User Identification Framework in Social Network Services Environment

Brijesh BAKARIYA, G.S. THAKUR
Department of Computer Application, Maulana Azad National Institute of Technology
Bhopal, India
brijesh_scs@yahoo.co.in, ganshyamthakur@gmail.com

Social Network Service is a one of the service where people may communicate with one another; and may also exchange messages even of any type of audio or video communication. Social Network Service as name suggests a type of network. Such type of web application plays a dominant role in internet technology. In such type of online community, people may share their common interest. Facebook LinkedIn, orkut and many more are the Social Network Service and it is good medium of making link with people having unique or common interest and goals. But the problem of privacy protection is a big issue in today's world. As social networking sites allows anonymous users to share information of other stuffs. Due to which cybercrime is also increasing to a rapid extent. In this article we preprocessed the web log data of Social Network Services and assemble that data on the basis of image file format like jpg, jpeg, gif, png, bmp etc. and also propose a framework for victim's identification.
Keywords: *Preprocessing, Web Log Data, Social Network Service, Data Mining, Cyber Crime*

1 Introduction

Now days we can see social networking services are growing day by day, First of all the term of social network is a social structure made of individuals (or organizations) called "nodes," which are tied (connected) by one or more specific types of interdependency, such as friendship, kinship, financial exchange, dislike or relationships of beliefs, knowledge or prestige [8].Social networking is the grouping of individuals into specific groups, like small rural communities or a neighbourhood subdivision, if you will [3]. Generally some social services focus on the privacy issue too. Due to which one person could not see the friend link of other one. There are various types of social network are available. Personal networks-These kinds of network allow creating profiles online and linked with other person, with a focus on social linkage such as friendship. For example, Orkut, LinkedIn, Facebook, MySpace are stage for communicating with contacts [9][4][13]. These kinds' networks involve users exchanging information with people. Statuses update networks-These types of social networks are developed to allow small status update to person in order to communicate with other persons. For example, Twitter, these types of network people may exchange messages or thoughts day by day. Location networks-These networks are developed to show one's real-time location, either as public information or as an update viewable to authorized contacts. Content sharing networks-These networks are designed as platforms for sharing content, such as music, photographs and videos. For example YouTube and Flicker. Shared interest networks-Some social networks are built around a common interest or geared to a specific group of people. These networks incorporate features from other types of social networks but are slanted toward a subset of individuals, such as those with similar hobbies, educational backgrounds, political affiliations, ethnic backgrounds, religious views, sexual orientations or other defining interests. For example LinkedIn [16].

DOI: 10.12948/issn14531305/18.2.2014.02

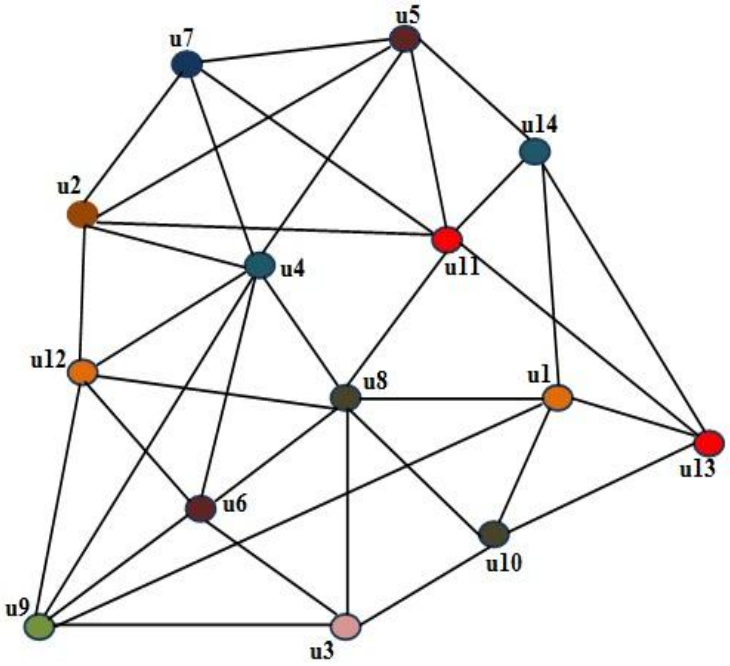

Fig. 1. Architecture of Social Network

Figure 1 shows that the social relationships among different users. In social networking services any person can interacts another person we have shown the person using dot in the give Fig. and maintained its relationships through lines. In this network any number of users (u1, u2, u3 ...) can communicate each other.But their exists some services that do not suffer from such kind of issues like LinkedIn Facebook, orkut etc. 'Mutual friend' is that feature existing on such services which provides a way of connecting through other person.

Mutual friends are the persons who are Social Networking Service (SNS) friends with both you and the person whose timeline you are viewing. Or we can say mutual friends are the common friends of a person. Suppose u1 is friend of u10 and u14 then u1 is mutual friend of u10, u14. According to Fig. 1 we are showing in the table1 of mutual friends of an individual user.

Table 1. Mutual Friends of a user

MUTUAL FRIEND	USERS
u_1	u_8,u_{13},u_{14},u_{10}
u_2	u_{12},u_7
u_3	u_9,u_{10}
u_4	$u_8,u_{12},u_5,u_2,u_7,u_9,u_6$
u_5	u_7,u_{14}
u_6	u_4,u_8,u_{12},u_9,u_3
u_7	u_2,u_6
u_8	$u_1,u_4,u_{11},u_{12},u_6,u_3,u_{10},u_{11}$
u_9	u_{12},u_3
u_{10}	u_3,u_{13}
u_{11}	$u_{13},u_{13},u_5,u_7,u_2,u_8$
u_{12}	u_2,u_9
u_{13}	u_{14},u_{10}
u_{14}	u_5,u_{13}

But now a day's these social sites are suffering from the critical problem of cyber crime. This thing is happening due to open space functionality of these sites. There are many internet applications which are working today. If any person commits a cyber crime and interact with the social networking sites then by identifying the different activities of that particular person we can jump to its complete group even those being included in that cyber crime[11]. For this work the log data of the person is collected which gets available on the server by reading this log data we can suspect the culprit. The log file is available at The

DOI: 10.12948/issn14531305/18.2.2014.02

Internet Traffic Archive sponsored by ACM SIGCOMM [5]. This log data later partitioned on the basis of its attributes which can be done by applying data mining techniques. By doing this if any person proves out to be criminal then its complete group who was included in that crime could be traced. For this work a special technique for data mining that is frequent pattern mining is used. When a particular user click or visit the particular link then the all activity which has been performing by the user are have shown on the Fig. 2 identified through log data and maintained by the server or database. Some social networking sites follow the unauthorized showing mechanism. And its overall content follows the mechanism or open space. Being an open source code any unauthorized user to these particular sites can click on any knowing or unknowing link[6][12][2]. Apart of it he can send any types of message too.

In a broad area cyber crime is also called as computer related crime this is an illegal work committed by any means, or in relation to a computer system or network. Including such kind of illegal distribution information by means of computer system or network. Cyber crime is of many types against any person, against any business or non business organization committing crime by targeting the government [17]. There are few more examples of it like financial claims, money laundering, credit card frauds, sale of illegal articles including weapon and wild life article, posting any article to your website brought from some other needs. Intellectual crime includes software piracy, copyright stuffs. Email spooling include sent the mail from one source to another source. Forgery include currency host stamp, making marks sheet through printer and scanner.

2 Literature Review

kim W. et al [8] attempt to organize the status, uses and issues of social websites into comprehensive framework for discussing, understanding, using building and forecasting the future of social websites. Chu H. et al [3] provide the real case review concerning the crime sense reconstruction with respect to the previous facebook session of the victim based on the digital evidence collected and analyzed via live internal data acquisition and they also identified victim according to its log data which is maintained by the server. They have converted that log data to hexadecimal number and also found its mutual friends who also involved with the victim. If a person involved in a crime, according its session their conversation may be happened and they mined that data and identified victim and his friends. They also developed a framework of social network according to facebook sessions.Wu H. et al. [16] analyze the characteristics of users in social networking websites as well as the related contents of the website. They provide a technique of social network analysis and web mining to illustrate the networks of the blog users and according to that they find out the interest group. Ortigosa A. et al [10] Shown the data from online social networks profiles, specifically facebook profiles, can be mined in order to build classifier able to predict user personality and they also developed TP2010 applications which is general questioners to get the user personality and also applied classification tree. Jin L. et.al [7] we show that it also raises significant privacy concerns as an adversary can use it to find out some or all of the victim's friends, although, as per the privacy settings of the victim, the adversary is not authorized to see his friend list directly. We show that by using mutual friend queries, an attacker can launch privacy attacks that we refer to as mutual-friend based attacks to identify friends and distant neighbours of targeted users. We analyze these attacks and identify various attack structures that an attacker can use to build attack strategies, using which an attacker can identify a user's friends and his distant neighbours. Through simulations, we demonstrate that mutual-friend based attacks are effective. For instance, one of the simulation results show that an attacker using just one attacker node can identify more than sixty percent of a user's friends.

DOI: 10.12948/issn14531305/18.2.2014.02

3 Proposed Architecture

Today's Social Networking Services are not restricted to the user. Any number of users can interact with it. And this number of user performs social activities daily on these sites like messaging, status updating etc [1]. These different social networking may also include criminal activity. Like any particular may misuse the information of the other person. Cyber criminals are the person of same type which do the same and misuse the desired information. Social Networking Services evidence is find out according to the session. Social Networking Services (facebook, orkut etc.) is a platform through which we perform the routine activities. In these social networking sites if any person gets register, during the registration his email address get identified, which works as a unique identifier for different users. Facebook orkut are the social networking services that follows global indexed architecture in which any user can Id be identified on the basis of its daily log activities. If any anonymous user performs any unethical task then his behaviour could be predicted through log and can know that at what time he performed? What all activities? Which could be an important thing in a digital evidence of a criminal case? For finding the digital evidence of any criminal case. Here we tried to analyze the log of social networking sites. Here a log file is split on the basis of its attributes, out of which we analyze one column access page. This column consists of collection of different pages visited by the user or hit by the user. Out of this different collection we only mentioned those pages which consist of image file format extension like .jpg, jpeg, .png etc. If any visitor likes or hit any image file of social networking sites on any particular time and it comes into light that particular person is involved in criminal case then we have to filed that particular session how many user hit that same image or it particular image? For such purpose we used the current time on which any criminal activity has been performed. For this purpose we converted this data into hexadecimal number because by converting this data to hexadecimal number it become memory efficient and makes the computation fast comparatively. We also used it by the point of view of cryptography so that any person can hit and interact directly with your data. In hexadecimal conversion bit modification becomes quite easy.

For analyzing the log data maintained by the server, or if any person is involved in criminal case, then for identifying that person, it is really necessary to find a common pattern in that. Which can only be possible by analyzing the different attributes of log data? In case of cyber crime in Social Networking Services play a vital role as a clue. It consists of many steps that could be helpful in this kind of investigation.

3.1 User Activity

When a user click or visit the particular link then the all activity which has been performing by the user are identified through log data and maintained by the server or database.

3.2 Web Log Collection

In this step, the time at which any crime had taken place on any social networking sites, the complete data of that particular duration is fetched. For example if on 15-Jan -2012 any crime get conducted then we can take the complete data of that day for investigation. In this data, different type of attributes are inbuilt like IP address, user name, timestamp, access request, status code, referrer etc.

3.3 Preprocessing of Web Log

According to the given time, for analyzing the current log, we will use the concept of data mining in which we will apply the data preprocessing technique[15][14]. Through preprocessing technique this log data could be simplified to a lot. Data can be cleaned through it or we can say noise could be removed from data.

3.4 Analyze Access Page Attribute

After splitting the log data separately, here only a column 'access page' is analyzed in which many type of information is present.

DOI: 10.12948/issn14531305/18.2.2014.02

Whenever any visitor visits any particular site then this column will collect all information regarding the activities of the visitors that on which all data that particular visitor has visited. Here we will take only those pages which consist of image file format. If nay image suppose xyz.jpg get involved in that criminal case, or it has been accessed or used by any victim then we have to look forward that xyz.jpg named image has been visited or hitted by what number of user? Or how many number of users?. This could be checked only according to particular session. This will help us gaining that how many such user are there that hit that particular image on any time duration?

4 Hexadecimal Conversion

Whichever images have been identified, the page related to that image will be converted to hexadecimal number due to which it could be easily processed. In terms of cryptography it would be easier for privacy and modification of bits is also easier for computation, and we will take only those image files which exist in that session and will also count the occurrence of that page that how many times that page has occurred in that session? We displayed a table here according to the user in which it is shown that how many users hit which all image? At what time?

5 Result

After counting an occurrence of all images we have to analyze user who visited or viewed that images and apply data mining techniques we have identified restricted users.

By applying all above steps we have proposed architecture which is shown in Figure 3.

6 Experimental Result

First of all we have rearranged our web log data according to its attribute as we shown in fiure4. Every user may registered in social networking services like facebook, orkut, LinkedIn etc. and that kinds of service provide a unique number or unique Id for a particular user. But through same IP address of a system any number of user can login to various social networking services, through IP address it is very complex to discriminate user exist in social networking services.

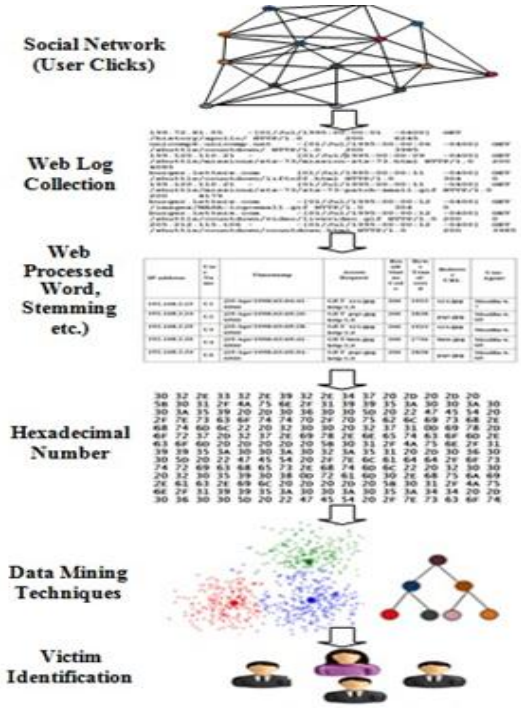

Fig. 2. Victim Identification Framework


192.168.2.1 U1 [24/Oct/1994:13:41:41 -0600] "GET index.html HTTP/1.0" 200 150
192.168.2.1 U2 [24/Oct/1994:13:41:41 -0600] "GET lmn.jpeg HTTP/1.0" 200 1210
192.168.2.1 U1 [24/Oct/1994:13:43:13 -0600] "GET index.html HTTP/1.0" 200 31
192.168.2.2 U2 [24/Oct/1994:13:43:14 -0600] "GET xyz.jpeg HTTP/1.0" 200 2555
192.168.2.2 U2 [24/Oct/1994:13:43:15 -0600] "GET pqr.jpg HTTP/1.0" 200 36403
192.168.2.17 U3 [24/Oct/1994:13:43:17 -0600] "GET abc.png HTTP/1.0" 200 441
192.168.2.22 U4 [24/Oct/1994:13:46:45 -0600] "GET index.html HTTP/1.0" 200 31
192.168.2.1 U2 [24/Oct/1994:13:46:45 -0600] "GET xyz.jpeg HTTP/1.0" 200 2555
192.168.2.2 U4 [24/Oct/1994:13:46:47 -0600] "GET pqr.jpg HTTP/1.0" 200 36403
192.168.2.22 U2 [24/Oct/1994:13:46:50 -0600] "GET abc.png HTTP/1.0" 200 441
192.168.2.12 U5 [24/Oct/1994:13:47:19 -0600] "GET lmn.jpeg HTTP/1.0" 200 1210
192.168.2.5 U2 [24/Oct/1994:13:47:41 -0600] "GET index.html HTTP/1.0" 200 318
192.168.2.67 U3 [24/Oct/1994:13:47:50 -0600] "GET 5.html HTTP/1.0" 200 2065
192.168.2.77 U2 [24/Oct/1994:13:48:11 -0600] "GET 6.html HTTP/1.0" 200 1124
192.168.2.89 U2 [24/Oct/1994:13:48:23 -0600] "GET 7.html HTTP/1.0" 200 1538
192.168.2.2 U3 [24/Oct/1994:13:48:51 -0600] "GET 8.html HTTP/1.0" 200 322
192.168.2.2 U4 [24/Oct/1994:13:49:01 -0600] "GET index.html HTTP/1.0" 302 -
192.168.2.67 U5 [24/Oct/1994:13:49:01 -0600] "GET index.html HTTP/1.0" 200 999

Fig. 3. Identify Images from Web Log

Converting web log in to hexadecimal number due to which it could be easily processed. In terms of cryptography it would be easier for privacy and modification of bits is also easier for computation as we shown in the Figure 4.

DOI: 10.12948/issn14531305/18.2.2014.02

Fig. 4. Hexadecimal Form of Image Page

Any user which is in social network can visit same or different image when same user hit a particular image again and again all activities are resides in the server or database in terms of web log with different timings . In table 1 we showed the visited image by the user with different timing. We are showing a snippet of visited image. Suppose a crime gets conducted 15-jan-2005 then we can take all logs on that day for investigation. An investigation depends on how many users visited a particular image? At what time? Etc.

For splitting our log data we have accomplished that task using java programming. Any user can interact with different kind of social networking services moreover using same IP we can interact with various social networking services but their timings are different.

As we shown that Fig. showing number of pages which had visited by the user. When a crime gets conducted then we have analyzed all pages of access page on that day and get how many people involved on that day? Page column contain all activities which is done by the user. Any number of users is directly or indirectly linked together. We have divided our log data one attribute name as timing. Whenever an user viewed images then log maintain their timing. If any criminal activities get conducted, then we have included all data with respect to its timing.

Table 2. An information of web log

SNS-ID	USERS	ACCESSED IMAGE	TIME
1001	u_1	xyz.jpg	01:01:02
1002	u_2	lmn.jpeg	01:03:05
1003	u_3	pqr.png	01:20:20
1004	u_4	abc.gif	01:15:20
1005	u_5	efg.jpeg	02:30:02
1006	u_6	rst.bmp	02:50:45
1002	u_2	xyz.jpg	03:10:10
1007	u_7	ghj.jpg	03:44:49
1003	u_3	xyz.jpg	03:55:55
1002	u_2	xyz.jpeg	04:12:33
1003	u_3	pqr.png	04:44:23
1008	u_8	uio.jpeg	05:27:33
1009	u_9	wer.jpg	05:50:26
1002	u_2	xyz.jpg	05:58:29
1003	u_3	lmn.jpeg	06:44:24
1002	u_2	lmn.jpeg	06:56:54
1004	u_4	xyz.jpg	07:34:02
1002	u_2	lmn.jpeg	07:40:06
1003	u_3	xyz.jpg	08:33:43
1002	u_2	xyz.jpg	09:10:51

After splitting log data to a column. We have taken page access column. We have taken only those access pages which contain images of format jpg, jpeg, gif, bmp, png etc. we have also counted the frequency or number of occurrences of these pages. Suppose there is an image xyz.jpg then we will see how many times that image have occurred in whole day. If a user hits on an image several time that also increase its count. This varies user to user. This will also help us to know how many times user u1 hit an image in a particular time period. Suppose an image xyz.jpg has been 20 times a day then that image is involved in criminal case and this user u1 is a restricted user. And this is also important to find out, how many mutual friends of u1 have visited the image xyz.jpg. That is why we have implemented everything in java. This increases the frequency of a particular page and also let us knows the occurrence of that image in particular session. In this way the image with the highest frequency will be selected and the user visited that image will be found out.

DOI: 10.12948/issn14531305/18.2.2014.02

Table 3. Frequency of images clicked by visitors

SNS-ID	USERS	AC-CESSED IMAGE	FRE-QUENCY
1001	u_1	xyz.jpg	12
1002	u_2	lmn.jpeg	25
1003	u_3	pqr.png	28
1004	u_4	abc.gif	10
1005	u_5	efg.jpeg	29
1006	u_6	rst.bmp	20
1002	u_2	xyz.jpg	12
1007	u_7	ghj.jpg	03
1003	u_3	xyz.jpg	26
1002	u_2	xyz.jpeg	12
1003	u_3	pqr.png	13
1008	u_8	uio.jpeg	22
1009	u_9	wer.jpg	23
1002	u_2	xyz.jpg	14
1003	u_3	lmn.jpeg	16
1002	u_2	lmn.jpeg	13
1004	u_4	xyz.jpg	14
1002	u_2	lmn.jpeg	22
1003	u_3	xyz.jpg	24
1002	u_2	xyz.jpg	19

In table3 we have shown the frequency of clicked image by the visitor. We have taken support of clicked image through visitor is 25. If it is greater than 25 and then we observe that u3 and u5 are the users who clicked more than 25 times the image. It means at that time u3 and u5 are involved in the image hitting. It shows that according to clicked frequencies, these users are the victims. In this Fig. , the clicked images through user and hit count are shown. In this user u3 on that day visit three images again and again. That images are pqr.png, lmn.png and ghi.jpg, pqr.png are clicked 28 times by user u3, lmn.png clicked 13 times by u3 and ghi.jpg clicked 16 time by user u3.If our support is greater than 25, then pqr.png image involved in it. Similarly user u5 has visited only 2 images i.e. efg.jpg and uio.jpg, efg.jpg clicked 29 time by u5 and uio.jpg clicked 11 times by u5.

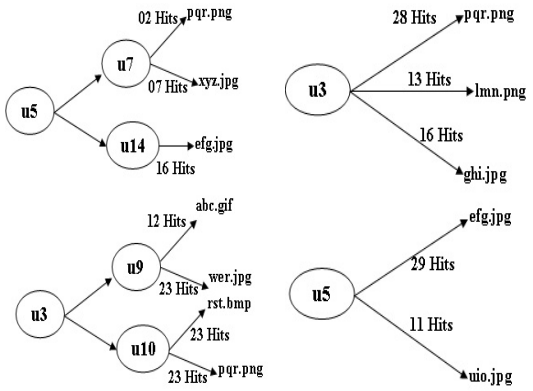

Fig. 5. Victim's Identification

The support we have decided above, according to that only efg.jpg image involved. After identify the visited image through user, now we have to search victim's mutual friends because we have to find that his mutual friends are also involved in it or not. For this, we will analyze the mutual friends of u3 and u5.

The image visited by mutual friends and the hit count. In this, we observe that u3 and u5 hit the pqr.png and efg.jpg image more of the times and also the mutual friend u7 and u10 have visited that image. From this we have concluded that rather than u3 and u5, the user u7 and u10 also visited image "pqr.png" again and again. Lastly u7 and u10 are the restricted user and finally four users (u3, u5, u7, and u10) as victims that are identified.

7 Conclusion

Now a days, large number of users are using social networking services and lots of social networking service providers are available to provide services at free of cost. Customers are participating in daily social life activities. In all types of service providers are playing the main role like to make user friendly websites provide user interactive features which attract customers, to provide effective communication between different users, provide protocols according to user needs. All concluded that cyber crime going to be increase rapidly. Social Networking Service include cyber crime means misuse of others information, unauthorized access, unethical

DOI: 10.12948/issn14531305/18.2.2014.02

posting of images etc. In this research, on the basis of web log we have identified the victim, for that we have analyzed different attributes of web log. If there is any criminal activity performed in any social site then we will take the web log of that day and from that log data it can be identified the user's behaviour on the basis of page click by users. For that we have preprocess that web log. In this we have find restricted user on the basis of number of hits on an image. Which is used as to restrict unauthorized access to social sites and it will also helpful for social network providers.

References

[1] B. Alsenoy, J. Ballet, Kuczerawy A., and Dumortier J., "Social networks and web 2.0: Are Users also Bound by Data Protection Regulations?", *Identity in the Information Society*, 2009.

[2] K. Andersen, and R. Medagli, "The Use of Facebook in National Election Campaigns: Politics as Usual*?, Lecture Notes in Computer Science*", 2009.

[3] H. Chu, D. Deng and J. Park, "Live Data Mining Concerning Social Networking Forensics Based on a Facebook Session Through Aggregation of Social Data" , *IEEE Journal On Selected Areas In Communications*, Vol. 29, No. 7, August 2011.

[4] B. Hajian and T. White, "Modelling Influence in a Social Network: Metrics and Evaluation", IEEE International Conference on Privacy, Security, Risk, and Trust, *IEEE International Conference on Social Computing*, 2011.

[5] *http://ita.ee.lbl.gov*, retrieved on March 12th, 2013.

[6] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang, Social contextual recommendation, *In KDD'12 Conference Proceedings*,

[7] L. Jin, D. Joshi and M. Anwar, "Mutual-friend based attacks in social network systems", *Computer and Security, Elsevier*, 2013.

[8] W. Kim, O. Jeong and S. WonLee, "On social Web sites", *Information Systems, Elsevier,* 2010.

[9] H. Lee, M. Choi, H. Park, and K. Park, "Social Networking Service based on Peer-to-Peer Network", *Third International Conference on Systems and Networks Communications, IEEE* 2008.

[10] A. Ortigosa , R. Carro and J. Quiroga (2013), "Predicting user personality by mining social interactions in Facebook", *Journal of Computer and System Sciences, Elsevier*.

[11] M. Pennacchiotti and S. Gurumurthy., Investigating topic models for social media user recommendation, *In WWW'11 Conference Proceedings*.

[12] J. Rana, J. Kristiansson, J. Hallberg, and K. Synnes, "Challenges for mobile social networking applications, in Lecture Notes of the Institute for Computer Sciences", *Social Informatics and Telecommunications Engineering,*.

[13] S. Sen, J. Vig, and J. Riedl, "Tagommenders: connecting users to items through tags", *In WWW'09 Conference Proceedings*, 2009.

[14] G. Thakur and B. Bakariya, "Preprocessing on Web Log Data in Web Usage Mining", *International Conference on Intelligent Computing and Information System (ICICIS)*, 2012.

[15] G. Thakur, B. Bakariya and K. Mohbey, "An Inclusive Survey on Data Preprocessing Method Used in Web Usage Mining", *presented in Seventh International Conference on Bio-Inspired Computing: Theories and Application, (BIC-TA 2012), ABV-Indian Institute of Information Technology and Management Gwalior*, December 14 - 16.

[16] H.Wu, I.Ting and K. Wang, "Combining Social Network Analysis and Web Mining Techniques to Discover Interest Groups in the Blogspace", *IEEE* 2009.

[17] M. Yuan, L. Chen, P. Yu and T. Yu, "Protecting Sensitive Labels in Social Network Data Anonymization", *IEEE Transactions on Knowledge and Data Engineering*, VOL. 25, NO. 3, 2013.

DOI: 10.12948/issn14531305/18.2.2014.02

Brijesh BAKARIYA received Graduation degree From Barkatullah University Bhopal MP in 2005 and Post Graduation Degree in Computer Applications from DAVV Indore in year 2009. He is currently pursuing the PhD. Degree in the Department of Computer Applications, Maulana Azad National Institute of Technology Bhopal. M.P. His Research interests include Web Mining and Clustering.

Ghanshyam Singh THAKUR has received BSc degree from Dr. Hari Singh Gour University Sagar M.P. in 2000. He has received MCA degree in 2003 from Pt. Ravi Shankar Shukal University Raipur C.G. and PhD degree from Barkhatullah University, Bhopal M.P. in year 2009. He is Assistant Professor in the department of Computer Applications, Maulana Azad National Institute of technology, Bhopal, M.P. India. He has eight year teaching and research experience. He has 26 publications in national and international journals. His research interests include Text Mining, Document clustering, Information Retrieval, Data Warehousing. He is a member of the CSI, IAENG, and IACSIT.

DOI: 10.12948/issn14531305/18.2.2014.02