

Ranking Influential Users

Mane Priyanka B., Rathid Sonali S., Sanap Deepali D. &
Shirude Bhavana S.

MCOERC, Nashik, Pune, India.

Abstract: *The popularity of online social media is spread day by day for various online community purposes. Now, it has been found that social media is become use as a tool for a harmful act in sophisticated manner. This things also done in web forum, chat room etc. some forum are used for a open discussions on a critical topics influenced by radical thoughts. The influential users dominate the mind of naïve users using their radical thoughts. Influential users compel the naïve users to do wrong things. This system identify radical influential user from the web forum and rank this user according to their comment on forum. According to the user rank this influential users are removed from the forum.*

Key-Words: *Social media analysis, Security informatics, Influential user identification*

1. Introduction

Online social media is very popular for different kinds of activities such as online chat etc. Now, it has been found that social media is become use as a tool for a harmful act in sophisticated manner. There are hundreds of multimedia websites present on internet, these multimedia websites, online chat rooms play serious threats who are society as well as national security. These website provide support psychological war and extension of their agendas, whereas chat rooms and forums encourage their strategies and ideology through interacting with naïve users.

Many users available on the social media but some users generally avoid going through every comment posted by others. There always exist few users which maintain relationship of trust with other user and their comment attract by other users. These users are called as active users or influential users. These users sometimes called as community leaders. The agenda of influential user is to dominate the naïve users to do the frauds as well as wrong things. The role of influential users is a dominating role in the community and the activity and comment of this user is mostly affect the mind of other users. For example, personal blog popularity is totally depending on the influence of users. The majority of the users remain silent viewer. Some political leader or product marketing or radical ideology propagations find an easy way to attract silent viewer and encourage their ideology.

In Dark Web forum, the leader of radical groups maintains their own strategy to attract the silent viewer using their convincing approach. It is a very difficult task to search these influential leaders of radical groups propagating through the dark web forum. To identify these influential users in a network using high connectivity in the network, leadership influence over the network and interest on the network domain.

This application will design for rank the influential users who are fanatically convince to other people about wars, hate. Comments of the users will match with manually created dictionary and then level of radicalness of these users will calculate.

A contingency table generation method used for finding a pair of users based on their communication which is used to define eleven different collocations based association metrics. The association measures along with radical measure are built in PageRank algorithm. Then generate the list of radically influential users using PageRank algorithm. To find the radicalness of users we will establish five different criteria. Using these criteria we will calculate the score of influential users.

2. Literature Survey

We studied the previous some paper details to design our system. Matthew Richardson and Pedro Domingos works on different way of marketing and social network sites by mining knowledge sharing. In a market, try to sell the product viral marketing uses the customers. This public advertising is more cost effective than previous methods. Use of internet is more popular from the past few years. Because of social media people are interacting with each other, this interaction is saved in archives. On knowledge sharing sites social interaction is in variety of forms. One is some form of explicit trust between users. The limitations of this system are multiple sources of related information will be available but this system extracts a network from a single source. This system introduces a marketing plan when the structure of network is unknown or partially known [1]. Jialun Qin, Yilu Zhou, Hsinchun Chen works on dark web collection building process is given. This system use a systematic content analysis tool called the dark web attribute system. This dark web attributes system which is used to analyze and compare these extremist

organizations. Low efficiency of manual analysis approaches is the disadvantage of existing dark web. To overcome this disadvantage newly invented approach examine an integrated approach for collecting and monitoring dark web content This proposed semi-automatic dark web crawling approach which is more efficient than manual approach which combined the accuracy of human expert. The limitations of this paper are not more accurate attribute system. This is an inefficient web crawling techniques. In dark web there is no any extremist domain expert to better result. The dark web is not automatic web crawling [2,7,9]. Ilham Esslimani, Armelle Brun, Anne Boyer explain the term of the behavior of good leaders can be finding out by similarity measures. We studied in this paper detect the leader in the circumstance of behavioral networks. When users have same navigational behavior then they are connected. Higher accuracy ratios are, the more consistent the behavioral leaders are. In the context of behavior networks, Detecting leader is important fact. To detect the reliable leader's capacity of propagating appropriate appreciation and high connectivity of network is needed. The limitations of this paper are no other methods for detection of leaders and analyze their performance. In this paper there is problem of coverage. There is no additional dataset for validate the generalization of detecting leaders behavioral networks approach [3]. Michael Trusov, Anand V. Bodapati, Randolph E. Bucklin Explain by log activities of users to determine influential users on Social network is given. The collection of user profiles is base of an social networking site. Users those are register on social network enter the information on their profiles which they want to share. These users involve in two activities: either they create new content like uploading music, adding photos, writing on blogs, sending messages, or they collecting the content of other people such as downloading music, looking at photos, reading blogs. Recently Google's try to improve ad targeting at MySpace.Com to identifying influential users. The limitations of this paper are blowup is a theoretical possibility [4]. Denzil Correa design a system for the radicalization of different types on the web is given. Link Based for making decision on detection of radical information this feature Identify the structure between documents. Hyperlink between web sites is included in this feature. It also includes relationship between users on web site. Content Based this feature provide content and structure of documents lexical, syntactic, graph based , structural features , link based feature is include in content based. This feature mainly used in link based bootstrapping algorithm, content based feature used in text classification techniques. The limitations of this paper the techniques rely more on visual inspection using graph layouts which can be misleading and

unreliable [5].

3. Proposed System

Radicalization means intelligently does the competition with similar people like political, racial or any other ideology. The people who are undergoing this radicalization those people do not have value of ethics. These people are counted as a radical user

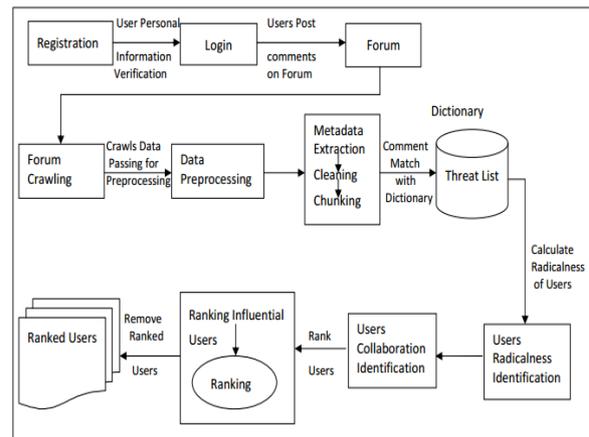


Figure1.System architecture

These people think about the war and hates because of they fill either unjust and discrimination or may not. They involve their personal, social issues in their act. The aim of our system is to rank these users and remove from the forum.

The proposed method starts with crawling and preprocessing the forum data followed by user radicalness identification, User collocation identification, and finally raking the user based on page rank algorithm.

A. Forum Crawling and Preprocessing:

The first step is data crawling and preprocessing. In this step URL of home page of the forum is passing to forum crawler. Then the crawler's crawls all related web pages and remove the redundancy. A platform-specific parser is in use to fetch meaningful part from the crawled web pages. This part of data is passed to the preprocessing module. To fetch related metadata the metadata fetch task works in congruence with parser module.

B. Measuring Radicalness:

In the previous work identify the radical users is based on discussion content. For this purposed the system uses the manually crafted list of threats word. To handle situation, updating of the list time by time

is required. The naïve bayes classification can be used to introduce new terms. In our system we will implement dictionary of words. System will continuously automatic check the comment of user with dictionary and then calculate the radicalness of the users according to their comments.

TABLE I

CONTINGENCY TABLE FOR A PAIR OF FORUM USERS (u_i, u_j)

	u_j	$U-u_j$	U
u_i	A	(b-a)	B
$U-u_i$	(c-a)	(d-c-b+a)	(d-b)
U	C	(d-c)	D

In the threat list Ω denotes the set of words. A measure of radicalness ρ is assigned to every user u_i of the forum, based on the existence of each word Ω_j

in each message post of u_i using equation (1), where $exists(\Omega_j, P_k^i)$ is a binary function which returns 1 if Ω_j exist in, otherwise 0.

$$\rho(u_i) = \frac{\sum_{p_k^i \in posts(u_i)} \sum_j exists(\Omega_j, P_k^i)}{\max\{\sum_{p_k^i \in posts(u_i)} \sum_j exists(\Omega_j, P_k^i)\}} \dots\dots\dots(1)$$

C. Identifying Collocations:

It has been found that there exists an intimate relationship between the users interacting in same thread, and in the context of Web forums the term collocation can be defined as the association of users co-interacting in same threads. Therefore we apply the collocation theory to study the associativity of different users, and estimate their influence while propagating an ideology through their interactions. To capture this information, a contingency table, shown in Table II, is constructed for each pair of users, where U is the set of users, and u_i and u_j represent two individual users. In this table, a denotes the number of instances (or threads) in which u_i and u_j have co-occurred, b denotes the number of instances (or threads) in which u_i has co-occurred with all other users in a thread, $(b - a)$ denotes the number of instances (or threads) in which u_i has co-occurred with all other users except u_j in a thread. Similarly, all other values in this table denote the number of instances (or threads) in which

interactions have taken place between the corresponding users.

D. Defining Association Metrics:

In this, we use cosine method.

Cosine:

Cosine is used to determine the strength of relationship between a pair of objects having attribute vectors. Cosine is formulated by $(X, Y) = \frac{|X \cap Y|}{\sqrt{|X|} \times \sqrt{|Y|}}$, Where X and Y represent the attribute vectors of same dimensions. We define the metric based on the contingency table to compute the association between two user's u_i and u_j using equation 2.

$$\mu_1(u_i, u_j) = \frac{a}{\sqrt{b} \times \sqrt{c}} \dots\dots\dots(2)$$

E. Ranking:

Page Rank algorithm is an important factor for radically influential user ranking method. For interaction between the users in a forum are used to construct a directed graph and adding every user in a forum as a node. Bi-directional links between each pair of commenter's and Uni-directional links from all commenter's to the thread initiator are recognized for each thread in the graph. Using PageRank algorithm every user node is initialized with the small value. This small value is considered as its PageRank score. To finding the PageRank Score of the Users we will compute the formula.

$$prob(p_i | p_j) = \frac{1}{out - deg\ ree(p_j)} \dots\dots\dots(3)$$

This equation is the transition probability from webpage p_j to webpage p_i . The transition process is continued until a convergence is achieved and the scores at that instance are accepted as their final PageRank scores.

4. Algorithm

Page Rank Algorithm

Cosine Method:

The cosine similarity between two vectors (or two documents on the Vector Space) is a measure that calculates the cosine of the angle between them. This metric is a measurement of orientation and not

magnitude, it can be seen as a comparison between documents on a normalized space because we're not taking into the consideration only the magnitude of each word count (tf-idf) of each document, but the angle between the documents. What we have to do to build the cosine similarity equation is to solve the equation of the dot product for the $\cos \theta$:
 Cosine is used to determine the strength of relationship between a pair of objects having attribute vectors. Cosine is formulated by

$$(X, Y) = \frac{|X \cap Y|}{\sqrt{|X|} \times \sqrt{|Y|}}$$

Where X and Y represent the attribute vectors of same dimensions. We define the metric based on the contingency table to compute the association between two user's u_i and u_j using equation 2.

$$\mu_1(u_i, u_j) = \frac{a}{\sqrt{b} \times \sqrt{c}}$$

Bayes Classifier:

Naive Bayes is one of the most effective and efficient classification algorithms. In classification learning problems, a learner attempts to construct a classifier

from a given set of training examples with class labels. Assume that A_1, A_2, \dots, A_n are n attributes. An example E is represented by a vector (a_1, a_2, \dots, a_n) , where a_i is the value of A_i . Let C represent the class variable which takes value $+$ (the positive class) or $-$ (the negative class). We use c to represent the value that C takes. A naive Bayesian classifier, or simply naive Bayes, is defined as:

$$c_{NB} = \operatorname{argmax}_{c_j \in C} P(c_j) \prod_{i \in \text{positions}} P(x_i | c_j)$$

5. Result Set

5.1 Blog



Figure 5.1: Blog

Above figure 6.1 is a blog page of project from which users and have options like registration. This page also show the comments posted by users.

5.2 Login Page



Figure 5.2: login form

In above form user should login by enter its correct user name and password.

5.3 Comment post

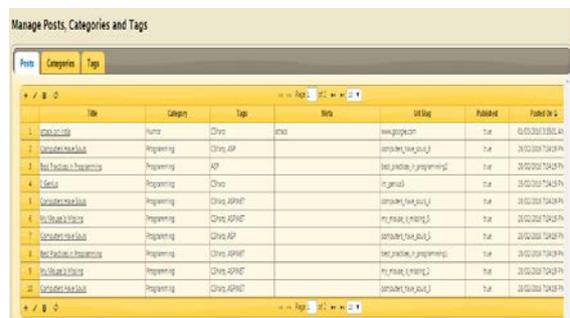


Figure 5.3: comment post by users

In this form users post the comment and also edit their comment. This page have categories and tags. Categories shows the human, programming etc. categories of comment.

5.4 Add Post

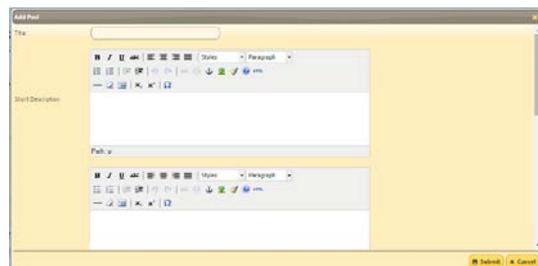


Figure 5.4: Add post by users

In this form user add new post with title, description, categories, tag, meta dada etc. user also give the url where comments are post.

5.5 Classification

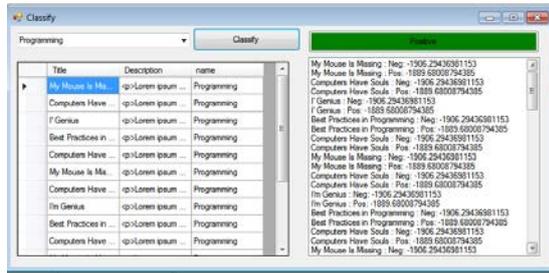


Figure 5.5: Classification of comments

In this form users comments are classify according to positive and negative status. User select the category of post they wants to classify.

6. Conclusion

In Ranking Radically Influential Web Forum Users we studied the advantages and feature of our system. And also study drawback of the existing system. In next phase we will start to implement the first module of our system which is forum crawling and preprocessing. This system is useful for removing the radical influential users from the web forum.

7. Acknowledgements

Expression of feeling by words, make them to lose their significance when it comes to make a statement of acknowledgement.

But, we are aware of the great sense of humanity and gratitude towards all those who have offered us their precious time and guidance.

Firstly, we would thank our Principal Dr. G. K. Kharate and HOD Dr. V. H. Patil for providing us with suitable media for performance and chance to prove our skills.

We would also like to express our gratitude to our project guide Prof. R. M. Gawande for his support, continuous volunteered guidance and tremendous contribution being so understanding and helpful throughout the project. We wish to express our sincere thanks to the project co-coordinator Prof. S. A. Bhavsar. We also thank to Prof. P. A. Andhale for his guidance. We thank with all our heart express our honor and deep gratitude for all staff members of Computer Engineering Department for helping us in every possible way for them.

8. References

- [1] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc. 8th ACM SIGKDD Int. Conf. KDD*, 2002, pp. 61–70
- [2] J. Qin, Y. Zhou, and H. Chen, "A multi-region empirical study on the Internet presence of global extremist organizations," *Inf. Sys. Frontiers*, vol. 13, no. 1, pp. 75–88, 2011.
- [3] Esslimani, A. Brun, and A. Boyer, "Detecting leaders in behavioral networks," in *Proc. Int. Conf. Adv. Soc. Newt. Anal. Mining (ASONAM)*, Aug. 2010, pp. 281–285.
- [4] M. Trusov, A. V. Bodapati, and R. E. Bucklin, "Determining influential users in Internet social networks," *J. Marketing Res.*, vol. 47, no. 4, pp. 643–658, Aug. 2010.
- [5] Denzil Correa, Ashish Sureka, "Solutions to Detect and Analyze Online Radicalization", in *Proc. IIITD PhD Comprehensive Report, Vol. V, No. N, November 2011*
- [6] T. Anwar and M. Abulaish, "Identifying cliques in Dark Web forums—An agglomerative clustering approach," in *Proc. IEEE ISI, Jun. 2012*, pp. 171–173.
- [7] H. Chen, W. Chung, J. Qin, E. Reid, M. Sageman, and G. Weimann, "Uncovering the Dark Web: A case study of Jihad on the Web," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 59, no. 8, pp. 1347–1359, 2008
- [8] C. C. Yang, X. Tang, and B. M. Thuraisingham, "An analysis of user influence ranking algorithms on Dark Web forums," in *Proc. ISI-KDD 2010, Art. ID 10*.
- [9] X. Tang and C. C. Yang, "Identifying influential users in an online healthcare social network," in *Proc. IEEE Int. Conf. ISI, May 2010*, pp. 43–48.
- [10] Mane Priyanka, Rathod Sonali, Sanap Deepali, Bhavana Shirude, "Ranking dominated web forum users," in *ijesr/jan 2016/ vol-6/issue-1/13-17*