Chapter 13
Characterization of Online Social Network: A Case Study on Twitter Data

Paramita Dey
Government College of Engineering & Ceramic Technology, India

ABSTRACT

The rapid growth of internet with large number of social network sites makes it easy to interconnect people from all over the world on a shared platform. Social network can be represented by a graph, where individual users are represented as nodes/vertices and connections between them are represented as edges of the graph. As social network inherits the properties of graph, its characterization includes centrality and community detection. In this chapter we discuss three centrality measures and its effects for information propagation. We discuss three popular hierarchical community detection measures and make a comparative analysis of them. Moreover we propose a new ego-based community detection algorithm which can be very efficient in terms of time complexity for very large network like online social network. In this chapter, a network is formed based on the data collected from Twitter account using hashtag(#).

INTRODUCTION

With the advent of online social network in last few years, there is a paradigm shift in information propagation. In online social network data are stored in the form of a graph where users and their interaction patterns are reflected through the network. To maximize information propagation for reaching targeted groups of people through online social network, it is required to evolve certain strategies based on the properties of the network. For this purpose, characterization of online social network is essential.

Popularity of on-line social networks like Twitter, Facebook, WhatsApp, Instagram are increased day by day. Active participation of society in social network opens a new arena of research opportunities. These networks are dynamic in nature, large scale and complex. For a long time, sociologists and economic analysts worked in this field with off-line social network data. But, the data is now available from the on-line social networks which are characterized by 4 v's- large volume, velocity, veracity and

DOI: 10.4018/978-1-5225-6023-4.ch013

variety. These characteristics forced computer science researchers to come up with new tools and algorithms and analyze these networks effectively and efficiently.

The goal of this chapter is to develop a unified framework to model social networks effectively and efficiently. A social network is viewed as a collection of relations between social users and their interactions. In this chapter, two methodologies of graph analytics, centrality and community detection are used for the characterization of online social network. Centrality measures indicate the influential nodes of the network. Thus propagation through centrality measures as seed nodes can increase influence in the network which is important for viral marketing. Thus blocking of important nodes and edges is helpful for rumor or misinformation blocking. Community detection algorithms are important for identifying clustering of similar minded users or groups.

For the case study of the chapter, Twitter data is used from the pool of all online social networks due to the wide range of text data and ease of availability. Twitter data can be extracted through different crawling software and API. For our test case, data are extracted from a Twitter account with the help of Node-XL and i-graph and they are processed though Hadoop based R software. All three software are free and open source softwares.

In a network, due to the hierarchical position some nodes play vital roles. These nodes are referred as centrality nodes of the network. Over the years, several measures of centrality have been proposed. The development of centrality measures helps to clarify the concept of importance of specifying nodes and their interrelationships between them. In the case of centrality, it is expected that nodes with higher centrality will play an important role in information decimation. Several centrality measures are often vaguely related to the intuitive ideas they purport to index. Many of these are so complex that it is difficult or impossible to apprehend what they are measuring. Another school of thought reflects that centrality measures do not have any significance for influence maximization or minimization in the network.

In this book chapter, we are using three popular centrality measures: node degree centrality, betweenness centrality and closeness centrality. Node degree centrality measures number of edges connected with this node. Betweenness centrality signifies number of occurrences of this node within all possible shortest paths of the network. Closeness centrality indicates how this node is connected with all other nodes with the shortest path. All these measures are based on the fact that though online social networks are scale free network but it exhibits the property of small world network. Small world property is based on the fact that all nodes of the network are connected with each other with a small number of separations (popularly known as six degree separation) or connected via other nodes.

Information propagation through the network is dependent on the nature of the propagator nodes. Based on the nature of the propagation of nodes, it can be classified into active, stiffler or ignorant nodes. Seed nodes or the original propagator nodes play important roles for the information propagation. Selections of seed nodes are important for influence maximization in the online social network. As centrality nodes have some hierarchical advantages over the network, when they are used as the seed node, information may be propagated more in the network. In this chapter, through an experiment, it is shown that when seed node is selected through highest betweenness centrality node, information propagation is maximised as compared to other nodes selected as seed node.

Now as information propagation is important for some applications like viral marketing, political campaigning etc., sometime it became important to minimise the influence of some particular information. As any user can post anything in the social network, it is possible that misinformation or rumour can spread through the network. Again it reflects the personal belief or word of mouth. Though different techniques are present for the rumour identification, it is not possible to identify them at the source. So,