IJCSN - International Journal of Computer Science and Network, Volume 6, Issue 6, December 2017
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Research Issues and Developments in Social Network Analytics

[1] Dwarapu Suneetha; [2] Mogalla Shashi

[1] Research Scholar, Department of CSE, JNTU Kakinada, India

[2] Professor, Department of CS &SE, Andhra University, Visakhapatnam, India

Abstract - As billions of individuals share data, opinions, images, videos, through social media, enormous data is getting accumulated attracting researchers towards Social Network Analytics which involves combination of structural and content analytics to mine patterns/knowledge from social media. This paper provides a comprehensive survey of various concepts, challenges, techniques, and outcomes of recent research on Social Media Analytics. The major research issues including partial information, scalability, heterogeneity, structural component and dynamically changing content call for specifically designed techniques for representation and analysis of social media content. Different types of node centralities to quantify the impact of an individual in social media are discussed. This paper provides useful insights on different types of network structures and models for propagation of opinions/influence in different types of networks. It also provides inputs on latest methods for dynamically changing link prediction and visualization for better understanding. The paper discusses successful methods for Sentiment Analysis. It also discusses the concept of Socio dimensions to identify the heterogeneity of connections between nodes of a social network to accurately predict the interests of an individual for targeted marketing etc. Recent research on dynamic topic detection from tweet stream based on a predefined threshold on Minimum Bounding Similarity is discussed. Extensive reference material on related concepts and techniques are briefly discussed in this paper and the citations are helpful for further readings.

Keywords: *Social Media Analytics, Sentiment Analysis, Minimum Bounding Similarity, Content analytics, Patterns, Knowledge*

1. Introduction

Social media is the dominating force in the modern world that can connect people, propagate opinions, influence minds, catch trends, increase sales and help to build business over the years. Social media sites have also proven to be effective platforms for marketing. Social media is the computer mediated tool that allows people, companies, and other organizations to create, share or exchange information, ideas, career interests and pictures/videos in virtual communities and networks. Social networking is the practice of expanding the social contacts by making connections with individuals.

Social networking sites offer people new and different ways to communicate via internet either through their PC, or their mobile phones[1]. They allow people to easily and simply create their own online page or profile and to construct and display an online network of contacts often called as "friends". The profile page functions as the users own web page and include profile information ranging from their date of birth, religion, gender, politics, hometown to their favorite books, quotes, and what they like doing in a spare time. It is important to note that the term "friend" as used on social networking site is different from its traditional meaning. Users of these sites can communicate via their profile both with their friends and with people outside their list of contents, as one to one basis (like email) or in a public way such as comment posted for all users.

With the rise of social media, the web has attracted billions of individuals all around the globe to interact, share, post opinions, and facts. Social media enables us to be connected and interact with each other anytime and anywhere; social media world has no geographical boundaries. This allows us to observe human behavior in a novel scale with a new lens. This social media lens provides us with golden opportunities to understand individuals at large, and to mine human behavior patterns which are otherwise impossible. Social media produces big user generated data and has a huge potential for social science research[4].

Valuable knowledge extracted from this huge amount of social media data motivates user to make informed decisions by drawing useful information from raw data sets, understanding and analyzing useful structures, patterns, and communications in networks, analyzing temporal and special dynamics in networks, analyzing social reputation, influence, and trust. Knowledge extracted from social media also supports user activity modeling, and profiling to build personalized recommender systems[2,3].

Attempts to mine huge data available in social media sites calls for the problem **"drowning in the data but thirsty for knowledge"[7].** Unfortunately social media data is significantly different from traditional data. Apart from enormous size, most of the user generated data is noisy and unstructured with abundant social relationships such as follower-followees, and friendships. It is also essential to combine social theories with computational methods to study how individuals interact, and how communities form.

The uniqueness of social media data calls for specially designed data mining techniques that can effectively integrate user generated content with social structure to discover human behavioral patterns. The study and development of these techniques is termed as Social Media Mining, an emerging discipline under the broad area of Data Mining. Social media mining cultivates a new kind of data scientists who are well versed in social and computational theories, to analyze social media data[6]. The following challenges are faced by the researchers of social media mining.

Partial information: Diverse relations are intertwined with connections in a social network. Connections in a social network can represent various kinds of relationships between users; user might communicate with friends, relatives, classmates, colleagues, or online buddies with similar hobbies. However, often the social media data may not include the relationships between communicating users. The missing relation information can limit the performance of some techniques when they are applied to extract patterns.

Scalability: Networks in social media can be huge, often involving millions of actors and hundreds of millions of connections, while traditional network analysis normally deals with hundreds of subjects or fewer. Twitter, for example, claims to have more than 100 million users, and face book 500 million active users. Existing network analysis techniques might fail to handle networks of this astronomical size.

Heterogeneity: In social media, it is highly likely that heterogeneous types of interaction exist between the same set of users. Moreover, multiple types of entities can also be involved in a network. For example YouTube includes data related to users, tags, videos. Analysis of social media networks with heterogeneous entities and interactions requires new theories and tools.

Dynamic: Social media emphasizes timeliness and the impact of people/theme changing dynamically over time. For example, people quickly loose their interests in most shared content and blog posts. It is common that new users jump in, new connections establish between existing members and old users become dormant or simply leave. Behind noisy interactions, communities can also emerge, grow, shrink, or dissolve. It is challenging to capture the dynamics of individuals and communities and to identify influential persons who act as the backbone of communities and determine the rise and fall of their communities.

This paper provides a comprehensive overview of various concepts and research developments related to social media analytics. Section 2 discusses various approaches to represent social media data, while Section 3 introduces types of network structures, Section 4 provides an overview of different measures for assessing the impact of a node in the network, While Section 5 discusses the propagation of information in social networks. Section 6 & 7 provide a brief summary of important research directions on structural and content based mining of social media data followed by a section that concludes.

2. Data Representation

The connected networks in social networking sites can be represented using graphs. Each individual can be represented using a node, and two individuals who know each other can be connected with an edge. Mathematically a graph is denoted as a pair G=(V,E) where $V=\{v_1,v_2,v_3,\ldots v_n\}$ and $E=\{e_1,e_2,e_3,\ldots e_n\}$. In a graph representing friendship the nodes represent people, and any pair of connected people denotes friendship between them. Edges are also represented by endpoints

$e(v_1,v_2)$ defines an edge e between adjacent nodes v_1 and v_2. While symmetric relationships like 'friend' are represented as undirected edge, asymmetric relationships like 'follower' are represented as directed edge. The number of edges connected to the node is called as **degree** of that node. A hypothetical social theory states that **"the more individuals you know the more influential you are"**[7] .This theory in the graph translates to FaceBook degree represents number of friends a given user has. On twitter, in-degree represents number of followers, and out-degree represents number of followees.

In an undirected graph summation of all node degrees is equal to twice the number of edges. Whereas in a directed graph, the summation of in-degree is equal to sum of out-degrees. On social networking sites, friendships relationships can be represented by a large graph. In this graph nodes represent individuals, edges represent friendship relationships.

Graphs can also be represented using adjacency matrix also known as Socio Matrix. Direct relationship with 1 representing between i^{th} and j^{th} individuals and 0 otherwise. Because of the relatively small number of direct relationships between pair of individuals in the matrix has more number of 0's compared to 1's and hence became sparse. Alternatively social network can be represented using adjacency list. In adjacency list, every node is linked with a list of all its adjacent nodes. Another simple method to store the large graphs is to store all edges in the graph. This is known as edge list representation. Since social media networks are sparse, the representations of adjacency list and edge list save significant space.

Multi graphs are frequently observed in social media, where two individuals can have different interactions with one another. Bi-partite graph is a graph where the node set can be partitioned into two sets such that, for all edges one end point is in one set and the other end point is in the other set. In social media affiliation networks are represented as bi-partite graphs. In these networks nodes in one part represents individuals, and nodes in other part represent affiliations. In other words edges connect nodes in these two sets, but there exist no edges between nodes that belong to the same set.

the individual represented by a node with the **maximum degree**. For any node v_i in an undirected graph the set of nodes it is connected to via an edge, is called its neighborhood which is represented as **$N(v_i)$.** In directed graphs, node v_i has in-degree (edges pointing towards the node v_i) and out-degree (edges points away from the node v_i) are also known as incoming and outgoing neighbors. In social media like

3. Basic Network Structure

Beyond nodes and edges, there are some basic structures that are important to know for describing and understanding social networks.

3.1 Subnetworks

One can consider entire graph as a network or a graph, by looking at how many nodes and edges it has. Often there are also parts of the network representing a subset of the nodes and edges in the graph which is called sub network. The simplest sub networks are **'singletons**' representing the nodes that have no edges. It is very easy to identify singletons in online social networks, often these represent, often these represent, people who signed up but never actively participated. Similarly two individuals interacting with each other but none else form a subnet known as Dyad and three such individuals form a subnet known as known as **Triad.** A **Clique** is a generalization to have any number of nodes of such special type of subnets, in which every node is directly connected to every other node

3.2 Clusters

Clusters are used to identify communities in a social media. But one can describe properties of clusters using some network connectivity measures like density. In the context of social network data analysis the density is one of the measure used to assess the structure of a subnet[5]. It is defined as the ratio as the number of edges in a sub graph to the number of possible edges. Density of a subnet is also called as "Local Clustering Coefficient".

3.3 Ego-Centric Networks

It is a sub network formed by selecting a node and all of its connections to its neighbors. In degree-1

ego centric networks neighbors are defined as nodes one step away from the ego node in the network. Degree-1 ego centric networks are extended by adding the interconnection, if any, existing between the neighbors of the ego to form its 1.5 egocentric network. It is called as 1.5 instead of 2-degree because we are not moving two steps away from the node in the network, we are going only one step but looking at the connection between the neighboring nodes.

4. Social Network Measures

The influence or impact of an individual in social media is quantified using centrality of the node representing the individual. Four types of centralities are defined to assess an individual in four different aspects. The first centrality that can be calculated is **Degree centrality**. In this the highest degree node is considered as the most central node. This is based on the notion that with more number of connections are more important. However high degree centrality of a node need not necessarily reflect how central the node is to the main group especially in networks having two or more clusters of nodes with weak inter cluster links. The second centrality that can be calculated is **Closeness centrality**, for each node it measures, the average distance to the rest of the nodes as shortest path. A lower value of closeness centrality indicates a more vital node since it is close to many nodes. The third centrality is **Betweeness centrality**. This measures how necessary a node is in a network for data to flow along shortest paths from one node to the other. The higher the value the more central the node is. The final centrality measure is **Eigen Vector Centrality;** this is calculated by taking into consideration not only the nodes importance, but also the importance of its neighbors. A node with important neighbors will be more central to a network than a node with less number of important neighbors. This measure is used in Google's Page Rank algorithm.

5. Propagation in networks

Social networks have changed the ways of spreading information or opinion in the modern world in the last few decades. It allows any kind of information to pass from one person to other person in a rapid way compared to olden days. Epidemic model for diseases and their spread states that, diseases can pass from infected individual (those who are carrying a disease) to susceptible individual [8](those who do not have the disease, can catch it) by many factors like age, area, immunity of a person etc. But it is very difficult to analyze and identify the factor that causes a susceptible individual to become infected. Compartmental models are designed to simplify the analysis. It divides persons into three categories, as S (Suspected), I (Infected),and R (Recovered). Susceptible individuals are those who do not have the disease, but possibly catch it. Infected individuals are those who are carrying the disease and Recovered individuals are those who are recovered from the disease. Similar models are designed to handle the information spread in networks like spread of diseases, rumors, frauds etc. So by combining the letters from the above three categories different models are designed to understand the spread of diseases like SI, SIR, SIRS, and SIS. The pattern of these disease causing contacts form a network. In SI disease predict a model, a person is susceptible to the disease, then became infected and never recovers. Where as in SIR model, a person who is not carrying the disease may become infected and recovers. In SIRS model, after recovery, a person is susceptible to become sick again. In all the above models, susceptible individual became infected, but some people will have more robust immune system. A person with more immunity is less likely to catch the disease, so stochastic model calculate the probability for passing a disease from an infected person to susceptible individual. Threshold model introduces a minimum threshold on the number of infected neighbors of the susceptible individual to become infected. If a person can be infected from one neighbor it is called as 1-threshold model. Similarly a susceptible person may be infected from k neighbors in a K-Threshold model.

6. Research topics in Social Networks based on Structural Analysis

6.1 Link Prediction In Social Networks

In social media Links are continuously changing over time, and hence dynamic in nature. Sometimes network errors may give rise to missing links even though there exists a relation between a pair of nodes. Link prediction aims to predict a link which is going to be formed in future. It also detects a missing link based on the network structure in its neighborhood. Various techniques for predicting link between a pair of nodes include:

Jaccard Index: This estimates the similarity between pair of users as the ratio of total number of

common friends to the total number of people who are friends of either the nodes.

Common neighbor: This estimates the similarity between a pair of users as the number of neighbors that two nodes share in common.

Adamic/adar: This is a variant of common neighbor method. Instead of simply counting on the common neighbors the pair of nodes has, this method gives higher weightage to the nodes with less number of friends than celebrities with large number of friends.

Preferential attachment: It calculates the degree of each node and predicts that nodes with a high degree will have a chance to get additional connections.

6.2 Visualization in Social Networks

Since social networks are large in size, it is difficult for us to identify clusters, cliques, to extract patterns, etc, from those large networks. So visualization provides an easier way to understand the network and summarize the information by providing different layout algorithms such as graph layout, circular layout, force-directed layout etc[2]. Using this layout algorithms user can easily understand the networks better by visualization.

7. Research on Content Based Mining Issues in Social Networks

7.1 Sentiment Analysis

Sentiment Analysis also known as 'Opinion Extraction' extracts individual opinions on different types of products, movies, political issues etc. With the increased participants in social networking sites like Twitter, Face Book etc, users are posting large amount of information, public opinions etc. . If an individual wants to buy a product there is no need to ask opinions of family and friends, because there are many public opinions, discussions about the product on the web. Instead of conducting surveys which are expensive business executives can capture public opinion from these social networking sites applying appropriate sentiment analysis techniques. Because of these applications sentiment analysis is extensively used in various fields like elections, events etc. Many research papers[9] have been published in the area of "Sentiment Analysis and Opinion Mining" .

7.2 Node Classification in Social Media

Social networking sites allow user to publish large quantities of data that is crucial for advertising domains. Some of the nodes but not all in the network are labeled, indicating the class or interest of the individual represented by the node. In addition, when two nodes are connected the labels of the two nodes may be correlated. In this scenario, it is beneficial to combine the known structural information and the labeled data to propagate their labels through semi-supervised learning. Such an ability to predict the labels/interests of individual's aids targeted marketing/advertising in order to propose recommend products/services of possible interests to the users.

Most of the researchers on network structure analyze considers connections between nodes as homogeneous, ignoring their heterogeneity. The connections between pairs of nodes represent the relationship between them and, in turn represent type of influence a node has on the other node. Lei Tang and Huan Liu[9] advocated that it is essential to make use of heterogeneity of connections for achieving accurate classification. They proposed a new frame work called 'Socio Dim' to find interests of a node based on the network structure differentiating the heterogeneous connections to extract social dimensions. There are different methods to obtain social dimensions of a node such as latent space models [10,11] modularity maximization[12], Spectral Clustering [13] etc. Socio Dim uses Spectral Clustering to obtain social dimensions of a node. They build a SVM classifier to determine the class labels of unknown individual.

7.3 Identifying Uniqueness of an individual based on a criteria

Yi-Hen Lo etal.,[14] has identified the significance of the problem of finding the uniqueness of a query node in a social network based on a given criteria. For example the set of persons who are equivalent to the query individual (v) based on a set (M) of skills/characteristics possessed by him is termed as Uniqueness Identification group of v (UID(V)) based on M. UID for a vertex v implies that v is unique based on M. They used ego-centric bi-partite graph to organize the data from social networking sites. They proposed three ways uniquely named as 1-HOP+, one neighbor method, multiple neighbor method. Some applications need

to identify a group of experts, a compact possess sharing a set of skills rather than just one expert defined to find MUID of a query vertex. The above mentioned three methods of UID are extended.

7.4 Summarization and Dynamic Topic Detection

From tweet streams Zhenhua Wang etal.,[15] takes up the task of dynamically summarizing tweet stream on a time line. Huge amounts of noise, redundancy and the bulk challenges the researchers as they deal with continuously arriving tweets with varied themes. Tweets are clustered based on the common keywords used in them representing a shared theme. Centroid of each cluster represents a theme and the cosine similarity of a tweet to its closest centroid represents its compatibility with one of the existing themes if the similarity is greater than minimum bounding similarity (MBS) otherwise this tweet becomes the seed for a new cluster representing a new theme different from the existing themes. The set of themes relevant to a particular time period defines the tweet summary or snapshot at that time period. Different snapshots of the evolving themes are maintained in a pyramidal time frame (PTF) structure such that the importance of various snapshots fades away with time. Incremental clustering is adapted to update the clustering solutions originally obtained using K-means algorithm for tweet clustering. Thus the proposed framework 'SUMBLR' generates both online summaries and historical summaries of arbitrary time durations.

8. Conclusion

According to Pew Research Center on Internet Science & Technology news bulletin dated 8th Oct '15, 65% adults are using social networking sites which is a tenfold increase in the last decade. Since lot of data is getting accumulated in social networks, they provide a rich source for data mining researchers to extract hidden patterns and knowledge useful to various domains. This paper proposes an overview of various concepts and measures related to social network analysis. It also through light on the recent research outcomes and applications of social network analytics.

References

[1] Mazin Abed Mohammed, Belal AL-Khateeb, Dheyaa Ahmed Ibrahim,"**Human Interaction with Mobile Devices on Social Networks by Young and Elderly People: Iraq a Case Study",*Indian Journal of Science and Technology,*** 2016 Nov, 9(42), Doi no: 10.17485/ijst/2016/v9i42/101281

[2] B. Akshaya, S. K. Akshaya, S. Gayathri, P. Saravanan,"**Investigation of Bi-Max Algorithm for On-Line Purchase Recommender System using Social Networks",*Indian Journal of Science and Technology,*** 2016 Nov, 9(44), Doi no:10.17485/ijst/2016/v9i44/98932

[3] R. Satish Srinivas, C. S. Anish Balaji, P. Saravanan,**"Online Product Recommendation using Relationships and Demographic Data on Social Networks",*Indian Journal of Science and Technology,*** 2016 Nov, 9(44), Doi no:10.17485/ijst/2016/v9i44/99896

[4] Nathaneal Ramesh, J. Andrews, "**Personalized Search Engine using Social Networking Activity",*Indian Journal of Science and Technology,*** 2015 Feb, 8(4), Doi no:10.17485/ijst/2015/v8i4/60376

[5] Kim Jong-Weon, Park Ki-Nam,"**A Study on Methodologies to Develop an e-Industrial ClusterHub System using Social Networks",*Indian Journal of Science and Technology,*** 2015 Sep,8(21), Dono:10.17485/ijst/2015/v8i21/78377

[6] Sayed Zakarya Taghavinezhad, Fariba Nazari, Zahed Bigdeli,"**Review of Factors Effecting Social Networks Acceptance among Graduate Students at Islamic Azad University of Ahvaz",*Indian Journal of Science and Technology,*** 2015 Sep, 8(21), Doi no: 10.17485/ijst/2015/v8i21/79091

[7] Sayed Zakarya Taghavinezhad, Fariba Nazari, Zahed Bigdeli,"**Review of Factors Effecting Social Networks Acceptance among Graduate Students at Islamic Azad University of Ahvaz",*Indian Journal of Science and Technology,*** 2015 Sep, 8(21), Doi no: 10.17485/ijst/2015/v8i21/79091

[8] Jennifer G. Analyzing the social web. Elseveir.

[9] Lei T, Huan L. Leveraging social media networks for classification. Springer ***DMKD 2010***.

[10] Hoff PD, Raftery A, Handcock M. Latent space approaches to social network analysis. 2002;1090-1098

[11] Sarkar P, Moore A. Dynamic social network analysis using latent space models. ***SIGKDD;*** 7(2):31-40.

[12] Newman M. Modularity and community structure in networks. ***PNAS*** 2006; 103(23):8577-8582.

[13] Luxburg U. A tutorial on spectral clustering. Stat comput; 17(4):395-416.

[14] Yi-chen L, Jhao L, Mi-yen Y, Shou L, Jian P. What distinguishes one from its peers in social networks?. ***DMKD 2013;*** 27() 396-420.

[15] Znenhua W, Lidan S, Chen K, Gang C, Sharad M. On summarization and time line generation for evolutionary tweet streams. ***IEEE Transactions on knowledge and data engineering***. 2015,May, 27(5), pp.1301-1315.