ARTICLE IN PRESS
JID: KNOSYS [m5G;October 27, 2016;16:30]
Knowledge-Based Systems 000 (2016) 1–10
Contents lists available at ScienceDirect
Knowledge-Based Systems
journal homepage: www.elsevier.com/locate/knosys

Social network pruning for building optimal social network: A user perspective

Sumith N.[a,*], Annappa B.[a], Swapan Bhattacharya[b]

[a] Dept. of CSE, National Institute of Technology Karnataka, Surathkal, India
[b] Dept. of CSE, Jadavpur University Kolkata, India

ARTICLE INFO

Article history:
Received 31 March 2016
Revised 27 September 2016
Accepted 22 October 2016
Available online xxx

Keywords:
Social networks
Pruning
Small-world properties
Influence maximization
Viral marketing

ABSTRACT

Social networks with millions of nodes and edges are difficult to visualize and understand. Therefore, approaches to simplify social networks are needed. This paper addresses the problem of pruning social network while not only retaining but also improving its information propagation properties. The paper presents an approach which examines the nodal attribute of a node and develops a criterion to retain a subset of nodes to form a pruned graph of the original social network. To authenticate feasibility of the proposed approach to information propagation process, it is evaluated on small world properties such as average clustering coefficient, diameter, path length, connected components and modularity. The pruned graph, when compared to original social network, shows improvement in small world properties which are essential for information propagation. Results also give a significantly more refined picture of social network, than has been previously highlighted. The efficacy of the pruned graph is demonstrated in the information diffusion process under Independent Cascade (IC) and Linear Threshold (LT) models on various seeding strategies. In all size ranges and across various seeding strategies, the proposed approach performs consistently well in IC model and outperforms other approaches in LT model. Although, the paper discusses the problem with the context of information propagation for viral marketing, the pruned graph generated from the proposed approach is also suitable for any application, where information propagation has to take place reasonably fast and effectively.

1. Introduction

Social networks are becoming a pedestal for applications such as viral marketing, poll analysis, recommendation system and so on. In this paper, the social network is discussed with the context of viral marketing application. This application depends on the information propagation process seen in social networks. However, huge size of social network is a challenge to understand the information propagation process. Social networks with millions of users and connections among them will be difficult to visualize and understand. Moreover, its disconnected structure may result in numerous outliers. As a result, an information initiated at random node[1] may not possibly reach target nodes. In such situation, a reduced social network with the best properties still retained, can be more helpful than the original large social network.

Various network simplification concepts are already available [1,2]. However, when these are used to simplify the social network, it risks in losing certain properties that may makes simplified social network unsuitable for application of interest. Moreover, impact of removing users may result in disturbing graph properties, those that are essential for information propagation. The aim of the paper is to develop an approach to reduce the size of the social network to an optimal consumer network. Such a pruned graph should have small world properties [3] such as higher average clustering coefficient, lower diameter, lower average path length, fewer number of connected components and lower modularity for effective information propagation. The proposed approach also supports the claim earlier made by Faust [4] and Butt [5], which reinforced the theory that social networks have little or no social structure. The term social structure is explained in terms of the presence of properties that are not explained by nodal tendencies. The proposed approach demonstrates that nodal attributes such as user interactions can be used to simplify social networks.

The contributions in this paper are as follows:

1. Simplification of social network on the nodal attribute by retaining the small world properties.

* Corresponding author.
E-mail addresses: s.nireshwalya@gmail.com (Sumith N.), annappa@ieee.org (Annappa B.), bswapan2000@yahoo.co.in (S. Bhattacharya).

[1] The words node and user are used interchangeably in this article.

http://dx.doi.org/10.1016/j.knosys.2016.10.020
0950-7051/© 2016 Published by Elsevier B.V.

2. Increased information diffusion in the pruned network for Independent Cascade and Linear Threshold models.

Few of the applications where the framework can be used are listed below:

- *Finding the contribution level*: It can be used to understand the contribution level of the users. One can fetch and analyze the users who contribute to the network.
- *Viral Marketing*: In this application the social network can be used for promotion of products. Since the advertising expenditure is increasing day by day, companies use the social network as a good platform for advertisements. Proposed framework can help to identify the important contributors thus finding the most influential users in the network for endorsing the products.
- *Feed Ranking System*: In this application the system recommends information to the user based on his/her past choices. The proposed concept can be used to fetch recommendation from the valid users. Thus the system would reflect the ground truth more accurately.

2. Related works

Social networks have been studied in various domains such as predicting user behaviors, community detection, advertisement campaigns, spread of information, trending behavior and so on. To understand the scope of the proposed work, the related work is partitioned into two segments (i) centrality measure and (ii) approaches for pruning the network under various constraints.

2.1. Centrality measures for evaluating nodes

Most common centrality measures to quantify the centrality of a certain node in social networks are degree centrality, closeness centrality, betweeness centrality and have been studied by Hilbrich [6]. The degree centrality assumes that a node that has many direct connections is at the center of the network and plays a very important role in the information diffusion process. The second measure, namely the closeness centrality focuses on how close a node is to all other nodes in the network. The betweeness centrality assumes that if a node is more often in the shortest paths between other nodes, it is more central to the network. Eigenvector centrality is yet another metric for measuring a nodes popularity in a network. A node's eigenvector centrality is proportional to the sum of eigenvector centralities of all nodes directly connected to it [7].

There are also other metrics such as PageRank [8] and HITS [9] that rank the nodes individually based on their importance but may not as such be used for pruning process. In their basic form, PageRank and HITS, value a node according to the graph topology[10].

Concept of hub is prevalent in identifying key users. Users who are in a hub position are characterized by a great potential for communication and interaction within network [11]. This centrality measure, indicates the importance of certain key users in the network. However, in real world networks, users who are connected to large portion of the network, do not show significantly high activity count [12]. It is here, that the concept of the hub fails to understand the structure of the social system. Hubs and brokers concepts are also used for pruning the social network [13]. This is considered as a structural pruning technique where all users who are hubs and/or brokers are retained in the sub-graph. However, such as sub-graph may not guarantee the diffusion and adoption of information. Similar to identifying hubs, gatekeeper centrality approach [14] evaluates node on Shapley value. nodes that have the capability to disconnect graph is identified in this approach. In contrast, the proposed approach aims to identify nodes that are actively involved in network, thus facilitating information diffusion. Therefore, gatekeeper strategy is not suitable for information diffusion application.

Borgatti [15] introduced the concept of group centrality measure and used graph fragmentation to define it. Smith and White [16] introduced personalization concept to understand how central a user is to given subset in social networks. Further, Estrada [17] introduces sub graph centrality that characterizes the participation of node in all sub graph based on spectral feature. Also, core centrality measure is coined by Evenett and Borgatti [18] to evaluate the extent to which a network revolves around a sub- network in social network.

The various centrality measures have a fundamental property [5], that two isomorphic networks should always attain the same centrality value on their nodes. Hence, only for isomorphic graphs the central node continues to be in dominant position. But, when graph is reduced to its sub graph, central node may lose its importance. Therefore, other metrics to evaluate nodes have to be considered. Centrality measures can be used to identify the least important nodes to be pruned. However, the result of such pruning techniques would result in sub graphs that may not benefit application of interest in the intended sense.

2.2. Pruning the network under various constraints

Structural equivalence, first proposed by Lorrain and White [1], is used to simplify graphs by means of identifying sets of node with similar structural properties. Structural equivalence, since then been used to understand various applications [19–23]. An attempt to model interactions in social graph is attempted by Heidemann [11]. Their approach retains a link that has even a single interaction. However, pruning the graph on this concept will not fetch significant results. Therefore, efficient criteria to keep an activity link has to be addressed.

Lossy network simplification approach to simplify the graph by removing edges resulting in loss of connectivity is attempted [24]. A trade-off between simplicity and connectivity has to be made in this approach. Further, pruning strategy to maintain connectivity [25] is also discussed to keep important edges without losing connectivity. This work defines the best path function and based on this, edges are given priority. This approach is applicable to probabilistic graph, flow graph and distance graph.

Source to link flow strategy [26] is proposed for flow network to reduce the computation time. Triangular inequality strategy [27] is used to reduce computation time of pathfinder algorithm. In their subsequent work, MST-Pathfinder [28] that prunes the edges, to retain edges on minimum spanning tree also is implemented on various networks including social networks. Modularity concept [29] is proposed to reduce graph to core components with the constraint of maintaining modularity. The *cut sparsifiers* [30] based on connectivity concept for undirected graph is also proposed to understand the strength of connectivity of the graph. Serrano et al. [31] and Foti et al. [32], focus on weighted networks and select edges that represent statistically significant deviations with respect to a null model. An application of pruning the graph on connectivity constraint is also seen in [33].

An interesting application of pruning is also seen in the domain of relational schema for multiple databases [2]. The sub-graph ensemble pruning technique prunes the relational databases to find correlation among the multiple databases. In this approach informative substructures are maintained. The criteria used for selecting information substructures for databases are not suitable for pruning the social graph. Collective inference technique [34] produced networks of smaller size that facilitates improved performance through collective inference. The approach discusses a sub

Please cite this article as: S. N. et al., Social network pruning for building optimal social network: A user perspective, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.10.020