Volume 2, Issue 10, October 2012 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Analysis of Social Networks Using the Techniques of Web Mining

E.Raju
Department of Computer Science & Engg
SR Engineering College
WARANGAL, A.P

K.Sravanthi
Department of Computer Science & Engg
University College of Engineering
Kakatiya University, WARANGAL.

***Abstract*—Web 2.0 technologies have brought new ways of connecting people in social networks for collaboration in various on-line communities. Social network analysis deals with the interactions between individuals by considering them as nodes of a network (graph) whereas their relations are mapped as network edges. Study of such structures lies on the intersection of different areas of research: sociology, graph theory and data mining. This paper studies issues around analysis of social networks using web mining techniques. Techniques and concepts of web mining and social networks analysis will be introduced and reviewed along with a discussion about how to use web mining techniques for social networks analysis. This paper also sets out a process for social network analysis using web mining. Comparison of social networks with other networks is also studied. Discussions of the challenges and future research are also included.**

***Keywords*— Web Mining, Social Networking, Social Network Analysis, Association Rule**

I. INTRODUCTION

Social networks analysis is an interesting research direction to analyze the structures and relationships of social networks, such as analyses of density, centrality and cliques in social network structures [24]. In recent years, on-line social networking has become a very popular application in the age of Web 2.0 [12], which allows users to communicate, interact and share on the World Wide Web [1].

Social networks, such as Facebook and Bebo, are essentially online communities that allow users to come together, communicate and share things such as photographs, music or other files; and, most prolifically, to create short messages, often in the style of a mobile phone text message but shared among a group. People use the sites to ask their friends questions, say how they feel today and what they are up to, to comment on something they have seen on someone's page. These sites are typical of the internet today in exploiting what are known as 'Web 2.0 technologies'.

Recently, the websites that are created based on the concept of web 2.0 have becoming the main stream in WWW especially for those social networking websites, such as Blog, friends making website, web album, etc. A social network is the network of relationships and interactions among social entities such as individuals, groups of individuals, and organizations[28]. Since the rise of Internet and the World Wide Web has enabled us to investigate large-scale social networks, there has been growing interest in social network analysis[16].

A social network is usually formed and constructed by daily and continuous communication between people and therefore includes different relationships, such as the positions, betweenness and closeness among individuals or groups [15]. In order to understand the social structure, social relationships and social behaviors, social network analysis therefore is an essential and important technique. Research on social networks could be traced back to sociology, anthropology and epidemiology. Social network analysis developed with the kinship studies of Elizabeth Bott in England in the 1950s and the 1950s-1960s urbanization studies of the University of Manchester group of anthropologists. Initially, the studies of social networks analysis focused on small groups and small social networks. However, it became harder and harder to manually analyse very broad social networks. Therefore, strong computation power and information technology has became a very important tool for social networks analysis and the direction of the research is therefore now moving from sociology to computer science. For social networks analysis, the analysis targets are mainly focused on resources from the web, such as its content, structures and the user behaviors. Application of data mining techniques to the World Wide Web, referred to as *Web mining*, can be used for the analysis of social networks [5]. In web mining, main analysis targets are from the World Wide Web, in the form of web content mining, web structure mining and web usage mining [4].

The structure of this paper is organized as below: In section 1, the background and introduction will be introduced. Some related literatures of social network analysis, the taxonomy and techniques of web mining will be reviewed in section 2. Section 3 introduces social network models. In section 4, a study of how web mining techniques can be used for social networks analysis will be included. In section 5, a general process for applying these techniques will be proposed. In section 6, a comparison of various kinds of social networks with other networks is reviewed. In section 7, there is a discussion of the challenge of using web mining for social networks analysis and some suggestions for future research directions.

II. LITERATURE REVIEW

In this section, related literature about social network analysis, the taxonomy and techniques of web mining will be reviewed, in order to present a broad view of these two topics for readers.

A. Social Networks Analysis

A social network is a social structure made up of individuals (or organizations) called nodes which are tied (connected) by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, dislike, relationships of beliefs, knowledge or prestige [29]. Most broadly, social network analysis conceptualizes social structure as a network with ties connecting members and channelling resources, focuses on the characteristics of ties rather than on the characteristics of the individual members and views communities as 'personal communities', that is, as networks of individual relations that people foster, maintain, and use in the course of their daily lives[30].

Social Network Analysis (SNA) is used to analyse the interpersonal relationships within an organization or community and can provide rich and systematic descriptions and interpretation of complex social relationships. SNA focuses on the interconnections of the actors, instead of on the peculiarities of the actors themselves.

Common concepts in social network analysis:

(i) Ties: Ties or links connect two and more nodes in a graph. Many human behaviors, such as advice seeking, information-sharing, and lending money to somebody are directed ties while co-memberships are examples of undirected ties [21].

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

(iii) Centrality: The measures of centrality identify the most prominent actors, especially the star or the "key" players, that is, those who are extensively involved in relationships with other network members. The most important centrality measures are: Degree centrality, Between-ness centrality and Closeness centrality.

(iv) Closeness :The degree an individual is near 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.

(v) Clique: A clique in a graph is a sub-graph in which any node is directly connected to any other node of the sub-graph.

(vi) Clustering coefficient : A measure of the likelihood that two associates of a node are associates themselves. A higher clustering coefficient indicates a greater 'cliquishness'.

(vii) Cohesion : The degree to which actors are connected directly to each other by cohesive bonds. Groups are identified as 'cliques' if every individual is directly tied to every other individual, 'social circles' if there is less stringency of direct contact, which is imprecise, or as structurally cohesive blocks if precision is wanted.

(viii) Density: Density is a measure of the closeness of a network. Given a number of nodes, the more links between them, the larger the density. Its formal definition is as follows. If the number of nodes in a network is n, and the number of links *l*, then the density is,

$$\rho = \frac{2l}{n*(n-1)} \qquad \text{for directed graph} \quad (1)$$

and

$$\rho = \frac{l}{n*(n-1)} \qquad \text{for undirected graph} \quad (2)$$

(ix) Path length: Nodes or actors may be directly connected by a line, or they may be indirectly connected through a sequence of lines. A sequence of lines in a graph is a "walk", and a walk in which each point and each line are distinct is called a path. The length of a path is measured by the number of lines which makes it up.

(x) Reach : The degree any member of a network can reach other members of the network.

In social networks analysis the main task is usually about how to extract social networks from different communication resources [15] [18]. The data used for building social networks is relational data , which can be obtained from different resources including the web, email communication, Internet relay chats, telephone communications, etc [3]. For example, email communications are a rich source for extracting and constructing social networks. By means of email social networks extraction, the relationship between email senders and receivers can be transformed by measuring the frequency of email communication and taking the communication behaviour (such as reply, forward, etc.) into account [6].

In additional to social network extraction, there are other measurements that can be used for social network analysis as well. For example, degree centrality in a social network is used to measure the betweenness and closeness of the social network [26]. The density measurement can be used to analyse the connectivity and the degree of nodes and links in a social network [17].

Path length and reachability can be measured to analyse how to reach a given node from another node in the social networks.

B. Web Mining

Web mining is an application of data mining, the technique of discovering and extracting useful information from large data sets or databases [13]. Web mining therefore can be defined as to discover or extract useful information from the web [4].

Different Types of Web Mining

The web mining techniques can be divided into three different types, which are *Web Content Mining, Web Structure Mining and Web Usage Mining [4],* shown in Fig. 1.

Web Content Mining is a web mining technique to analyze the contents in the web, such as texts, graphs, graphics, etc [2]. Recently, most of web content mining researches are focused on the text data processing and few are focused on other multimedia data. Natural language processing (NLP) is therefore the main technology that used in this area. The concept and techniques of Semantic Web and Ontology are also relevant here [19].

Web structure mining is a technique that can be used to analyze the links and structure of websites. Graph theory is usually the main concept and theory for web structure mining to analyze and explain the structure of websites. In addition, the extraction of the structure of websites is always essential in this research area [9][10]. The main concern in this area is how to design and implement a crawler (or spider, bots) to extract and construct the structure of websites, such as in the research area of Deep-web.

Web usage mining is a web mining technique that can be used to analyze how the websites have been used, such as the navigation behavior of users [7]. The server-side clickstream data (logs file) is the main sources that used for web usage mining. Client-side data (such as client-side logs file, cookies) is sometimes to be used due to some research concerns, such as in order to record more complete behavior of users [27][4]. Different web usage mining analyses include basic statistical analysis of the navigation behavior of users in a website: such as how many times the website has been browsed, where the users come from, etc.

Fig.1. Taxonomy of Web Mining

C. Web Mining Techniques

Traditional data mining techniques can also be used for web mining, such as classification, clustering, association rule mining, and visualization. In web mining, classification algorithms can be used to classify users into different classes according to their browsing behavior, for example according to their browsing time. After classification, a useful classification rule like "30% of users browse product/food during the hours 8:00-10:00 PM" can be discovered. The

criterion by which items are assigned to different clusters is the degree of similarity among them. The main purpose of Clustering is to maximize both the similarity of the items in a cluster and the difference between clusters [14].

The association rule technique can be used to indicate pages that are most often referenced together and to discover the direct or indirect relationships between web pages in users' browsing behavior [23]. For example, an association rule in the web usage mining area could take the form "the people who view web page index.htm and also view product.htm the support=50% and the confidence=60%".

III. SOCIAL NETWORK MODELS

A. Using formal methods to show Social Networks

One reason for using mathematical and graphical techniques in social network analysis is to represent the descriptions of networks compactly and systematically. A related reason for using (particularly mathematical) formal methods for representing social networks is that mathematical representations allow us to apply computers to the analysis of network data.

B. Using Graphs to Represent Social Relations

Network analysis uses (primarily) one kind of graphic display that consists of points (or nodes) to represent actors and lines (or edges) to represent ties or relations. When sociologists borrowed this way of graphing things from the mathematicians, they renamed their graphs as "sociograms".

Visualization by displaying a sociogram as well as a summary of graph theoretical concepts provides a first description of social network data. For a small graph this may suffice, but usually the data and/or research questions are too complex for this relatively simple approach.

C. Using Matrices to Represent Social Relations

The most common form of matrix in social network analysis is a very simple one composed of as many rows and columns as there are actors in our data set, and where the elements represent the ties between the actors. The simplest and most common matrix is binary. That is, if a tie is present, a one is entered in a cell; if there is no tie, a zero is entered. This kind of a matrix is called an "adjacency matrix" (Figure2).

	Bob	Carol	Ted	Alice
Bob	---	1	0	0
Carol	1	---	1	0
Ted	1	1	---	1
Alice	0	0	1	---

Fig.2 Using Matrices to Represent Social relations

IV. WEB MINING TECHNIQUES FOR SOCIAL NETWORK ANALYSIS

This section will show how the three different types of web mining along with the other web mining techniques introduced in section 2 can be used for social networks analysis.

A. *The Three Web Mining Types for Social Networks Analysis*

Web content mining, text mining or natural language processing are very useful in social network analysis. For example, web content mining can categorize or classify documents on an on-line social networking website, especially articles on blogs or text forums. Article categorization is usually the first task for many social networks analyses or applications.

Web content mining can also be used in social networks analysis to analyze users' reading interests, and determine their favorite content. However, for most on-line social networks analysis tasks, it is usually necessary to utilize all three types of web mining and techniques together.

Web usage mining also plays an important role in social networks analysis. It is useful for the analysis of social networks extraction discussed in section 2 of this paper. The usage data and user communications on an on-line social networking website can be transformed into relational data for social-networks construction [22]. In addition, web usage mining is also a tool for measuring centrality degree. For example, the closeness of blog users can be measured by:

$$Closeness = (f*(w*b)) + (f*(w*r)) + (f*(w*i))$$

In the equation above, *f* denotes the frequency of a blog behavior, and *w* is the weight of closeness for each blog behavior. The three blog behaviors are *b=browsing*, *r=reading* and *i=interaction*. This is just a simple example of web usage mining, but the techniques allow many possible means of social networks analysis.

Web structure mining is the third kind of web mining and it is also useful for extracting and constructing social networks to extract the links from WWW, e-mail or other sources. It can also be used to analyze path length, reachability or to find structural holes, which are very basic and traditional social networks analyses. Web structure mining usually uses graphs and visualized means to represent the data about social networks, enabling the analyst to easily understand and analyze social networks [11].

B. *Web Mining Techniques for Social Networks Analysis*

There are many other different kinds of web mining techniques. In this section, examples will be given of two web mining techniques used for social networks analysis. The two techniques are clustering and association rule mining.

Clustering:

In social networks analysis, finding the group of closest people in a network is usually the main task, and is normally achieved by using a visualization technique in a small social network. Thus, the clustering technique can be used with a large social network in identifying more groups and clusters. Moreover, it can provide more detailed information than just using visualization can [25], including the closeness of a group, detailed information of the members in a group and the relationship between groups in a social network.

Association rule mining:

In social network analysis, association rule mining can help discover the hidden relationships between nodes in a social network or even cross networks. For example, an association rule for social networks analysis may be "person A knows person B and also knows person C, the support is 0.9 and the confidence is 0.5" or "the person who read person A's blog article and also read person B's blog article, the support is 0.9 and the confidence is 0.5". In addition, association rule mining is helpful for the application after social networks analysis, such as recommendation systems or information filtering systems [20].

V. THE PROCESS OF WEB MINING FOR SOCIAL NETWORK ANALYSIS

A general process to use web mining for social networks analysis is shown in figure 3. Its steps include selection of analysis targets, selection of social networks analysis, data preparation, web mining techniques selection, results presentation and interpretation, recommendation and action.

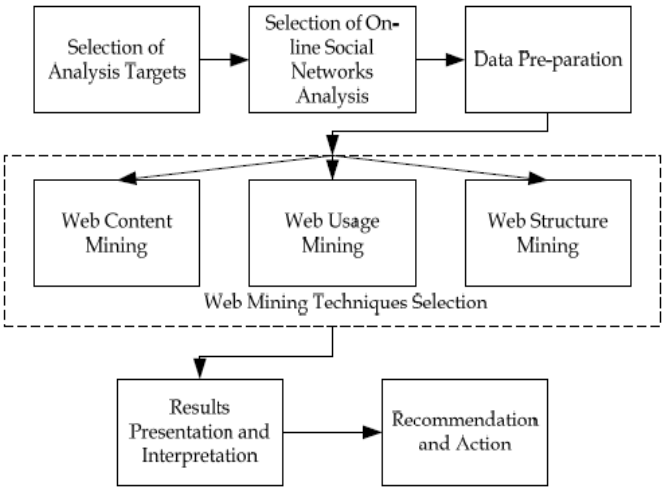

Fig.3 The general process of social network analysis

The first step is the selection of the analysis targets, such as web, email, telephone communications, etc. Sometimes, more than one target will be selected. After this step, we then select what of kind social networks analysis we will proceed with.

Once the analysis targets and the social networks analysis methodology have been selected, the next step is data preparation. In this stage related data will be collected for analysis, then cleaned and formed as the final format to store in database.

The next step is selecting the web mining techniques to be used and then proceeding with them. More than one technique may be selected and sometimes a combination of techniques is necessary. The selected suitable techniques are then used to analyze the data collected and prepared in the third step of the process.

The results of the analysis after web mining are then presented and interpreted either manually or automatically. Visualization techniques are sometimes used to assist the presentation of the results of the analysis, such as the extracted social networks.

The last step of the general process to use web mining for social networks analysis is recommendation and action. This is an optional step, and the process may be terminated after the analysis results have been generated. The recommendation and action step deals with the results of the analysis; for example, if structure holes in a social network have been discovered recommendations about how to fill the hole may be manually or automatically generated and then appropriate action can be taken.

This general process of using web mining for social networks can be a continuous process. In some research projects, it start again after recommendations have been generated and action taken. The process starts again to perform the action performance evaluation or to do a new analysis.

VI. COMPARISON OF VARIOUS PARAMETERS OF SOCIAL NETWORKS

A. Comparison of Social Network with Other Networks

TABLE I
COMPARISON BETWEEN SOCIAL NETWORK AND OTHER NETWORKS

Type of Network	Conceptual Definition	Operatrional Measure	Content of relation/Link
Social Network	A set of people (or organizations or other social entities) connected by a set of relationships	Individual, Group, Organization, Nation-State	Any Kind of Social Relation
Communication Network	A network composed of interconnected individuals linked by patterned flows of information	Same as above, but generally focuses on individual people	Communication and Information
Computer Mediated Network	A specific type of communication network in which individuals are interconnect	Same as above, but also includes computer systems	Same as above, but restricted to computer as channel of information flow
Internet Network	A communication network connected by the Internet among computer systems	Same as above, but focuses on Internet users	Same as above, but restricted to Internet as channel of information flow
Hyperlink Network	An extension of traditional communication networks in that it focuses on the structure of a social system based on the shared hyperlinks among websites	Same as above, but focuses on websites which represent Individuals, Groups, Organizations, Nation-States	Same as above, but Restricted to hyperlink as channel of information flow

B. *Comparison of Different Kinds of Social Networks*

TABLE II
COMPARING BASIC PROPERTIES OF DIFFERENT KINDS OF SOCIAL NETWORKS

Social Network Type	Actors	Ties	Direction
Friendship Network	People in Society	Friendship relations between people	Undirected
Web's Social Network	Web Pages	Links Between Web Pages	Directed
Semantic Web Social Networks	Semantic Web Documents or Concepts in Them	Semantic Relations Between Documents or Concepts, such as *foaf* : *knows*	Directed or Undirected

VII. CONCLUSION AND FUTURE RESEARCH

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 [30].

This paper studies the application of the concept and techniques of web mining for social networks analysis, and reviews the related literature about web mining and social networks analysis. Moreover, how to use web mining and a general process of using web mining for social networks analysis have also been studied.

Social networks analysis carried out using the techniques of web mining is an interesting area of research. However, there are several challenges in this research area to be overcome. For example, data sampling is a big issue when using web mining for social networks analysis. In social networks analysis, it becomes a difficult task to select suitable samples representative of the real social networks. Other challenges include finding communities in social networks, finding patterns in social networks and analysing overlapping communities.

We will shift our future research focus to overcome the challenges discussed above, such as the how to reduce the data size and not affect the characters of the social networks. In addition, we will focus on how to apply the web mining techniques to some real on-line social networking websites, such as blogs and on-line photo albums. The enormous amount of scope is available to develop new algorithms to study social interactions, particularly to detect community structures, key nodes and other useful information. Less researched data mining techniques in the field of social networks are stochastic based techniques that capture the uncertainty and dynamic aspects of social interactions.

REFERENCES

[1] Adamic, L. A., and Adar, E. "Friends and Neighbors on the Web" Social Networks, Vol. 25, 2007, pp. 211-230.
[2] Joshi, A., and Joshi, K. "On Mining Web Access Logs", in *Technical Report*, CSEE Department, UMBC, 1999.
[3] Cai, D., Shao, Z., He, X., Yan, X. and Han, J. "Mining Hidden Community in Heterogeneous Social Networks" In *Proceedings of LinkKDD 2005 Conference*, August 21, 2005, Chicago, IL, USA, pp. 58-65.
[4] Cooley, R., Mobasher, B. and Srivastave, J. "Web Mining: Information and Pattern Discovery on the World Wide Web*" In Proceedings of the 9th IEEE International Conference on Tool with Artificial Intelligence*, 1997, pp. 558-567, Newport Beach, CA, USA.
[5] Chakrabarti, S. "Mining the Web: Discovering Knowledge from Hypertext Data" Morgan Kaufmann Publishers, USA, 2003.
[6] Chin, A. and Chignell, M. "Finding Evidence of Community from Blogging Co-Citations: A Social Network Analytic Approach" In *Proceedings of the IADIS International Conference on Web Based Communities* 2006, San Sebastian, Spain, February 26-28, 2006.

[7] Buchner, A. G. and Mulvenna, M., D. "Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining", *ACM SIGMOD Record*, 27(4):54-61,1998.
[8] Garton, L. and Haythornthwaite, C. and Wellman, B., "Studying Online Social Networks", *Journal Of Computer - Medicated Communication*, June 1997.
[9] Spiliopoulou , M. "Web Usage Mining for Web Site Evaluation", Communication of the ACM 43, 8, 2000, pp. 127- 134.
[10] S.Chakrabarti, "Data mining for hypertext: A tutorial survey", ACM SIGKDD Explorations, 1(2):1 - 11, 2000.
[11] S., M. "Advances in Exponential Random Graph (*p**) Models Applied to A Large Social Network" Social Network, Vol. 29, 2007, pp.231-248.
[12] Goth, G. "Are Social Networking Sites Growing Up?" IEEE Distributed Systems Online, Vol. 9, No. 2, February 2008.
[13] Hand, D., Mannila, H., and Smyth, P. "Principles of Data Mining", MIT Press, Cambridge, MA, 2001.
[14] Jain, A. K., Murty, M. N., Flynn, P. J. "Data Clustering: A Review" ACM Computing Surveys, Vol. 31, No. 3, pp. 264-323, 1999.
[15] Jin, Y. Z., Matsuo, Y., and Ishizuka, M. "Extracting Social Networks among Various Entities On the Web" In *Proceedings of the Fourth European Semantic Web Conference, 2007*.
[16] McCallum, A., Corrada- Emmanuel, A., and Wang, X. 2005. "Topic and role discovery in social networks." In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 786–791.
[17] Lento, T., Welser, H. T., Gu, L., and Smith M. "The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Web logging System" In *Proceedings of the 15th International World Wide Web Conference*, May 23-26, 2006, Edinburgh, Scotland.
[18] Matsuo, Y., Tomobe, H., and Nishimura, T. "Robust Estimation of Google Counts for Social Network Extraction" In *Proceedings of Twenty Second Conference on Artificial Intelligence (AAAI-07)*, July 22-26, 2007, Vancouver BC Canada.
[19] Mika, P. "Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks" Web Semantics, Vol. 3, Issue 2-3, October 2005, pp. 211-223.
[20] Mishne, G. "Using Blog Properties to Improve Retrieval" In *Proceedings of ICWSM* 2007, Boulder, Colorado, USA.
[21] G. Plickert, R. Cote, and 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, 2007, pp. 405–429.
[22] Nowson, S., and Oberlander, J. "Identifying More Bloggers" In *Proceedings of ICWSM* 2007, Boulder, Colorado, USA.
[23] Pierrakos, D., Paliouras, G., Papatheodorou, C., Spyropoulos C. D. "Web Usage Mining As A Tool for Personalization: A Survey", User Modelling and User Adapted Interaction, Vol.13, Issue 4, pp. 311-372, 2003.
[24] Ting, I. H. "Web Mining Techniques for Online Social Networks Analysis" In *Proceedings of International Conference on Service Systems and Service Management*, Melbourne, June 30-July 2, 2008, pp 1-5.
[25] , B. L., Tatemura, J., and Wu, Y. "Tomographic Clustering to Visualize Blog Communities as Mountain Views" In *Proceedings of World Wide Web 2005 Conference*, Chiba, Japan, May 10-14, 2005.
[26] Wang, Y., and Li, X. " Social Network Analysis of Interaction in Online Learning Communities" In *Proceedings of 7th IEEE Internationall conference on Advanced Learning Technologies (ICALT 2007)*, Niigata, Japan, July 18-20, 2007.
[27] Hui-Ju, W., Ting, I.H., and Wang, K,Y. "Combining Social Network Analysis and Web Mining Techniques to Discover Interest Groups in the Blogspace", In *Proceedings of 4th IEEE International Conference on Innovative Computing, Informatin and Control (ICICIC)* , Kaohsiung, Taiwan, Dececmber 7-9, 2009, pp 1180-1183.
[28] Kimura, M., Saito, K., and Nakano, R. "Extracting Influential Nodes for Information Diffusion on a Social Network", *ACM DMKD* Explorations, vol. 20, issue 1, January, 2010, pp. 70-97.
[29] Wen-jun, S., and Hang-ming, Q. "A Social Network Analysis on Blogspheres" *In Proceedings of the 15th IEEE International Conference on Management Science and Engineering*, 2008, pp.1769 - 1773, Long Beach, CA, USA.
[30] Borko, F. "Handbook of Social Network Technologies and Applications (1st ed)", Springer publication, 2010.

