AN IMPROVISED MODEL FOR IDENTIFYING INFLUENTIAL NODES IN MULTI-PARAMETER SOCIAL NETWORKS

Abhishek Singh[1] and A. K. Agrawal[2]

[1,2]Department of Computer Engineering, IIT(BHU) Varansi
abhishek.singh.cse09@iitbhu.ac.in
akagrawal.cse@iitbhu.ac.in

ABSTRACT

Influence Maximization is one of the major tasks in the field of viral marketing and community detection. Based on the observation that social networks in general are multi-parameter graphs and viral marketing or Influence Maximization is based on few parameters, we propose to convert the general social networks into "interest graphs". We have proposed an improvised model for identifying influential nodes in multi-parameter social networks using these "interest graphs". The experiments conducted on these interest graphs have shown better results than the method proposed in [8].

KEYWORDS

Viral Marketing, Community Detection, Influence Maximization

1. INTRODUCTION

In today's era of the internet, the enormous growth and penetration of social networks into people's daily lives has brought a number of opportunities and challenges. It is not only a way to connect to the rest of the world but has also become an integral part of business, economy, politics and almost all such fields.

Identifying community structure [10] has been a central area of research in identifying groups in the network based on multiple factors such as common interests, friendship, organizations etc among the 'actors'. These communities are important as they can be viewed as a platform for sharing knowledge, data, emotions, sentiments etc. Another application of social network analysis has been viral marketing [] which is a very crucial area of research in business analytics. Viral marketing is an advertisement technique where one identifies a subset of 'actors' of the social network so as to obtain a "word of mouth" effect in promoting the product.

A major task in tackling both these problems is identifying influencers in the network, as whether it is the survival of a community or spreading of an innovation/idea/product in a network, *there's always a need for certain 'actors' who have influence over the rest of the community.* This can be seen in viral marketing as companies trying to identify a seed set of individuals to introduce their product so as to have as much spread of word as possible. In terms of community detection, a

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 25–30, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4503

certain set of individuals that share interests can be seen influencing each other. For example, an individual who shares interests in terms of his movie preferences with another individual would be compelled to watch a new movie if he/she sees positive feedback from the other. When this happens, the community detection algorithms increases the parameter used to represent the common interests among a certain set of individuals which results in detecting the community. So, although both community detection and viral marketing are different areas, the underlying problem in both the cases is the same. This is the problem of maximizing the spread of influence in a network.

Formally, the problem of Influence maximization involves finding few initial users in an online social network to adopt an innovation and spread the information, so that the influence of innovation or product in the network is maximized. Influence maximization is a problem applied not only to tasks related to social networks but can be used for different other applications.'

Finding these 'few initial users' in these large networks is the major challenge of the problem and for which, huge amount of data is needed to be processed. Not only the processing of huge amount of data is required, timeliness of the processes are also important. For this, the time complexity of the process should be small. The most popular approaches in this area are greedy algorithm and/or optimizations to the greedy algorithms.

It was observed that the individuals are connected to many others based on different interests. So a product/idea/event, which is to be 'spread' in the network, belong to a particular community of the individual only. So, instead of targeting the entire network to find 'influential seeds', one can find the community first, where the probability of spreading is high and subsequently finding the 'seed'/'seeds'. In this paper, we propose an improvement of the three methods for choosing the seeds by using community detection methods. The approaches are discussed in the next sections. In our approach, we emulate relationship between community detection and viral marketing.

The rest of the paper is organized as follows: section 2 discusses about the related works, section 3 describes our method, section 4 provides experimental results and section 5 concludes and discusses the future aspect of the work.

2. RELATED WORK

The concept of spreading of an idea, or an innovation or influence for that matter was first studied in the field of economy, giving birth to viral marketing. Several models were proposed to simulate this process. There are two models in particular have gained widespread acceptance. These are the Linear Threshold Model and the Independent Cascade Model. The Linear Threshold model states that every node in a network has some threshold which is needed to be achieved after which it can become active. The Independent Cascade model on the other hand gives a probabilistic methodology to this. It proposes that any active node in a network would get a dingle chance to activate an inactive node, which it can do with certain probability. The problem of Influence Maximization was first studied by Domingos and Richardson [1][2]. Although there attempts at solving this problem were probabilistic. Since Domingos and Richardson studied the problem of Influence Maximization, the other researchers [5][6][7] have proposed greedy approaches rather than the probabilistic approach suggested by the formers. In [3], Kempe et.al, proposed a greedy approach to solve what they viewed as a discrete

optimization problem. After this several modifications and improvements have been proposed over this original approach.

In 2010 Tieyun Qian et.al [8] proposed a different approach towards studying the same problem by identifying seed nodes in implicit social networks. The suggested approach uses the Reverse Nearest Neighbours logic presented by [4], and build on that by defining Social Network Potential of an individual in a network. Inspired by this idea, we propose a new model to identify influential nodes and understand the spread of influence in a network where the connections between actors incorporate various factors such as the reasons for these connections and how strong these connections are.

3. PROPOSED METHOD

In this work we propose a model based on the method of [8]. As discussed in section 1 it was observed that the actors in the social networks are connected to each other for various interest/reasons. So a product/idea/event, which is to be 'spread' in the network, may be of interest to a particular set of individuals only. Here we propose a model for the Influence Maximization problem taking into account the specific area which may be of interest to one who wants to influence the network, out of the available areas.

Generally, the social networks are represented as graphs where the nodes are actors and the edges represent the connections between these actors. This representation of the edge is a accumulation of multiple parameters on which the social network is based. These parameters are of diverse nature like location, interests, likes etc. The application of any method traditionally involves considering all the above mentioned parameters instead of focusing on a particular interest. So in our model we converted the traditional social graph into “interest graph”, which represents a particular interest(s) of the network. By this the volume of the network and the parameters of the network are reduced subsequently bringing down the time complexity and the computational overhead of applying the method on the original graph. An overview of the proposed model is shown in Figure 1.

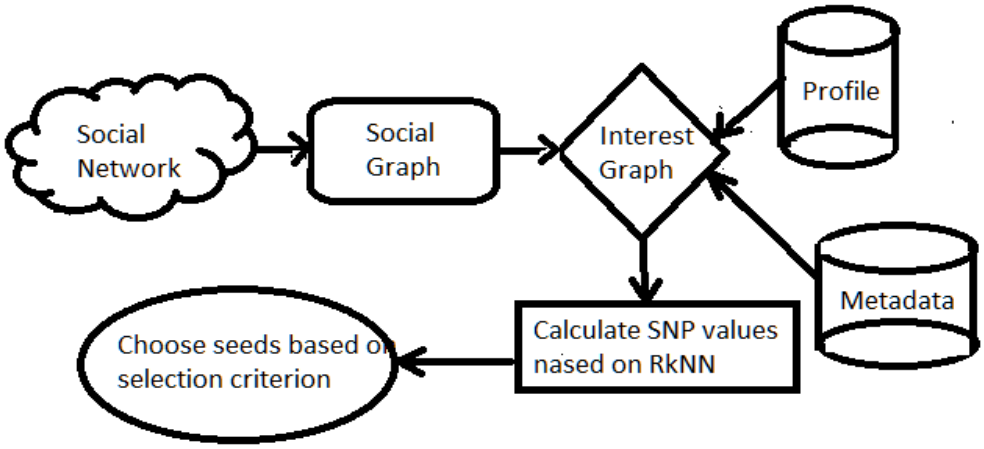

Figure 1. Pictorial description of the proposed model

The model shows a social network, for example a blogger network. Now actors in a blogger network may be connected to each other through links that represent multiple shared interests.

The graph extracted from such network has edges that account for a number of parameters which may or may not be of interest. Hence it is required that the graph be filtered to form an interest graph containing edges with weights specific to the interest. Once we have this graph we can move forward with the basic algorithm as proposed by [8]

4. EXPERIMENT

The data used in [11], forms the basis of our experiments. The data contains information about blogs and bloggers from a particular organization. The author of the blog use different tags for each blog entry they make. These blogs are represented as a graph where the blogs are the nodes and the edges are defined as the relationship between the blogs. This relationship is defined by the common tags giving weightage to particular tag which in this case represents the interests of the authors.

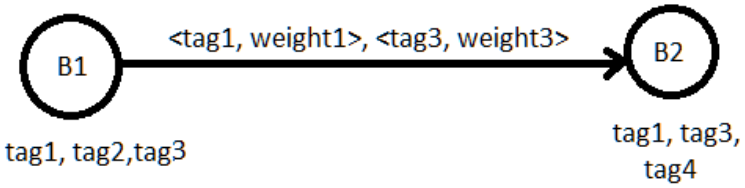

Figure 2. An example edge in the multi-parameter network

Any graph which represents this blog network can be defined by representing the blogs as nodes and the tags can be used to represent the weightage between the nodes. For converting this graph into "interest graph" we have given weightage to a particular tag in which one is interested. For example, let there be two blogs b1 and b2 having the tags{t1,t2,t3} and {t1, t4} respectively as shown in fig 2, then the nodes represent b1 and b2, and the weightage of the edge between b1 and b2, will be calculated as:

The graph so prepared is of 500 nodes and 87436 edges, by taking the first 500 nodes and their connections from the original dataset. The weights of this newly formed graph are then used as similarity measure for the nearest neighbourhood algorithm [4]. Then we use the methods explained in[8], to calculate the SNP values for the actors and find the desired set of seed nodes.

5. RESULTS

Table 1. Top 5 users for k=1

User-Id	R1NN	SNP
#211	8 22 41 45 61 100 107 137 157 169 171 175 186 209 227 247 249 280 295 296 297 298 312 316 337 359 380 391 429 435 437 455 461 481	34
#12	5 8 9 23 34 40 41 108 114 129 162 169 171 209 216 217 229 232 243 245 297 311 337 345 385 391 408 428 455 481	32

#119	2 29 41 142 169 171 209 217 239 243 245 247 263 271 297 300 337 355 368 370 383 391 401 409 417 421 425 454 455 481 485	31
#65	2 26 33 39 43 79 82 115 122 136 203 205 240 243 250 264 271 295 311 333 336 347 377 398 399 472 475 485	28
#327	4 16 23 27 32 47 56 92 94 103 129 130 154 163 181 182 186 197 217 228 283 339 375 385 418 427 457	27

6. CONCLUSION

In this work, we have proposed an improvised model for Influence Maximization in a general social network with edges having weights that are combination of multiple parameters. In such networks, connection between any two actors may be seen due to multiple shared interests. Thus, application of any algorithm on such a graph directly would not be completely realistic. Our work has addressed this problem by isolating these parameters by creating "interest graphs".

The results so obtained are not very different from those obtained on the original graph, but the isolation of parameters has shown improvements in terms of the computational and time complexity requirements of the algorithm. Although the original algorithm would work perfectly in case of single parameter connections, we have proved that using "interest graphs" improves the accuracy and efficiency of the algorithm in multi parameter connection graphs.

As mentioned in the introduction, the fields of viral marketing and community detection are connected. Hence, as part of our future work we plan to use this hypothesis and work towards improving our model by incorporating community detection in this work.

REFERENCES

[1] P. Domingos & M. Richardson. "Mining the network value of customers", 2001
[2] M. Richardson & P. Domingos. "Mining Knowledge-Sharing Sites for Viral Marketing", 2002.
[3] D. Kempe, J. Kleinberg, & E. Tardos. "Maximizing the spread of influence through a social network", 2003
[4] S. M. Flip Korn & S. Muthukrishnan. "Influence sets based on reverse nearest neighbour queries", 2000
[5] W. Chen, Y. Wang, & S. Yang. "Efficient influence maximization in social networks", 2009.
[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos,J. VanBriesen, & N. Glance. "Cost-effective outbreak detection in networks".
[7] W. Chen, Y. Wang, & C. Wang. "Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks", 2010.
[8] Tieyun Qian & Jiangbo Liu"Influence Maximization through Identifying Seed Nodes from Implicit Social Networks", ICUIMC'10 proceedings
[9] en.wikipedia.org/wiki/Viral_marketing
[10] http://en.wikipedia.org/wiki/Community_structure

[11] Nitin Agarwal, Huan Liu, Lei Tang & Philip S. Yu "Identifying influential blogger in a community", in proceedings of the 1st International Conference on web search and data mining(WSDM '08), PP 207-218,Feb 11-12 2008, Stanford, California

AUTHORS

Abhishek Singh is a post graduate student at the department of computer engineering, IIT-BHU. His area of interests are social networks, data mining, graph theory.

Prof. A. K. Agrawal is the head of department of Department of Computer Engineering at IIT-BHU. His areas of interest include database systems, theory of computation, compiler design and graph theory.