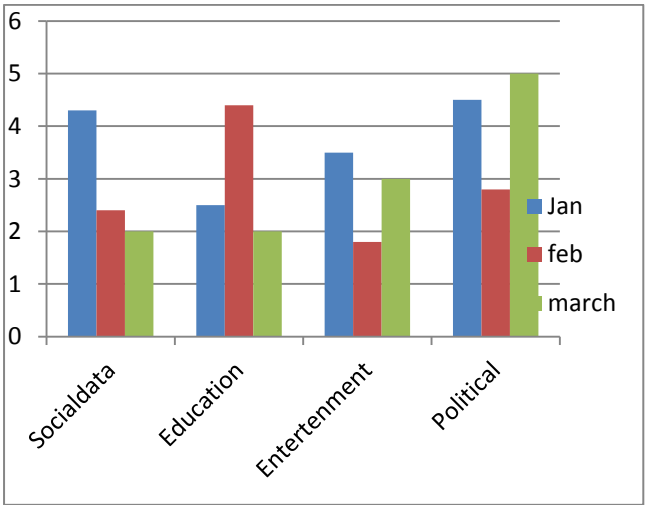
VI. Experiment Results

In this section, we first examine how prediction performances vary with social dimensions extracted following different approaches. Then we verify the sparsity of social dimensions and its implication for scalability. We also study how the performance varies with dimensionality. Finally, concrete examples of extracted social dimensions are given.

6.1 Prediction Performance

The prediction performance on all data is shown in Tables 5-7. The entries in bold face denote the best performance in each column. Obviously, EdgeCluster is the winner most of the time. Edge-centric clustering shows comparable performance to modularity maximization on BlogCatalog network, yet it outperforms ModMax on Flickr. ModMax on YouTube is not applicable due to the scalability constraint. Clearly, with sparse social dimensions, we are able to achieve comparable performance as that of dense social dimensions.

6.2 Our System Result



VII. Related Work

Classification with networked instances are known as within-network classification [9], or a special case of relational learning . The data instances in a network are not independently identically distributed (i.i.d.) as in conventional data mining. To capture the correlation between labels of neighboring data objects, typically a Markov dependency assumption is assumed. That is, the label of one node depends on the labels (or attributes) of its neighbors. Normally, a relational classifier is constructed based on the relational features of labeled data, and then an iterative process is required to determine the class labels for the unlabeled data. The class label or the class membership is updated for each node while the labels of its neighbors are fixed. This process is repeated until the label inconsistency between neighboring nodes is minimized. It is shown that [9] a simple weighted vote relational neighbor-hood classifier works reasonably well on some benchmark relational data and is recommended as a baseline for comparison.

However, a network tends to present heterogeneous relations, and the Markov assumption can only capture the local dependency. Hence, researchers propose to model network connections or class labels based on latent groups. A similar idea is also adopted in [2] to differentiate heterogeneous relations in a network by extracting social dimensions to represent the potential affiliations of actors in a network. The authors suggest using the community membership of a soft clustering scheme as social dimensions. The extracted social dimensions are treated as features, and a support vector machine based on that can be constructed for classification. It has been shown that the proposed social dimension approach significantly outperforms representative methods based on collective inference.

There are various approaches to conduct soft clustering for a graph. Some are based on matrix factorization, like spectral clustering and modularity maximization [3]. Probabilistic methods are also de-veloped . Please refer to for a compre-hensive survey. A disadvantage with soft clustering is that the resultant social dimensions are dense, posing thorny computational challenges.

Another line of research closely related to the method proposed in this work is finding overlapping communities. Palla et al. propose a clique percolation method to discover overlapping dense communities. It consists of two steps: first find out all the cliques of size k in a graph. Two k-cliques are connected if they share k 1 nodes. Based on the connections between cliques, we can find the connected components with respect to k-cliques. Each component then corresponds to one community. Since a node can be involved in multiple different k-cliques, the resultant community structure allows one node to be associated with multiple different communities. A similar idea is presented in , in which the authors propose to find out all the maximal cliques of a network and then perform hierarchical clustering.

Gregory extends the Newman-Girvan method to handle overlapping communities. The original Newman-Girvan method recursively removes edges with highest betweenness until a network is separated into a pre-specified number of disconnected components. It outputs non-overlapping communities only. Therefore, Gregory proposes to add one more action (node splitting) besides edge removal. The algorithm recursively splits nodes that are likely to reside in multiple communities into two, or removes edges that seem to bridge two different communities. This process is repeated until the network is disconnected into the desired number of communities. The aforementioned methods enumerate all the possible cliques or shortest paths within a network, whose computational cost is daunting for large-scale networks.

Recently, a simple scheme proposed to detect over-lapping communities is to construct a line graph and then apply graph partition algorithms. However, the construction of the line graph alone, as we discussed, is

prohibitive for a network of a reason-able size. In order to detect overlapping communities, scalable approaches have to be developed.

VIII. Conclusions And Future Work

Evaluating user preferences of web search results is crucial for search engine development, deployment, and maintenance. We present a real-world study of modeling the behavior of web search users to predict web search result preferences. Accurate modeling and interpretation of user behavior has important applications to ranking, click spam detection, web search personalization, and other tasks. Our key insight to improving robustness of interpreting implicit feedback is to model query-dependent deviations from the expected "noisy" user behavior. We show that our model of click through interpretation improves prediction accuracy over state-of-the-art click through methods. We generalize our approach to model user behavior beyond click through, which results in higher preference prediction accuracy than models based on click through information alone. We report results of a large-scale experimental evaluation that show substantial improvements over published implicit feedback interpretation methods.

Acknowledgments

We are done project on collective behaviour of social networking sites our college Jspm's BSIOTR.

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 behavior 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, "Ho-mophily 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 Rela-tional Learning. The MIT Press, 2007.
[12] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised 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.