Leveraging Graph Theoretic Approach For Online Social Network Analysis

[1]Javaid Iqbal Bhat, [2]Kaiser J. Giri
[1,2]Department of Computer Sciences,
[1]University of Kashmir, Srinagar, J&K, India.
[2]Islamic University of Science & Technology, Kashmir, J&K, India.
[1]javaidonnet@gmail.com, [2]kaiserjaveed@gmail.com

Abstract:

A social network is a structure made up of set of individuals connected by the set of one or more specific type of social relations. Knowing the exquisiteness of availability of increasing real-world social network data, social networks are receiving perpetually lot of attention from scientific communities. The purpose of this paper is to explore various facets of Social Network Analysis (SNA) using data mining techniques associated to World Wide Web (www), normally referred as Web Mining. In this paper, the significant properties of social networks will be deliberated along with the review of various techniques of web mining. The prime objective of this study is to provide a roadmap for researchers who are engaged in using data mining techniques for realization of different utility trends in social network data along with the emphasis to explore the possibility of representation of social networks by an alternate, other than the Adjacency Matrix.

Keywords: Social Network, Social Network Analysis, Web Mining, Clustering Coefficient, Web Structure.

I. Introduction:

In the current state of technology, we have been witnessing a complete shift from the traditional online technology, normally referred as Web 1.0 to Web 2.0 that describes the changing trends in the use of World Wide Web and is aimed to significantly improve collaboration among internet users, content providers and enterprises. In this decade we have also been able to realize the beauty of participative instead of individual publishing, two way instead of single way communication and end user generated instead of independently developed content. The impact of this is completely visible in the ways software developers and end users utilize the web in the active consumption of content. As a result of all this, the development of social networks and its communities with effective information sharing and improved web functionalities has acquired huge popularity and signifies one of the most important social and computer science phenomena of these years, largely based on 'architecture of participation'. This has happened because of many factors including the popularity of Online Social Networks (OSNs), availability of large volumes of OSN log data, effective representation & analysis of Social Network Graphs (SNGs) and predominantly the market interests of social networks. Social networking sites have skyrocketed in popularity in a very short span of time as shown in Fig 1. [1] Facebook, Twitter, LinkedIn, Wikipedia, YouTube have been able to make it to the top 15 global websites.

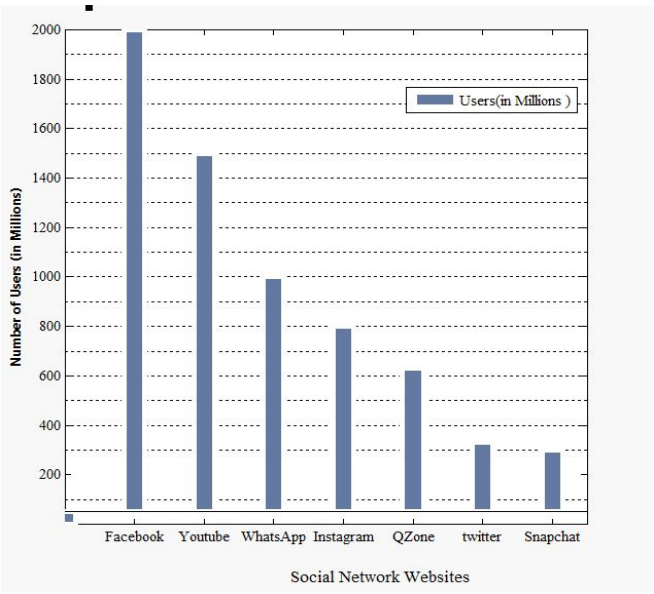

Figure 1. Global rank of some major social networking sites

II. Representation:

To provide the basis to represent the pattern of ties with their network actors, graphs and matrices [2] are the two standard mathematical structures used by the Social Network Analysts (SNAs).

A. Graph Based Representation:

Graphic models are used in resolving the problems of almost every conceivable discipline [3]. In network analysis a special kind of graphic display in the form of point and a line is used in combination. These points normally called as nodes are used to represent the network actors and the lines which usually are known as edges, are a reflection of relations or ties. To represent it explicitly, we have G= (V, E), where V consists of a non-empty set of nodes (or vertices) and E is a set of edges. Every edge is supposed to connect its points and necessarily has at least one or utmost two nodes, called as end points, associated to it. A graph is characterized by some values like degree of nodes, average path length of the nodes, clustering coefficients, triangle and isomorphism numbers and many more. These values are derived to generate statistical details for various purposes. On the similar pattern we can use theses graph models only to represent the people in terms of relationship, where a simple graph can designate whether two people originally are known to each other or not. People in the particular group are represented by the vertices and the undirected edge that represents the relationship between the two end points of it reflects that these two persons know each other very well. However, in the networks no multiple edges or loops are used except when we are interested to include the self-knowledge, for that we do use the loops. This type of undirected graph is used by the social network sites including the Facebook and LinkedIn. Moreover the directed edges are also a kind of acquaintanceship but is found in the social networks like Twitter, where a person follows the other person, based on the concept of ‘following’. In the present state, where the acquaintanceship of people if represented in a graph will emerge as the graph having more than eight billion vertices and almost one trillion edges, a reasonable number of social scientists have agreed that by and large any pair of the people in this world is linked to a chain of four or utmost six people. This intern represents that all the people if represented by the vertices in an acquaintanceship graph shall be linked by a path of at the most four. The play ‘Six Degrees of Separation’ is also based on the very same concept.

B. Matrix Based Representation:

Among all the standard mathematical structures, matrix representation is considered one of the finest way of representing the social network data, particularly with regard to analysis and extensive understanding of the data. One of the significant reasons behind this structure is that it is highly convenient rather pragmatic for most of the mathematical operations and more importantly has an inherent property to be always treated as a graph like structure. In addition to this, the original graph can be easily rebuild without losing any data item from such structure.

The information within a graph G= (V, E) can be stored in the linked, indexed or in any other common convenient matrix form. However, one of the highly suitable storage matrix that represents the real relationship of vertices and edges in a graph *G= (V, E)* is known as *Adjacency Matrix*. This adjacency matrix normally denoted by *A*, contain all the necessary entries of a graph to sufficiently designate whether any pair of vertices is adjacent to each other or not. This type of matrix is also known as *Sociomatrix* in the social network analysis domain. This square matrix *A* of size *m* can very effectively describe an un-weighted, undirected graph *G= (V, E)*. The rows and columns of such matrix *A* designate the indices of each vertex of the graph and is usually labelled as *1, 2…. N*. The data values of this Sociomatrix a_{ij} is always equal to 1, iff the specified pair of vertices identified by their indices are adjacent to each other and 0 otherwise .This simply means that $(i,j)^{th}$ location will hold a value 1 in case a_i and a_j are adjacent to each other. In case the graph is undirected the matrix is symmetric with respect to its diagonal values, then $a_{ij} = a_{ji}$, $V\ i \neq j$. Sociomatrices are used extensively for undirected network structures in terms of storage and effective analysis, predominantly because of various properties. These social networks arise to the sparse Sociomatrices and hence are fully conversant to adapt the existing techniques of compact matrix decomposition for effectively storing the data [4].

The other standard representation of an undirected graph *G= (V, E)* using the matrices is called *Incidence Matrix*, usually denoted by *I*. The value of this matrix reflect the sufficient information about the edges of the graph incident on the various vertices, i.e. it gives the details of edges which are incident on some vertices, by indexing the edges on columns and the vertices on rows . The result of this is a matrix dimension *I* as *VXE*. A matrix value I_{ij} is always equal to 1 if the value V_i is directly incident on the edge E_j, and is 0 otherwise. Both the matrices, Adjacency and Incidence are sufficient enough to fully describe the graph representation. Since Incidence Matrix in contrary to Adjacency Matrix, does not demand to be necessarily a square matrix provides enough flexibility. Moreover the Incidence Matrix also need not to be symmetric as is the case in Adjacency Matrix of simple undirected graphs. In addition to all this when a simple graph is sparse i.e. the number of edges is comparatively very less in number, then the Adjacency matrix can be replaced by *Adjacency List*. This brings lot of improvement in terms of computational complexity, i.e. in a graph of *N* vertices where each vertex has a very less values of in and out degrees compared to the value of *b*, where *b* is a constant much smaller than the *N*. Therefore there are likely not higher than *bN* items in the Adjacency List while as if same represented in the Adjacency Matrix, there is always N^2 value of the matrix which is higher than *bN* and is computationally always poor and wasteful in terms of storage. In the light of all this the considerable efforts need to be worked out while analyzing the suitable data structures along with the algorithm consideration particularly keeping in view the purpose of introducing Adjacency List and Incidence Matrices.

III. Social Network Properties:

There are some significant number of social network properties that include the following four [5] as deliberated below:

a. Diameter:

In a social network, reasonable pair of people are connected by very short chains of friends. As a result of it, it is found that most of the real world graphs genuinely exhibit a very small diameter relatively. A graph has diameter *'d'* if each pair of nodes are connected by a pair of length of utmost *'d'* edges. However, the diameter of graph is very ambiguously used by the different authors at different occasions. At least four of its definitions found in the Social Network Analysis literature are as follows:

1) The longest shortest-path length, which is the true graph theoretic diameter but which is infinite in disconnected networks.

2) The longest shortest-path length between connected nodes, which is always finite but cannot distinguish the complete graph from a graph with a solitary edge.
3) The average shortest-path length, and
4) The average shortest-path length between connected nodes

b. Navigability:

Based on the concept of small world model by Walton, Dodds and Newman [6], social networks exhibit the same phenomena where the navigability is possible. The model given by them is purely based on the multiple hierarchies including geography, occupation, hobbies etc., into which the people fall and an attempt is made by an algorithm based on greedy strategy to move closer to the target at every step in any dimension. However no theoretical results have been established despite the fact that the simulations have remarkably shown the working of algorithm as well as the model to allow navigation. Not only there exists the shortest paths connecting the most pairs of the people but instead the people in the network can generate and then construct the shortest paths from source to destination using the local information and the limited knowledge of global structures.

c. Clustering Coefficient:

The measure of the probability that the two individuals who have common friend will be friends themselves, is informally considered as the clustering coefficient of the network. This coefficient for a node $u \in V$ in graph is the fraction of edges that is present between the two nodes adjacent to u and in other ways within the neighborhood of *u*. Moreover the average clustering coefficient taken over all the existing nodes in a graph is considered as clustering coefficient of the entire network.

d. Power and Centrality:

Power is considered as a fundamental characteristic of the social structure, as also agreed by almost all the social scientists and psychologists. However, there is also a serious disagreement about the definition of the power and its description. More importantly how can we describe and analyze the causes and consequences of power. Some of the prime approaches that the social network analysts have developed to study the power and concept of centrality is summarized below:

i) Degree:

Degree is the number of ties for an actor, where a tie connects two or more nodes in a graph. Ties can be direct or indirect. Many human behaviors such as advice seeking, information sharing are direct ties while co-memberships are examples of undirected ties [7]

ii) Closeness:

The degree an individual is near to all other individuals in a network (directly or indirectly). It reflects the ability to access information through the "grapevine" of network members. Thus closeness is the inverse of the sum of the shortest distances between each individual and every other person in the network.

iii) Betweenness:

The extent to which a node lies between other nodes in a network. This measure takes into account the connectivity of the node's neighbors, giving a higher value for nodes which bridge clusters.

iv) Density:

It is the measure of the closeness of a network. Given a number of nodes, the more links between them, the larger the density. If the number of nodes in a network is *n*, and the number of links is l, then its density is given by:

$p = \frac{2l}{n*(n-1)}$; For directed graphs

$p = \frac{l}{n*(n-1)}$; For undirected graphs

IV Web Mining:

Information and communication technology (ICT) on one side has facilitated large amount of storage of data as a dynamic information source in the form of structurally complex and even ever growing social network and also on the other side provides a fertile ground for the application of data mining principles or web mining. This web mining discipline encompasses a wide range of concerns and issues purely aimed at deriving actionable knowledge from the web sources and thereby involves researchers from information retrieval, database technologies and artificial intelligence [8]. Oren Etzioni [9] and many others of course have introduced the term *'Web Mining'* formally but term web mining is subjective and the different authors slightly mean differently from web mining. For example, Jaideep Srivastava and colleagues [10] define it as: "The application of data-mining techniques to extract knowledge from web data, in which at least one of the structure or usage data (web log) is used in the mining process (with or without other types of web data)". Web mining consists of three main categories based on the web data used as input in web data mining. Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM).

a) Web Content Mining:

Web content mining is the application of data mining techniques to extract content published on the Internet, usually as HTML (semi structured), plaintext (unstructured) or XML (structured) documents. In other words it may be defined as the procedure of retrieving the information from the web into more structured forms and indexing the information to retrieve it quickly [11]. Table I summarizes the concept of web content mining:

Property	***Details***
View of Data	Unstructured, Semi-Structured & Structured
Main Data	Text Documents &Hypertext Documents
Representation	Edge Labelled Graph& Relational
Method	Proprietary Algorithms &Association Rules
Application Categories	Categorization, Clustering & Website Schema Discovery

Table I. Web Content Mining

b) Web Structure Mining:

Web structure mining operates on the web's hyperlink structure. This graph structure can provide information about a page ranking [12] or authoritativeness [13] and enhance search results through filtering. In other words web structure mining may be defined as the process by which we discover the model of link structure of the web pages. Its aim is to generate structured abstract about the website and web page. Table II summarizes web structure mining:

Property	***Details***
View of Data	Link Structure
Main Data	Link Structure
Representation	Graph & Web Page Hits

Method	Proprietary Algorithms & Web PageRank
ApplicationCategories	Categorization & Clustering

Table II. Web Structure Mining

c) Web Usage Mining:

Web usage mining analyzes results of user interactions with a web server, including web logs, click streams and database transactions at a web site or a group of related sites. It is used to identify the browsing patterns by analyzing the navigational behavior of user. Web usage mining tries to make sense of the data generated by the web surfer's sessions or behaviors, whereas web-content mining and web-structure miningutilize real or primary data on the web. Web usage mining introduces privacy concerns and is currently the topic of extensive debate.

Property	***Details***
View of Data	User Interactivity
Main Data	Server Logs (log-files) & Browser Logs
Representation	Relational Table, Graph & User Behaviour
Method	Machine Learning and Statistical & Association Rules
Application Categories	Site Construction Adaptation & Management, Marketing and User Modelling

Table III. Web Usage Mining

V Web Mining Techniques:

a) Association Rules:

To uncover relationships between seemingly unrelated data in relational database or any other transactional and simple information repository, we make use of if / then statements. These if / then statements are normally called as Association Rules. Association rule mining was introduced in 1993 by R Agrawal and et all [14] to represent a data mining technique with ultimate objective to extract fascinating correlation, recurrent pattern and unpremeditated structures of various sets of items. In the transactional database or any other relevant repository, the target of this technique is always an interesting relationship among a large rather vast data items. A survey of the association rules may be presented in [15]. Moreover the sophisticated and intelligent techniques are required in the present scenario as the traditional simple marginal and conditional profanities are not sufficient to describe the causal relationships. Association rule mining is easy to use and implement. The patterns discovered with this data mining technique can be represented in the form of association rules. Rule support and confidence are two measures of rule interestingness. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts. In social network analysis, association rule mining can help discover the hidden relationships between the different nodes of a network.

b) Classification:

Classification is the process to automatically build a model that can classify a class of objects so as to predict the classification or missing attribute value of future objects (whose class may be unknown) [16]. In the first stage of this two stage process, based on the collection of training data set, a model is constructed to describe the characteristics of a set of concrete concepts or data classes. Since both concepts and the data classes are predefined, the initial stage of this process is also known as supervised learning. In other words the training sample belongs to which particular class is agreed in advance. In the second stage, the model is exclusively used to predict the classes of future objects or data. There are reasonable number of techniques for classification. Classification based on decision tree has been well researched out and lot of algorithms have been developed. A complete survey of classification using decision trees is given in [17]. Bayesian classification is another technique mostly used and can be found in [18]. Nearest neighbor methods are also discussed in many statistical texts on classification, such as [19]. Many other machine learning and neural network techniques are used for constructing the classification models at the moment.

c) Clustering:

An unsupervised learning process very similar to the classification one, for grouping of a set of physical or abstract objects into classes of similar objects, ensuring that the objects within the same cluster turn out to be similar to greater extent and simultaneously dissimilar to the objects of other different clusters [16]. Contrary to the classification process, where in record-class association is almost predefined, the clustering process does not involve any such class, rather objects are purely grouped together based on similarities. Similarity functions are actually used to define the similarity or dissimilarity between the objects and are quantitatively specified as distance or other measures by corresponding domain experts. A survey of clustering techniques and algorithms can be found in [20]. In social network analysis, discovering the closest people in the network is usually the prime functionality and is generally achieved by using a visualization technique in a small social network. Thus clustering may emerge a potential technique for identifying more clusters and groups in large social networks. Besides, it can also offer more meticulous information than visualization [21] including the closeness of a group, detailed information of members in a group and the relationship between groups in a social network.

VI Conclusion and Future Scope:

In conclusion, the social network research displays four prime features like Structural Intuition, Systematic Relational Data, Graphic Images and Computational Models [22]. After presenting the comprehensive view of all these four significant features, this paper deliberates upon the formal representation methods of social networks along with the significant properties of such networks. The representation of social networks using adjacency and incidence matrices was also extensively conversed keeping in view its mathematical beauty and background along with its progressive possible applications in the modern day computational science in near future. Besides all this, the computational complexity of association rule mining needs a serious contemplation by the present day researchers in order to reduce its limitations. Finally this paper shall provide all the strong basis for the researchers of social network analysis and enrich the domain of applications and studies of web mining with a focus on impending challenges.

VII References:

[1] Alexa, “Alexa the Web Information Company”, 2014. Online Accessed on Dec. 25, 2014]. Available: www.alexa.com/comparison/ 2014.
[2] A. Hanneman and M. Riddle, “Introduction to Social Network Methods”,[Online].Available: http://www.faculty.ucr.edu/~hanneman/nettext/ 2005.
[3] Kenneth H. Rosen, “Discrete Mathematics & its Applications with Combinatorics and Graph Theory”, TMH, 2012.

[4] V. Snasel, Z. Horak, J. Kocibova, A. Abraham, "Reducing social network dimensions using matrix factorization methods", International Conference on Advances in Social Network Analysis and Mining, pp. 348 – 351, IEEE , 2009.
[5] D. Liben-Nowell, "An Algorithmic Approach to Social Networks", Ph. D Thesis, Massachusetts Institute of Technology, June 2005.
[6] D. J. Watts, P. Sheridan Dodds, and M. E. J. Newman, "Identity and Search in Social Networks", Science, 296:1302 - 1305, 17 May 2002.
[7] G. Plickert, R. Cote, B. Wellman, "It's Not Who You Know. It's How You Know Them: Who Exchanges What with Whom?", Social Networks, Vol. 29, No. 3, pp.405-429, 2007
[8] P. Kolari, A. Joshi, "Web Mining: Research and Practice", IEE CS, July- August, 2004
[9] O. Etzioni, "The World Wide Web: Quagmire or Gold Mine?" Comm. ACM, vol. 39, no.11, pp. 65–68, 1996.
[10] J. Srivastava, P. Desikan, and V. Kumar, "Web Mining: Accomplishments and Future Directions," Proc. US National Science Foundation Workshop on Next-Generation Data Mining (NGDM), National Science Foundation, 2002.
[11] Z. S. Zubi, "Ranking Web Pages Using Web Structure Mining Concepts", Recent Advances in Telecommunications, Signals and Systems, 2010.
[12] L. Page et al., "The PageRank Citation Ranking: Bring Order to the Web", Tech. report, Stanford Digital Library Technologies, Jan. 1998.
[13]. J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", Proc. 9th Ann. ACM–SIAM Symp. Discrete Algorithms, ACM Press, pp. 668–677. 1998.
[14] R.Agrawal, T. Imielinski, and A.N. Swami, " Mining Association Rules between Sets of Items in large Databases", In Proceedings of the ACM SIGMOD International Conference on Management of Data,. Washington, D.C., 207–216. 1993.
[15] Q. Zhao, S. Bhowmick, "Association Rule Mining: A Survey", No. 2003116, Technical Report, CAIS, Nanyang Technological University, Singapore, 2003.
[16] J. Han, M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann, 2000.
[17] S. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary Survey", Data Mining and Knowledge Discovery 2, 4, 345–389, 1998.
[18] R. Duda, T. Hart, "Pattern Classification and Scene Analysis.", Wiley & Sons, Inc., 1973.
[19] M. James, "Classification Algorithms", Wiley & Sons, Inc., 1985.
[20] P. Berkhin, "Survey of clustering data mining techniques" Technical Report, Accrue Software, San Jose, CA, 2002.
[21] B. Tatemura, Y.Wu, "Tomographic Clustering to Visualize Blog Communities as Mountain Views", In Proc. Of WWW Conference, Japan, 10- 14 May, 2005.
[22] F. Borko, "Handbook of Social Network Technologies and Applications", Springer Publications, 2010.