Collective Behaviour Prediction Via Social Dimensions Extraction

Mrs. Kanchan Jadhav
SKNCOE Pune,
University of Pune.

Prof. Nalini Mhetre
Dept. Of computer Engg.
SKNCOE Pune,
University of Pune

Abstract

Day by day the clicks are increasing in a particular network. Behaviour is nothing but to know the interest and requirement of users. Collective behaviour, which indicates the group of data generated on a large scale. In this framework affiliations of actors are capture by extracting social dimensions and then classify the actors using extracted dimensions. As existing approaches to extract social dimensions are not scalable and can not handle network of huge size. We solve these problem by sparsifying social dimension to make this extraction scalable by using edge-centric clustering scheme and k-means variant algorithm. In social media, multiple modes of actors can be involved in the same network, resulting in a multimode network. In this work, we attempt to harness the predictive power of social connections to determine the preferences or behaviours of individuals such as whether a user supports a certain political view, whether one likes one product, whether he/she would like to vote for a presidential candidate, etc.

Keywords: Collective Behaviour, Affiliations, Scalable learning, Edge Clustering, k-means variant.

1. Introduction

Human intention is the valuable source in the present world. As the trends in the technology are increasing, there is a need to know and resolve the issues of the visitors who leave their footprints in their fascinated areas. We have so many social networking websites. If analysis is made on the fascinated areas of different people, the united behaviour of the people can be identified. In Social networking sites and enterprises collaborative software coupled with advancement in computing communication technologies are enabling people to share information in innovative ways from wikipedia, facebook to microsoft share points the numerous advancement which are helping using sharing information. These tools and software provides ampere opportunities to study human Interactions and collective behaviour on unprecedented scale[1]. In social media usually network consist of millions of actors. Hence size of extracted social media is huge. It is not possible to hold such amount of huge information in RAM and hence it is extremely difficult to do mathematical operations. These problems are solve by sparsifying social dimension to make this extraction scalable. In this work we present edge centric approach to extract sparse social dimensions[2].

In this work, we study how networks in social media can predict some human behaviour and individual performances, that means when behaviour of some Individuals in social media is given we can conclude or predict behaviour of other Individuals in networks.

Connections in social media networks are not homogeneous for e.g. One user may have multiple connections. One for his friend, family, classmates and colleagues. Unfortunately this relative information is not fully available in reality. We also don’t know why the users are connected with each other. To study this heterogeneity a new framework is designed and it is called as “social dimensions”. as shown in Fig 1.

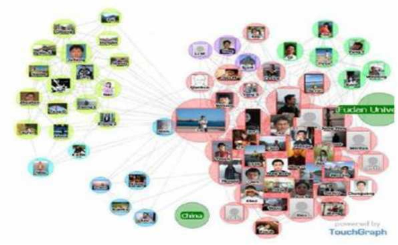

Fig 1: Contacts of one user in a Social Network

2. Collective Behaviour

Collective behaviour is a behaviour of individuals in social networking environment but it is not addition of individual behaviours[3]. In connectual environment behaviour of one person is usually influenced by behaviour of friend. Another observation is that actors of same affiliation tend to connect with each other for e.g. If a person likes shoes then chances of some one who likes football can also likes shoes. it is not simply the aggregation of individual behaviours. This study of collective behaviour is to understand how individuals act in a social networking environment. Our approach follows a social-dimension based learning framework. Social dimensions are extracted to represent the potential affiliations of actors before discriminative learning occurs.

3. Sparse Social Dimensions

The latest trend in the social network enables us to study intersection of behaviour or interests on a large scale. The interests include connecting a person, clicking an add, joining in a group, marking friends as buddies, becoming a fan of a celebrity, searching for friends, sending gifts for their special days etc. We have a given network with the behavioural outcome of the remaining users within the same network[3].

Existing methods to extract social dimensions can be categorized into node-view and edge-view. Node-view methods concentrate on clustering nodes of a network into communities. It take one node and their affiliation and find out in how many communities that node is involved. So it required larger memory space for these all nodes.

Edge-view methods concentrate on clustering edges of a network into communities. It find out in how many communities that edge is involved

We implement an edge centric view using k-means clustering by using following methodologies. Table 1 shows how affiliation is represented.

Table1: shows how an affiliation is represented

Actors	Affiliation-1	Affiliation-2	.....	Affiliation-k
1	0	1	.....	0.8
2	0.5	0.3	.....	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

4. Communities In An Edge Centric View

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. SocioDim with soft clustering for social dimension extraction verified promising results, but its limitation is scalability. A network may be sparse (i.e., the density of connectivity is very low), whereas the extracted social dimensions are not sparse.

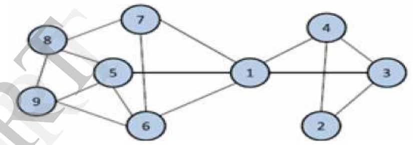

Fig 2: Toy example

Let’s look at the toy network example with two communities in Figure 3. Its social dimensions following modularity maximization are shown in Table 1. There is no one of the entries is zero. Then extracting dimensions from the large network may be in big number and maintaining these many dimensions also a big problem to the persons who are taking care of this work. And the memory required to do this is also a problem. Maintainability is also one of the problem for this.

Hence, we need minimum measurable dimensions. Hence, it is imperative to develop some other approach so that the extracted social dimensions are sparse.

Table2:Social dimension of Toy example

Actors	Modularity Maximization	Edge- centric Clustering	
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

An actor is considered associated with one relationship if one of his connections is assigned to that affiliation. There are the two communities in Figure 2. In Figure 3 the dashed edges represent one affiliation, and the remaining edges denote the second affiliation.

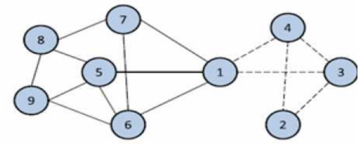

Figure 3: Edge Clusters

In Table 2, where an entry is 1 (0) if an actor is (not) involved in that corresponding social dimension. To extract sparse social dimensions, we partition edges into disjoint sets rather instead of nodes. In addition, the extracted social dimensions following edge partition are guaranteed to be sparse. This is because the number of one's affiliations is no more than that of her connections. The density of extracted social dimension is find by following theorem.

$$density \le \frac{\sum_{i=1}^{n} \min(d_i, k)}{nk}$$

$$= \frac{\sum_{\{i|d_i \le k\}} d_i + \sum_{\{i|d_i > k\}} k}{nk}$$

Where k is number of social dimensions to be extracted, m is number of edges, n is number of nodes and di is the degree of node. Moreover, for many real-world networks whose node degree follows a power law distribution, the upper bound in above equation can be approximated as bellow.

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

Where α > 2 is the exponent of the power law distribution.

5. Edge Partition Via Line Graph

In order to partition edges into disjoint sets, one way is to look at the "dual" view of a network, i.e., the line graph. A graph represented communities in a network can be drawn with respect to edge partition based on edge connections. An edge represents the connectivity of two vertices. Hence an edge based connections are very useful to calculate approximately the relationships of a user with different communities.

In a line graph L(G), each node corresponds to an edge in the original network G, and edges in the line graph represent the adjacency between two edges in the original graph. The line graph of the toy example is shown in Figure 4.

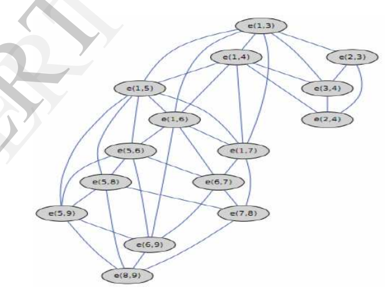

Figure 4: The line graph of Toy example

For instance, e(1, 3) and e(2, 3) are connected in the line graph as they share one terminal node 3. They are adjacent to each other. Each node in the line graph corresponds to an edge in the original graph. Following equation is use to increase many more edges in the graph.

$$N = m, \quad M \geq m\left(\frac{2m}{n} - 1\right)$$

Where n denotes the number of nodes, m denotes number of connections in a network. N & M denotes number of nodes and connections in its line graph respectively.

6. Edge Partition Via Clustering Edge Instances

By using edge centric view it easy to identify which nodes are connected each other .Then we apply edge clustering methods for finding the similarity between the edges. Edge-centric clustering essentially treats each edge as one data instance with its ending nodes . Then a typical k-means clustering algorithm can be applied to find out disjoint partitions. One apprehension with this scheme is that the total number of edges might be too huge.
Then, clustering algorithm like k-means clustering can be applied to Table 3 to find disjoint partitions[1].

Table3: Edge centric view

Edge	Feature
	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
.	…….
.	…..
.	…

K-Means clustering can be implemented abiding following algorithm

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

Output: $\{idx_i\}$

Procedure:

```
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 $|1 \leq i \leq m\}$, number of clusters k
    update centroid $C_{idxi}$
until change of objective value < $\varepsilon$
```

Algorithm of Scalable k-means Variant

As a simple k-means is used to extract social dimensions, it is easy to update social dimensions if a given network changes. If a new member joins the network and a new connection emerges, we can simply assign the new edge to the corresponding clusters. The updation of centroids when the new connection is appear also straightforward. This k-means scheme is mainly appropriate for dynamic large scale networks[1].

The detailed algorithm is summarized.

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

Output: labels of unlabeled nodes.

Procedure:

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

Algorithm for Learning of Collective Behaviour

7. Regularization On Communities

After clustering of the edges we construct classifier based on the social dimensions. We designed a new classifier

The extracted social dimensions are treated as features of nodes. In order to handle large-scale data with high dimensionality and vast numbers of instances, we adopt a linear SVM, which can be finished in linear time generally; the larger a community is, the weaker the connections within the community are. so we build an SVM classifier for the classification.

8. Conclusion

We propose an edge-centric clustering scheme to extract social dimensions and a scalable k-means variant to handle edge clustering. ,in the propose 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.

With this edge-centric view, we guaranteed that the extracted social dimensions are to be sparse. 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 feasible solution to effective learning of online collective behaviour on a large scale.

In social media, multiple modes of actors can be involved in the same network, resulting in a multi-mode network. For instance, in YouTube, users, videos, tags, and comments are intertwined with each other in co-existence. Extending the edge-centric clustering scheme to address this object heterogeneity can be a promising future direction.

9. References

[1] Lei Tang, Xufei Wang, and Huan Liu, Scalable Learning of Collective Behavior, 2012.
[2] L. Tang and H. Liu, “Toward predicting collective behavior via social dimension extraction,” IEEE Intelligent Systems, vol. 25,pp. 19–25, 2010
[3] 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.
[4] M. Newman, “Power laws, Pareto distributions and Zipf’s law,” Contemporary physics, vol. 46, no. 5, pp. 323–352, 2005.
[5] T. Evans and R. Lambiotte, “Line graphs, link partitions, and overlapping communities,” Physical Review E, vol. 80, no. 1, p.16105, 2009
[6] L. Tang, S. Rajan, and V. K. Narayanan, “Large scale multilabel classification via metalabeler,” in WWW ’09: Proceedings of the 18th international conference on World Wide Web. New York, NY, USA: ACM, 2009
[7] M. Newman, “Finding community structure in networks using the eigenvectors of atrices,” Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), vol. 74, no. 3, 2006.
[8] X. Zhu, “Semi-supervised learning literature survey,” 2006.
[9] F. Sebastiani, “Machine learning in automated text categorization,” ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, 2002.
[10] S. Fortunato, “Community detection in graphs,” Physics Reports, vol. 486, no. 3-5, pp. 75-
174, 2010.
[11] J. Bentley, “Multidimensional binary search trees used for associative searching,” Comm.
ACM, vol. 18, pp. 509–175, 1975.
[11] 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, vol. 24, pp. 881–892, 2002.
[12] P. Bradley, U. Fayyad, and C. Reina, “Scaling clustering algorithms to large databases,” in ACM KDD Conference, 1998.