ISSN: 2321-7782 (Online)
Volume 4, Issue 3, March 2016
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

# A Survey on Data Mining Techniques for Analysis of Social Network

**R. Adaikkalam**[1]
Research Scholar
P.G. and Research Department of Computer Science
Khadir Mohideen College
Adirampattinam
Thanjavur (Dist), Tamil Nadu, India

**Dr. A. Shaik Abdul Khadir**[2]
Associate Professor
P.G. and Research Department of Computer Science
Khadir Mohideen College
Adirampattinam
Thanjavur (Dist), Tamil Nadu, India

***Abstract: Data mining is the extraction of projecting information from large data sets, is a great innovative technology which helps corporations focus on the most important information in their data stockrooms. Data mining makes use of various statistical, machine learning and graphical methods and separate the knowledge in to a form which is very much useful for many real world applications. Social network analysis has become a very popular field of research as it is useful for many applications. In this research paper we have overviewed various data mining techniques used for social network analysis.***

***Keywords: Data Mining, Social Networks, Online Social Networks, Web Mining, Social Network Analysis.***

## I. INTRODUCTION

Social networks are defined as virtual spaces where people of all ages can make contacts, share information and ideas, and build a sense of community. A social network community represents people and connects them. It has also provided a way of keeping in touch with friends, constructing personal profiles, view others profiles and connections, communicate and share personal information. Members of social network communities manage their identity through their profile, they meet new friends and like-minded people in the community, they connect with each other, rate peers and objects, ask questions, get answers and discuss topics. Facebook, twitter, MySpace, and bebo, can be taken as most commonly accessed social network sites. Social networks can be used in many business activities like increasing word-of-mouth market ing, marketing research, General marketing, Idea generation & new product development, Co-innovation, Customer service, Public relations, Employee communications and in Reputation management.

Social networks analysis, the analysis targets are mainly focused on resources from the web, such as the contents of the web, the structures of the web and the usage behaviors of users in the web. Among the information techniques that can be used for the analysis of social networks, Data mining is claimed to be the most suitable one. Therefore, it is more than suitable to use the data mining techniques for social networks analysis, and it is also the focus of this paper.

## II. REVIEW OF LITERATURE

In this section, related literatures about social networks analysis and Data Mining will be reviewed, in order to present a broad view about these two topics for readers.

### *2.1 Social Networks Analysis*

In the research area of social networks analysis, it is usually the main task about how to extract social networks from different communication resources. The data that used for building social networks is relational data, which can be obtained and transferred from different resources including the web, email communication, internet relay chats, telephone communications, organization and business events, etc. For example, the email communication is a rich source for extracting and constructing social networks. In the issue of email social networks extraction, the relationship between email senders and receivers can be transformed by measuring the frequency of email communication with take the communication behavior (such as reply, forward, etc.) into account. The transformed relational data can then be used for social networks construction.

In the past three decades, social network analysis has developed a range of concepts and methods for detecting structural patterns, identifying patterns of different types of relationship interrelate, analyzing the implications that structural patterns for the behavior of network members, studying the impact on social structures of network members and their social relationships.

### *2.1.1 Types of Social Network Analysis*

A social network has a set of relations of ties, which can be viewed in two different ways. One approach focus on an individual, called ego-centered network, and put it at the centers of the network. Members of the network are defined by the relations with the ego. Ego-centered network analysis can show the range and breadth of connectivity for individuals and identify those who have access to diverse pools of information and resources. The ego-centered approach is useful when the population is large, or the boundaries of the population are hard to define. The second approach considers the whole network based on some specific criterion of population boundaries such as a formal organization, department, club or kinship group. Whole network analysis can identify those members of the network who emerge as central figures or who act as bridges between different groups. This approach requires responses from all members on their relations with all others in the same environment, such as the extent of email and video communication in a workgroup.

### *2.1.2 Key Concepts of Social Network Analysis*

Network analysis provides a rich and systematic means of assessing such network by mapping and analyzing relationships among people, teams, departments or even the entire organization. A network is composed of three elements—(1) actors (2) relations between actors, and (3) the linkages among actors. Actors and their actions are viewed as interdependent rather than independent, autonomous units. Actors can be persons, organizations, or groups, or any other set of related entities. Relations between actors are depicted as links between the corresponding nodes. A tie connects a pair of actors by one or more relations. Pairs may maintain a tie based on one relation only or a multiplex tie based on many relations. Thus, ties also have characteristics like content, direction and strength, but they are often referred to as weak or strong. Social network analysts have found that multiplex ties are more intimate, voluntary, supportive and durable. In addition, the linkages among actors have several characteristics, which are direction, degree, and content. The direction of linkages covers symmetrical and asymmetrical relations; the degree of linkages means the strength of relations, and the content of linkages includes friendship, information, power, and influence, etc. Owing to complex properties of nodes, relations, and linkages, scholars utilizing the concept of network in their studies have different definitions of network.

### *2.1.3 Social Network Analysis Techniques*

Visualization is also a hot topic of social network analysis, and it is a suitable technique in this area. Through the visualization of social networks, the characters of social networks can be understood easily, such as the structure of networks, the distribution of nodes, the links (relationships) between nodes and the clusters and groups in the social networks.

In additional to social network extraction and visualization, there are other measurements that can be used for social network analysis as well. For example, centrality degree of a social network is a measurement that is used to measure the betweenness and closeness of the social network. Betweenness centrality indicates the extent to which a node lies on the shortest path between every other pair of nodes. Closeness centrality analyzes centrality structure of a network based on geodesic distances among nodes in a social network. Cluster coefficient is a measurement to discover the clusters in a social network and to measure the coefficient of the clusters. The density measurement can be used to analyze the connectivity and the degree of nodes and links in a social network.

The measurements path length and reachability can be used to analyze how to reach a node from another node in the social networks. Structural hole is also a measurement of social network analysis, which can be used to discover the holes in a social network and by this to fill the hole and expand the social network. These sparse regions are structural holes that prove opportunities for brokering information flows among actors. Thus, maximizing the structural holes spanned or minimizing redundancy between actors is an important aspect of constructing an efficient, information-rich network.

### *2.2 Data Mining*

### *2.2.1 Overview of data mining*

It is the process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Data warehouses are being made use of in order to store large amounts of data. The growth of commercial databases has had a huge impact on the necessity of data mining in organizations. Data mining allows organizations to proactively respond to problems that may arise in future by forecasting about specific occurrences. As illustrated in figure 1, the first step is data preparation. Data is selected, processed under the knowledge of a domain expert. Second, a data mining algorithm is used to process the prepared data. The third phase is to analyze whether important facts were generated by the data mining algorithms.

Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be use for purposes such as measuring and improving program performance.

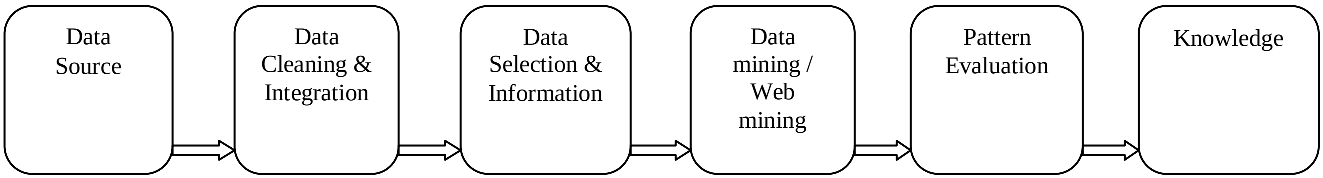

Figure 1: Data Mining Steps

### *2.2.2 Data Mining Algorithms*

Data mining techniques is dividing in to two approaches; direct approach is used in prediction where it tries to predict a state of a new value by looking at the known values. The Second approach, non -direct approach is used to identify new patterns by looking at the past values. Before start creating mining models data should be cleaned and prepared. The Mining models can be created on following Algorithms.

***Association Rules*** - This algorithm can be used in marketing base analysis like identifying cross -selling opportunities. This takes multiple items in a single transaction, scans the data and counts the number of times the items appear in the transaction so it can be used to identify the relationships in the large data sets.

***Clustering*** **-** This algorithm groups the data according to their similar characteristics. This can be used to identify the relationship of the characteristics among a group. When a new data is introduced, the characteristics of it can be mapped with the relationships, it can be used to predict the behavior of the new data. Clustering can be used to find anomalies of the data as well. This is commonly used in systems of fraud detection and Customer relationship Management.

***Decision Trees*** **-** This is the simple and one of the most commonly used algorithm. This is used to predict discrete and continues variables.

***Liner regression*** **-** This is predicting only continues variables using single multiple liner regression formula.

***Logistic Regression*** **–** This algorithm uses a neural network without hidden layers.

***Naïve Bayes* –**This can be used to calculate probabilities for each possible state of the input attribute when a state of a predictive attribute is given. This can be used as the starting algorithm of the predicting process.

***Neural Networks* –** This algorithm has been adopted from artificial intelligence. This can be used to search nonlinear functional dependencies. This will perform non liner transformations on the data in layers from input layers to the hidden layers and finally to the output layer.

***Sequence Clustering* -** This looks for cluster based models than the similarity if the data. The model use sequence of events by using hidden Markov chains. The states are models to a matrix and the probabilities of transiting from one state to another in the cells of the matrix. With these probabilities the probabilities for sequence of transition can be calculated by multiplying probabilities of state transitions in the sequence. The chains of highest probability can be used to model the clusters.

***Time Series*** - This is used to forecast continues variables. This is a combination of two algorithms called auto regression trees and Auto regressive integrated moving average.

## III. Social Network Data Mining Algorithm

### *3.1. Overview of the Algorithms*

When mining social network data it should be a combination of web structure mining and web content mining. Analyzing the structure of the Social network is known as Social Network Analysis. Social Network analysis where was a hot topic among the researchers from 1994 especially in the fields of psychology, anthropology, economics, geography, biology and epidemiology, anthropology, economics, geography, biology and epidemiology. Several tools has been introduced in the area of social network analysis like Graph Characterization Toolkit, TweetHood, Meerkat, NetDriller, HiTS/ISAC Social Network Analysis Tool ,D-Dupe and X-RIME, a cloud-based library for large scale social network analysis. Web content mining was more popular in marketing and advertising research.

### *3.2. Analysis of the Algorithms*

In social network data mining, existing data mining algorithms cannot be used directly because of the dynamic behavior. When analyzing the literature on social network data mining techniques; it was found that each algorithm has strengths and weaknesses. The following section explains about the existing algorithms in detail.

### *3.2.1. Graph mining algorithms*

Most popular data mining technique in Social Network Analysis is using Graph mining algorithms. World Wide Web including social networks is a collection of interconnected hypertext documents. These are interconnected by hyperlinks. So we can be considered as a directed graph, where nodes will be the hypertext documents and edges will be hyperlinks. Web structure analysis based on graph algorithms has been analyzed in many researches in past years. Lahiri and Berger-Wolf (2008) have created and tested methods combining network, quantitative, semantic, data processing, conversion and visualization-based components. They have introduce a new graph mining algorithm "periodic subgraph mining, or the discovery of all interaction patterns that occur at regular time intervals" taking into consideration of the dynamic behavior of Social Networks. The Algorithm is based on frequent pattern mining in transactional and graph databases with periodic pattern mining in unidimensional and multidimensional sequences.

Bourquiet.al (2009) presented a framework which is based on dynamic graph discretization and graph clustering. This framework is capable of detecting the dynamic changers of the social network structure and identifies events analyzing temporal dimension and exposes command hierarchies in social networks. The particular algorithms treat the network as a graph but it minimize the clustering problems and graph partitioning problems. As a solution minimum spanning trees can be used to identify users having similar profile pages and strong relationships. Zhang et.al (2010) have conducted an experiment on the applicability of general greedy, hill-climbing and centrality-based algorithms on dynamic social network data to identify key users for target marketing by mapping the network to a graph. They have proposed a new approximation searching algorithm based on the heuristics information from the above algorithms.

Even though the graphs map the connection or the relationship between the nodes it does not show the relationship strength. One interesting tools has been developed called SocialViz to provide h frequency information on social relationship among multiple entities in the networks by using a Frequent Pattern Visualization Approach.

### *3.2.2. Classification*

Classification is the method of categorizing data in to one of many categories. This can be apply in web data mining to classify user profiles based on profile characteristics. Most popular classification algorithms in data mining are decision trees, naïve Bayesian classifier and neural networks. Surma and Furmanek (2010) introduced an interesting algorithm called C&RT, combining classification and regression tree algorithms to determine rules to identify target groups to market. This can be used in real social network data.

### *3.2.3. Clustering*

Clustering is grouping a set of items such a way that items in the same group are more similar to each other than to those in other groups. These groups are known as a cluster. Clustering is mainly used in information retrieval in web mining. Based on past research clustering will increase the efficiency in information retrieval. Graph based clustering is comely used in web structure mining as explained in early section. Text based clustering is most commonly used in web content mining whether you create clusters based on the content of the web document. Bartal et.al (2009) introduced an interesting method combing social network analysis and text based clustering to predict the nodes of a social network would be linked next.

### *3.2.4. Associations*

Association rule mining is used to find frequent patterns and correlation among data set. Nancy et.al (2013) had use association rules to mine social network data using 100 Facebook university pages. The research focused on the formulation of association rules using which decisions can be made and uses Apriori Algorithm to derive association rules.

### *3.2.5. Semantic Web and Ontology*

Semantic Web is a new research area where it tends to give meaning to Web data. This enables machine and humans to interact intelligently and exchange information. There are many researchers has been carried out in this filed like using semantic geo catalogues and recovery in mental health information. Zhou et.al (2008) explains applying statistical learning methods on semantic web data. It has used an extended FOAF (friend-of-a-friend) ontology applied as a mediation schema to integrate Social Networks and a hybrid entity reconciliation method to resolve entities of different data sources]. Tushar et.al (2008) explains the usage of Semantic Web technology to detect the associations between multiple domains in a Social Network. Opuszko and Ruhland (2012) introduced a novel approach of using semantic similarity measure based on pre-defined ontologies for classify social network data. Ostrowski (2012) has developed an algorithm to retrieve information in social networks to identify trends. The Algorithm has use semantics for determine the relevancy of networks using unstructured data. The algorithm was tested on twitter messages.

### *3.2.6. Markov models*

Markov chains are is a mathematical algorithm that undergoes transitions from one state to another, among a finite or countable number of possible states. It is a random process where the next state depends only on the current state and not on the sequence of events that preceded it. Markov models can be used in web mining to predict users' next action.

The Social network can be mapped to a where nodes will be users previous visits. So based on the node information by using Markov models users next visit can be predicted. When analyzing the literature it proves that most of the research has been carried out in web structure mining and less in web content mining. Most of the researches are tested on static networks; they do not consider the dynamic behavior.

## IV. Conclusion

Data mining is an interesting field which can be used to produce new knowledge by analyzing large collection of data. In order to apply the traditional data mining techniques the data should be stored in data warehouses in a structured manner. Social Network Analysis is the study of social structure. The social network analysts are interested in how the individual is embedded within a structure and how the structure emerges from the micro-relations between individual parts. As an approach to social research, SNA displays four features: structural intuition, systematic relational data, graphic images and mathematical or computational models.

This research paper has specifically focused on the techniques used to mine social network data. Most of the algorithms are developed to mine the structure of the social network where mapping the network to a graph.

## References

1. E.K Clemons, " The Future of Advertising and the Value of Social Network Websites: Some PreliminaryExaminations".Minneapolis, Minneosta, USA, 2007.
2. D. Boyd., "Social Network Sites: Definition, History, and Scholarship", Journal of Computer-Mediated Communication 13 (1), 2007.
3. "Social Network Marketing: The Basics" Available: httpe_Basics.pdf [Aug 01, 2013].
4. J. Rennie, G. Zorpette, "The Social Era of the Web Starts Now," IEEE Spectrum, June 2011.Available:http://spectrum.ieee.org/teleco m/internet/the-social-era-of-the-web-starts-now [Aug 01, 2013].
5. J.W. Seifert. "Data Mining: An Overview", CRS Report for Congress, 2004 Available:http://www.fas.org/irp/crs/RL31798.pdf [Sep 05 2013].
6. E.Veerman, T.Lachev, D.Sarka, Microsoft SQL Server 2008- Bussiness Intelligence Development and Maintenance, Microsoft, PHI Learning Private Limited, pp 372 – 380, 2009.
7. C. Yu, X. Ying, "Application of Data Mining Technology in E-Commerce". In IEEE nt.Conf. on International Forum on Computer Science-Technology and Applications, pp.291-293, 2009.
8. G. Lappas G. "From Web Mining to Social Multimedia Mining.", IEEE International Conference on Advances in Social Networks Analysis and Mining, pp 336 – 343, 2011.
9. J. Srivastava, Cooley R., Deshpande M., and Tan P.N., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKKD Explorations, 2000.