2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)
19-20 December 2015, BUET, Dhaka, Bangladesh

PFU: Profiling Forum Users in Online Social Networks, A Knowledge Driven Data Mining Approach.

Vasanthakumar G U, P Deepa Shenoy, Venugopal K R

Department of Computer Science and Engineering,
University Visvesvaraya College of Engineering, Bangalore, India.
E-mail: vasanthakumar.gu.in@ieee.org

***Abstract*—Online Social Networks (OSNs) provide platform to raise opinions on various issues, create and spread news rapidly in Online Social Network Forums (OSNFs). This work proposes a novel method for Profiling Forum Users (PFU) by exploring their behavioral characteristics based on their involvement in various topics of discussion and number of posts in respective topics posted by them in OSNFs dynamically. Modeling the proposed method mathematically, the PFU algorithm is illustrated for its adequacy and accuracy.**

***Keywords*—Data Mining; Forum; Opinion; Online Social Network; Profiling; Topic of Discussion;**

I. INTRODUCTION

The use of internet and social networking has been exponentially increasing since a decade. The emerging trend of social networking has evolved with huge transactions between the individuals in mass over the internet. Opinions of individuals, their interaction patterns and actions on the weblogs over the social networking sites like Facebook, Twitter, Flickr, Myspace, LinkedIn, Google etc., are generating huge data every minute. Various data mining techniques [1] [2] may be used to profile Forum Users over the social networking sites.

Online forums provide platform for social network users to create their own network of users and to share their opinions and have discussions based on their topic of interest. Huge set of discussions happen on various topics every day. Here, the topic of discussion is not pre-defined, instead depends on users interest. Since there is no restriction on the topics, any user can share their opinion or knowledge on any number of topics at a time. The main challenge lies in identifying the topic of discussion and the set of users involved in those discussions.

Unlike previous generations, present youth share their opinion and discuss about various issues and matters in public using forums designed for online discussions. Internet forum, also sometimes called as Message board, is a platform where users hold conversations on various topics in the form of posted messages. The conversations here are completely different from what happens in chat rooms, since the conversations here will be long and users share their opinion clearly without any sparse. In forums, users will have access levels based on which they post their opinions while moderators verify and approve the post before publishing. Few sites provide free access to users and while few require registration and login before accessing the site. An unregistered user, called guest is granted the privilege to read through the discussions of registered users but not to either post or to alter the database.

A tree-like hierarchical structure is maintained with respect to the discussions in the forums. The main discussion may get branched into many sub forums with new discussions and topics evolving over time. Sub forums further can have many sub forums and so on. If any user posts a thread in the present topic, a new discussion can start based on that new thread, when others in the forum reply to that thread. As replies, opinions increase, evolution of new topics may increase. When a post is submitted, the details like date, time and user names etc., are logged. The collection of posts, a user would have submitted from oldest to latest is termed as *thread*. The information present in the thread includes all the posts and helps in tracing the topics of interest over time of an user.

Here the discussions/threads/posts submitted by an Influential User [3] may influence others in the network and may also lead to change an individual's opinion. The political, national, women harassment etc., movements in society may also influence the forum discussions. The topics in the forum discussions may die quickly or extend longer, which purely depends on the users' interest. Forums do not restrict its users on the topic of discussions or there is no time limit on the discussions, hence the number of topics under discussions may vary over time.

Motivation: As the users in the social network increase, the chances of their discussions on different topics also increase relating to politics, cinema, technology, nature etc., which motivated us to carry out this work. In few forums, the users start discussions after creating titles and posting related messages. But from the replies to the posted messages, new topics may evolve resulting in new sub forums, which initiates the need for identifying such new topic(s) of discussion and the involvement of users in them.

Contributions: This paper attempts to explore the behavioral characteristics of Forum Users in Online Social Network Forums. Considering a set of users in the forum discussion, an attempt is made to Profile the Forum Users based on their involvement in various topics of discussion and number of posts in respective topics posted by them in OSNFs dynamically as presented in PFU algorithm.

978-1-4673-8786-6/15/$31.00 ©2015 IEEE

The rest of the paper is organized as follows: Section-II gives a brief review of the literature survey. The problem definition is discussed in Section-III. The proposed system model is presented along with the illustration of PFU approach in Section-IV. The mathematical model is formulated in Section-V. The Profiling Forum Users (PFU) algorithm is presented in Section-VI and the Conclusions are drawn in Section-VII.

II. LITRATURE SURVEY

Adrian M.P et al. [4] presented a tool to extract the data of individuals from different social networking sites providing different perspectives to visualize their personal, professional or even social interests over time. Wang et al. [5] proposed a visual analytical system to depict the five W's concept in investigative analysis. Identifying strategic patterns, trends and exploring tactical incidents are easier with this tool. Kempe et al. [6] proposed an approximation algorithm based on degree and distance centrality which is useful in identifying the influence of a node in social networks.

Das et al. [7] proposed Automatic Clustering Differential Evolution (ACDE) algorithm requiring no prior knowledge of the data to be classified. Experiments conducted on Iris plant Database, Glass Wisconsin and Breast Cancer data sets shows that ACDE performs much better than classical DE-based clustering scheme. Based on the density, Bicici et al. [8] proposed Locally Scaled Density Based Clustering (LSDBC) algorithm having ability to identify clusters of arbitrary shape on noise backgrounds that contain significant density gradients. Experiments conducted shows that LSDBC can be used as a tool for summarizing the inherent relationships within the data as well as summarize and segment images. Phan et al. [9] proposed an approach to classify short and sparse text and web with hidden topics from large scale data collection. Experiments conducted on Wikipedia with likely words and data sample topics of Ohsumed Medline data shows that the approach attains more accuracy even with less training samples than other approaches.

Martin et al. [10] proposed Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm which discovers clusters in large spatial databases with noise. Experiments conducted on synthetic sample database and SEQUOIA 2000 database shows that the run time of DBSCAN is slightly higher than linear in the number of points and is more effective than Clustering Large Applications based on RANdomized Search (CLARANS) algorithm in discovering clusters of arbitrary shape. Based on two important directions:- clustering point objects, spatial and their non-spatial attributes, Generalized Density-Based Spatial Clustering of Applications with Noise (GDBSCAN) algorithm is proposed by Ram et al. [11]. Results obtained after experimenting shows effectiveness and efficiency of GDBSCAN on large spatial databases. For web opinion clustering, Dcouth et al. [12] proposed a technique called Scalable Distance-based Clustering (SDC) and also an interactive visualization tool which displays social network development, the network topology, similarity between topics, and the similarity values between participants. Experiments conducted on "Al Qaeda", a social network data set shows SDC achieves good performance with its limitations in clustering web opinions than DBSCAN, and that the SDC performs better with both micro accuracy and macro accuracy.

Ertoz et al. [13] proposed an approach to find clusters of different sizes, shapes and densities with noise by developing Shared Nearest Neighbor Clustering (SNN) algorithm. Experiments conducted on NASA earth science data and KDD Cup 99 network intrusion data shows SNN clustering algorithm achieving good results with respect to the Jarvis-Patrick approach to detect the nearest close node to form clusters. Iakovidis et al. [14] proposed a model describing knowledge extracted from the lowest-level of data mining process where information is represented in multiple. The clustering is performed by Non-negative Matrix Factorization (NMF) and the domain knowledge is used to automatically annotate two unlabeled clusters per feature space. Bollegala et al. [15] proposed a method that considers page counts and text snippets returned by a web search engine using i) Extract patterns from snippets and ii) GetFeatureVector(A,B) algorithms. Experiments were conducted with different kernel types on Benchmark data set and found that it performs efficiently. Mei et al. [16] defines a Probabilistic approach to model subtopic themes and spatiotemporal theme patterns simultaneously. The experiment shows that the proposed model performs well for different types of topics and can reveal interesting spatiotemporal patterns in weblogs.

By utilizing the network information features to extract Latent Social Dimensions based on Network Information and Discriminative Learning, Lei Tang et al. [17] proposed a Relational Learning Framework. Experiments conducted on BlogCatalog6 and Flickr7 data sets demonstrate that the proposed social dimension approach outperforms alternative relational learning methods. Cetints et al. [18] proposed to identify relevant and irrelevant micro-blogging questions asked in a classroom. The micro-blogging tool, called HotSeat is used to automatically identify relevant and irrelevant questions. Mei et al. [19] focused on Evolution Patterns (EPs) in sequences of documents which identifies and discovers event episodes together with temporal relationship that occurs frequently. To embed undefined relevant weblogs gathered through topic-specific exploration by analyzing and visualizing weblog social network, a framework is proposed by Tang et al. [20] which helps in identifying the influencing blogs in the network. Experimental results obtained by conducting on Xanga.com and Google blog data sets show that this method is tremendous to trap directly into the millions of minds. Veena H Bhat et al. [21] presented a spam filter to predict the category of mails by studying the content of mail as well as its behavioral characteristics.

III. PROBLEM DEFINITION

Based on various issues, the interests of users for posting their opinions in OSNFs change, which considerably depends on their day-to-day life activities. Another point which directly influences on the topics under discussion is the *time*, with different topics evolving over time. The challenge in forum discussions is in finding the topics on which the discussions are going on. There are many methods available to find the topics underneath the discussion, but the existing methods have assumed few pre-defined topics of interest because of which, the new topics the users may start to discuss with becomes difficult to capture and hence those methods are not so efficient in finding evolving topics of discussion.

The main objective of this work is to develop a method which identifies dynamically the evolving topics under discussion in OSNFs initially with zero topics and growing based on the words used in the posts considering a set of users in the forum discussion and Profile the Forum Users according to their involvement in various topics of discussion.

IV. PROPOSED SYSTEM MODEL

The topics under discussion depend on the users' interest, national issues, political matters or health tips etc., but some discussions may influence others in the network. As discussions go deeper, individuals may change their opinions or few users who have confusions get clarified by involving in the discussion. Hence there exists necessity to define / identify the characteristics / behavior of the user in OSNFs.

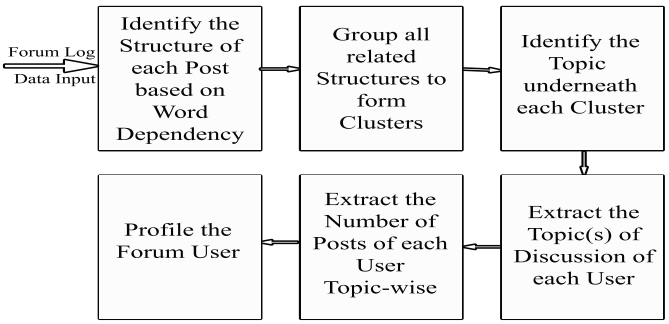

Fig. 1: System Architecture

Our proposed PFU method is applied on the collected forum log data set and the related structures of posts in the data set are grouped together to form clusters and then the topics underneath the clusters are identified. Over the time, if some structures occur relating to already defined/identified clusters, then those structures are automatically moved to the respective clusters, whereas other structures which do not belong to any of the pre-identified clusters, such structures are again grouped to form clusters based on their structure similarities. Hence new clusters are formed and their underneath topics are identified. The topics of discussion in which the users are involved for that period of time are extracted, which gives their interested topics of discussion. Based on their involvement in various topics and number of posts in respective topics posted by them in OSNFs for the considered period of time, the forum users are profiled. Fig. 1 shows the proposed system architecture with systematic flow.

A. Illustration

For the purpose of illustration, a case study is considered where an user from each cluster is considered and their involvement in various topics of discussion are computed and analyzed over a period of time. Table-I shows the number of posts posted by each forum user in various topics of discussion. As can be observed in the table, the Forum User with identification FU-1 is involved in three different topics of discussion having posted 19, 58 and 7 posts in each of the cluster Topics T-2, T-3 and T-5 respectively. Thus the user FU-1 profiled to be fond of topic T-3, while having additional interest in T-2 and T-5 topics as well.

TABLE I: PFU Approach Illustration

Forum Users / Topics	T-1	T-2	T-3	T-4	T-5
FU-1	0	19	58	0	7
FU-2	4	25	0	6	0
FU-3	0	0	0	0	75
FU-4	35	9	5	0	0

V. MATHEMATICAL MODEL

The table of notations shown in Table-II describes the symbols used in mathematical modeling of the proposed method.

TABLE II: Table of Notations

$Symbols$	$Description$
P_f	Forum Post
P_n	Name of the User who posted the Post
q	Number of Posts
S_p	Structure of the Post
K_s	Prominent Stemmed Keyword(s)
m	Number of Structures
C_s	Cluster of Structures
T_d	Topic of Discussion
n	Total number of Topics / Clusters
U_f	Forum User
L	Number of Users
U_p	Profile of User

Number of Users involved in considered cluster is:

$$L = \sum_{j=1}^{m} [P_n \cup P_n(j)] \tag{1}$$

Number of Posts posted by an User is:

$$q[U_f(i)] = \sum_{j=1}^{m} [P_n(i) \cap P_n(j)] \tag{2}$$

where i=1 to *L*.

Structure of the Post is the most prominent stemmed keyword(s) of that post based on word dependencies:

$$S_p = prominent[K_s] \tag{3}$$

Clusters are formed based on the similarity between the structures:

$$C_s = \sum_{i=1}^{m} \sum_{j=1}^{m} sim[S_p(i), S_p(j)] \tag{4}$$

Users involved in various Topics of discussions is:

$$U_f(i) = \sum_{C_s=1}^{n} [P_f(i) \cap P_f(C_s)] \tag{5}$$

Profile of Forum User is:

$$U_p = \sum_{k=1}^{n} CumulativeAnalysis(q[U_f(k)], T_d(k)) \tag{6}$$

The cumulative analysis of the number of posts and the respective topics of involvement of the user helps in profiling the forum user.

VI. ALGORITHM

Algorithm 1 *Profiling Forum User (PFU) Algorithm*

```
1: while True do
2:     for Every Post in Forum do
3:         Figure out Structure of Post based on Word Dependencies
4:         if clusters already exist then
5:             if Structure of Post matches to existing clusters then
6:                 Move the structure to the relevant cluster
7:             else Group separately
8:             end if
9:         else Group separately
10:        end if
11:    end for
12:    Form new clusters according to Structures Similarity
13:    Analyze newly formed clusters and identify their topics
14:    for Every Cluster do
15:        Identify the users involved
16:        for Every Identified User in the Cluster do
17:            Compute their number of posts
18:        end for
19:    end for
20:    for Every Distinct User do
21:        Compute Number and Topics of involvement of the User
22:        Profile User based on their Number of posts in each Topic.
23:    end for
24: end while
```

The algorithm runs on the data set collected and the structure of each and every forum post is identified and are clustered based on their word dependencies. The topic of discussion of each cluster is identified. In each and every cluster, the users involved are identified along with the number of posts posted by them. The algorithm is repeatedly run over the collected data every hour, and later, the forum users are profiled based on their involvement in various topics of discussion over time.

VII. CONCLUSIONS

The paper presents PFU algorithm for profiling forum users based on their involvement in various topics of discussion and number of posts in respective topics posted by them in OSNFs dynamically. The behavioral characteristics of forum users in OSNFs are explored to discover the knowledge out of it. The proposed algorithm identifies dynamically the evolving topics under discussion based on the words used in the posts and profile the forum users accordingly. Modeling the proposed method mathematically, the PFU algorithm is illustrated for its adequacy and accuracy.

The PFU algorithm when applied to Online Social Network Forums, helps in identifying the criminal activities along with the criminals involved with it. Our proposed algorithm can also be used by News Media to instantaneously get the news from across the world and broadcast before it can be known to any other media, while increasing their viewer's rating. The avenues for future work are in profiling the forum users based on the time stamp of the posts with their geographical location.

REFERENCES

[1] P Deepa Shenoy, Srinivasa K G, Venugopal K R and Lalit M Patnaik, "Evolutionary Approach for Mining Association Rules on Dynamic Databases," *Advances in Knowledge Discovery and Data Mining*, pp. 325–336, April 2003.

[2] P Deepa Shenoy, Srinivasa K G, Venugopal K R and Lalit M Patnaik, "Dynamic Association Rule Mining using Genetic Algorithms," *Intelligent Data Analysis*, vol. 9, no. 5, pp. 439–453, September 2005.

[3] Vasanthakumar G U, Bagul Prajakta, P Deepa Shenoy, Venugopal K R and Lalit M Patnaik, "PIB: Profiling Influential Blogger in Online Social Networks, A Knowledge Driven Data Mining Approach," *Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015), Procedia Computer Science, Elsevier B.V.*, vol. 54, pp. 362–370, August 2015.

[4] Adrian M.P. Braoveanu, Alexander Hubmann-Haidvogel, and Arno Scharl, "Interactive Visualization of Emerging Topics in Multiple Social Media Streams," *ACM Conference*, May-2012.

[5] S.Wang, E. Milar and W.Ribarsky, "Investigative Visual Analysis of Global Terrorism," *EUROGRAPHICS 2008 IEEE - VGTC Conference on Visualization*, vol. 27, no. 03, pp. 919–926, 2008.

[6] D. Kempe, J. Kelinberg and E. Tardos, "Maximizing the Spread of Influence through a Social Network," *Department of Computer Science, Cornell University, Ithaca NY*, pp. 137–146, 2003.

[7] S. Das, A. Abraham and A. Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm," *IEEE Transactions on System, Man, and Cybernetics-Part A: Systems and Humans*, vol. 38, Jan-2008.

[8] E. Bicici and D. Yuret, "Locally Scaled Density Based Clustering," *Koc University Rumelifeneri Yolu Sariyer Istanbul, Turkey*, Apr-2007.

[9] X.H. Phan, L.M. Nguyen and S.Horguchi, "Learning to Classify Short and Sparse Text and Web with Hidden Topics from Large Scale Data Collection," *WWW 2008 / Refereed Track: Data Mining Learning*, Apr-2008.

[10] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *2nd International Conference on Knowledge Discovery and Data Mining*, 1996.

[11] A. Ram, S. Jalal, A.S. Jalal and M. Kumar, "A Density based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases," *International Journal of Computer Applications*, vol. 03, no. 06, Jun-2010.

[12] J.R. Dcouth and T. Mohanraj, "Analyzing and Extracting Social Mining Trends Through Web Opinion Developments via Density Based Clustering," *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*, vol. 02, no. 01, Jan-2013.

[13] L .Ertoz, M.Steinbach and V.Kumar, "Finding Clusters of Different Sizes, Shapes and Densities in Noisy, High Dimensional," *Department of Computer Science, University of Minnesota*, Feb-2010.

[14] Iakovidis D and Smailis C, "A Semantic Model for Multimodal Data Mining in Healthcare Information Systems," *Stud Health Technological Inform*, 2012.

[15] D.Bollegala, Y. Matsuo and M. Ishizuka, "Measuring Semantic Similarity Between Words using Web Search Engines," *The University of Tokyo*, pp. 757–766, May-2007.

[16] Q. Mei, C. Liu and H.Su, "A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs," *Department of Computer Science, University of Illinois at Urbana-Champaign, Department of EECS, Vanderbilt University*, pp. 533–542, May-2006.

[17] Lei Tang and Huan Liu, "Relational Learning via Latent Social Dimensions," *ACM-KDD*, Jul-2009.

[18] S. Cetints, L .Sio and K. Bowen, "Micro Blogging in a Classroom: Classifying Students Relevant and Irrelevant Questions in a Micro blogging-Supported Classroom," *IEEE Transactions On learning Technologies*, vol. 04, no. 04, pp. 292–300, Oct-2011.

[19] Q. Mei and C. Zahi, "Discovering Event Evolution Patterns from Document Sequences," *IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans*, vol. 39, pp. 850–863, Jul-2009.

[20] L. Tang and H. Liu, "Terrorism and Crime Related Weblog Social Network Link, Content Analysis and Information Visualization," *IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans*, vol. 38, pp. 55–58, Jan-2007.

[21] Veena H Bhat, V R Malkani, P Deepa Shenoy, K R Venugopal and L M Patnaik, "Classification of Email using BeaKS: Behavior and Keyword Stemming," *IEEE TENCON*, 2011.