# Ranking Web Forum Users Based On User Posts

**Kavitha.D[1], P.Sathiyapriya[2]**

*[1]PG student, Dept. of Computer Science Engineering, Sir Isaac Newton College Of Engineering technology, Pappakoil, Nagapattinam-611001, India*
*[2] Assistant Professor, Dept. of Computer Science Engineering , Sir Isaac Newton College Of Engineering technology, Pappakoil, Nagapattinam-611001, India*

---***---

**Abstract** - *Data mining is the extraction of implicit previously unknown and potentially useful information from data. The exploration and analysis by automatic and semi automatic means of large quantities of data inorder to discover meaningful patterns. The web is being used as a medium by extremist groups to practice several kinds of mischievous acts with concealed agendas and promote ideologies in a sophisticated manner. Some of the web forums are predominantly being used for open discussions on critical issues influenced by radical thoughts. The influential users dominate and influence the newly joined innocent users through their radical thoughts. Here an application of collocation theory to identify radically influential users in web forums is developed. The radicalness of a user is captured by a measure based on the degree of match of the commented posts with a threat list being created as a database. When a post from an user is found to have a match with the threat list, the user is considered as a radical user. These users are finally embedded and grouped with a customized Page Rank algorithm to generate a ranked list of radically influential users. Collocation theory is more effective to deal with such ranking problem than the textual and temporal similarity-based measures. Along with the measure of radicalness, the trustfulness of the message posted by an user and his activeness in the web forum is also analysed.*

***Key Words***: Social media, National security, Online information, User posts, user identification, user ranking.

## 1.INTRODUCTION

### 1.1 Introduction

Data is the most elementary descriptions of things, events, activities and transactions. The organized data that has meaning and value is called an information. The extraction of useful information from large clusters of data is known as Data mining. The analysis and retrieval of information from large quantities of data in order to discover meaningful patterns can also be Data mining. The task of discovering interesting patterns from large amounts of data is also data mining.

The Web is used to practice activites that are not in a good cause to the society by several groups. They use the web to promote their ideas, and hence identification of these groups, and terrorist organizations on the Web is needed. The discussions of these groups in online forums lead to the spread of negative radical thoughts into the minds of various users of the forum. Hence the identification of the influential user and ranking them from the group is essential.

The one type of web known as the dark web is specifically a web forum with the objective to support terrorists groups. The Grey Web forums on topics that disturb the society or threaten the public safety. Some of the issues included in the Grey web forum are topics like gambling, spiritualism, pirated CDs and many more of the same kind.

### 1.2 Influential Users

Social media sites allows a huge amount of User generated content to be available where a portion of such information is in the form of noise. Some users develop an interest in some posts which are posted by the same user, henceforth developing a relationship with those users. They influence the users who trust them with their comments and thoughts. They are called as influential users. They use this relationship to induce other innocent users with their objectives.

### 1.3 Role Of Radical Users

Radicalization is defined as inducing people by thoughts beyond the normal limit on various aspects that create endanger to the society. The posts by these people are in the motive to promote radical thoughts to the innocent user. These people have no personal values for ethics and the growth of the society.

Their ideas may be against a race, or a political party, or a religion, or a nation, or any organization with a mass of followers. These radical users induce their thoughts on the other innocent members and form a group along with them and cause danger to the society. The web forums are used as a medium to serve this purpose by these radical users where they can interact with vast amount of users. Various radical thoughts are related to attacks and destroy globally. The destruction caused them affect innocent people lives.

### 1.4 Overview Of The Project

Ranking of the users to identify their radicalness in the web forums based on their posts is tried. It is a great nesesscity to identify such users as they cause a great danger to the society and remain as a challenge to the national security.The user posts are compared against a threat list maintained in the database and if any match found they are considered as radical and the user with more radical posts is considered as more influential and ranked at the top.Two kinds of algorithm is used for the identification. Page rank algorithm and mean reciprocal algorithm. This implementation is done on a stand alone application which is accessed only by the employees of the web forum.

## 2. LITERATURE SURVEY

Allodi.L [1] emphasis that to perform Cybercrime activities the people related to it have also designed a structure that is similar to the legal online web forums. The cybercrime markets run by criminals who keep themselves anonymous. An online market should be filed and has its own regulations. Users should be given a status when they register and as they proceed their status should change from good or worse or banned. This type of regulation is mainly used for credit card forgery. For online shopping the messages between the seller and buyer should be of the private type. A successful forum should follow this. For a successful forum it should have an admin who is the top head. To publish news and events it should have a redaction, a moderator to manage and run the forum. A trustee who are the moderators and admin of other forums. Specialists are advanced users of the forum and users are the normal users of the forum. The users who are found to be guilty are termed as ripper. Users who are found to be continuously guilty are categorized as banned users. Regulation mechanisms of underground markets follow the same regulations as that of a successful well off online forums.

Jialun Qin, et.al, [2] discuss that Extremists organizations are using the web to widely spread their ideas. Here a research is done to understand their technical knowledge and plans. It starts with the automatic crawling of data, then a systematic analysis tool DWAS (Dark Web Attribute System) is used to compare the plans of different extremists organizations. There are many extremists groups from the Global extremists organizations, as of the US domestic racists, Militia groups to Latin America, Islamic groups etc… They perform fund raising, arms distribution, training and functional operations through the web. DWAS automatically extracts the appearance of specific attributes and assign three scores to indicate their technical knowledge, content and web usage. The weightage of the scores are given by web experts. Based on the score weightage the efficiency and advancement of the groups on the web is determined.

Jingtian Jian, et.al, [3] develops FOCUS expanded as the Forum Crawler under supervision is a highly efficient automatic supervised web forum crawler. FOCUS is used to crawl web pages to get the content from them. The forum thread have all the information hence the target of FOCUS is Forum threads. All the forums have a different layout and style but the internal working is the same in all. Due to this the navigation path to the URL will also be the same, henceforth the reduction of the forum crawling problem to that of a URL type recognition problem. URL is uniform resource locator, an protocol identifier givrn by http. It is a reference to resources on the internet. A web forum mainly has four types of pages given as entry page, index page, thread page and other page. There are four URL types givrn as index URL, thread URL, page flipping URL and other URL. FOCUS applies the technique of EIT path and ITF regex where EIT path is the navigation path of the crawler and ITF regex is the regular expression to recognize ITF on the EIT path.

Kanagavalli, et.al,[4] finds that users lookup for reviews and opinions on the forum to take upon a decision. The task is to analyze the behavior of an user by his posts. The posts can be categorized into three as positive, negative or neutral posts. These posts decide the behavior of individuals . The posts are classifies and then clustering of them is performed. Encoding the posts to a numeric value of a positive or negative is performed to analyze the user behavior and sentiment analysis. First step in analysis is Forum crawling for the collection of information from the posts is done automatically by using specified tool. Second step followed by the Pre Processing of the crawled data to have an identity and title for further classification . Data cleaning is done to remove noise and irrelevant data. Third step is to create a graph where a node is created with forum topics and a link is created with them. They are then stored in the database for later retrieval. Final step is to classify the data which is clustered in four dimensions as posts, sentiments, positive and negative aspects and user perception.

Larry [5] proposes an algorithm to rank web pages based on their importance. Page rank is taken as a numeric value that points out the importance of a page. It uses the concept of hyperlinks, which is a link from a page to another. It can either be an inlink a link in the same website or an outlink a link to another website. The page that links to another page, is casting a vote for that other page. A page with more votes is considered to be the important page with manu number of links. To formulate page rank the web is treated as a directed graph.

Michael Trusov, et.al, [6] proves the success of social networking site rely on the activity levels of the users of that site. A method is developed to the influential user based on their activity level. The calculation is done based on the user site login, since the users activities are recorded and can be tracked. To identify the influential users their login details , activities of all the users are taken into account and a poisson regression is created using a poisson distribution. The poisson parameters are taken as self effects and friend. Self effects are user specific activities and friend are the posts of others to the user. The problem here is the activity of the friends are often missed out if a user has a large group of friends.

Nitin Agarwal, et.al, [7] proposes that using a blog as a popular means a user can deliver his content or information on the web globally.The users using a blog are called as bloggers. Bloggers write posts which contain their ideas, views, thoughts, likes and dislikes. Here the challenges faced in the identification of an influential user is identified. A blog acts as a social networking medium which connects many together. Blogs can be categorized as individual and community blogs. Individual blogs are that of a single persons ideas and the content is provided by that person and can be viewed by others. Community blogs are the one where all the users can post their ideas, opinions, thoughts and suggestions. To identify an influential user, the integration of all the information is required. A model is developed to face the issues in the identification process. Data collection is one of the major issue faced by the model.

## 3. EXISTING SYSTEM

The system performed ranking from a data taken from a workshop. The different users from existing forums were collected and the ranking was performed. The discussions among extremist groups lead to talks with cold culture followed with abusive languages, and hence the evolution of online hate and violence. Web forums are considered as the fastest and quick easy way to have discussions. The evolution of hate and violence did not provide a way for the web forum discussions to happen in a well to go manner.

### 3.1 Disadvantages Of Existing System

The Previous system is not effective to retrieve the user expectation and ranked page from existing web forum.The Disadvantage of the system is the ranking of influential user in a web forum is not performed effectively.

## 4. PROPOSED SYSTEM

It comprises the specification and procedures for crawling of data and the steps necessary to extract them for the identification and ranking of the users in a web forum. The most important point is that the direct source information to the user should be trustworthy and illegal thought should not be passed on to them. Efficient and intelligent filtration of such data is performed so that no radical thoughts get posted in the forum. Along with the identification of the radical users , the trustness of the information is obtained and the web are ranked.

The proposal is to analyse and find the influential user. The user can spread negative thoughts which could create disturbances in the society. The truth of information of each user is also planned to be identified and to rank the web pages in an effective manner. The identification and ranking of influential radical users is performed. Along with this the identification and ranking of the trustworthy information from the users in also done. To perform this a database is created and the posts of each user is verified. The identification and elimination of dead users in the web forum is also proposed.

### 4.1 Advantages Of Proposed System

The Advantage of this system is the ranking of influential user in a web page is done effectively. How much a message posted by an user can be trusted is identified. The users are identified as active/dead users based on the time stamp.

## 5. WEB FORUM INTRODUCTION

The efficient evaluation of the system modules starts with a clear understanding about the structure and working of the web forum. The Web Forum is an online internet discussion site used by various people to speak out their thoughts. Web Forums allow people to interact with other people through their posted messages. A posted message needs approval by a moderator before it becomes visible on the forum. A forum follows the DOM, Document Object Model directory structure. A thread is a collection of posts. A post is a message submitted by the user. Users can format their own posts, they can edit or delete their posts. A web forum keeps track of the user posts. Users with more posts and those with less posts can be recorded and tracked.

### 5.1 User Groups

The User group is categorized as three groups. The people who register and login and post their comments in the forum are termed as the Members of the forum. Another group called as the visitors do not register in the forum but are allowed to access some pages of the forum. The moderators are also referred to as mod are users or employees of the forum. The people involved in the backend of the forum are called as moderators and administrators. The moderators can allow or delete posts and hence keep the forum clean from posts that induce other people. All the information about each user is known by them. The moderators are the people who have all the access to every posts in the forum. The administrators play an important role and are responsible for the efficient working of the forum. The administrators take the important decisions for the functioning of the forum. All the technical decisions and the removal or adding up of pages in the forum can be done only by the administrators.

## 6. IMPLEMENTATION

### 6.1 Module Description

The different modules used in the application be stated as the following : Forum Crawling and parsing, Data pre-processing, User Radicalness Identification,User Trustfulness Identification, User Collocation identification, User Ranking.

### 6.1.1 Forum crawling and parsing

Crawling is done automatically where the data is taken from the thread. Automatic traversal of the data is known as crawling. After the retrieval of data the noise and special characters are removed. Huge collections of data are available in the web. When a user writes a post, an automatic crawler tool collects the data by downloading and storing them. The collected data is broken down into smaller chunks by applying a set of rules so that it can be easily interpreted managed or transmitted.

### 6.1.2 Data Preprocessing

The data that results after crawling works in coordination with the parser to extract all the meaningful information from it. The data is organized as a collection of threads where each thread had an identity and a title. Each thread has a collection user posts along with their timestamp. Even after the data is being organized, it is again processed for data cleaning to remove noise and get meaningful content.

### 6.1.3 User Radicalness Identification

The relationship between the users interacting in the forum is to be identified. When a user posts a message, it is checked by the moderator before it becomes visible in the forum. The Moderator verifies the post against the radical database formed and used by the web forum and if any match found with radical influential thoughts, the post is blocked by him and the user is ranked as a radical user. Whenever identify the user rank, we have to match with the list of existing horsing words and if we found any match against user post, then we have to ignore that post thread and remove owner of those post from the rank.

### 6.1.4 User Trustfulness Identification

Similar to the radical user identification, the post of an user can be verified and ranked against the truthness of the message. A Database is created in the Webforum with the data that gives a good identity to the member of the webforum. The user posts are matched against this dataset and the users are identified hence the ranking.

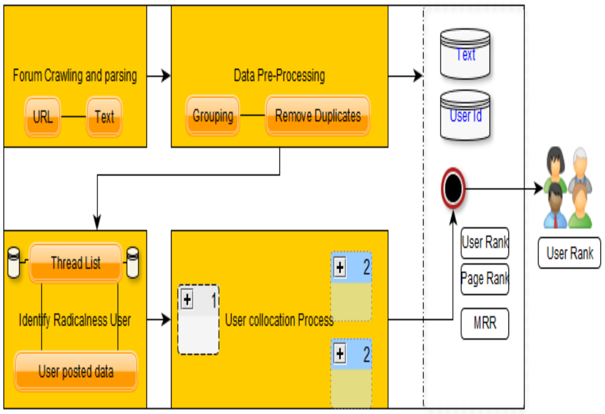
### 6.1.5 User Collocation Identification

The users of the web forum build an intimate relationship when they interact in the same threads. Collocation is defined as the interactions of various users when communicating in the same threads. Collocation theory is applied to study the influence of an user on the other by the means of their interactions. The Contingency coefficient measure is the most efficient collocation measure to identify the influence between users. The user posts are grouped based on this collocation theory and hence the radical users are identified.

### 6.1.6 User Ranking

Once collocation identified, the user with more radical posts or the excellent words posts are identified or ranked compared with the users against WEB Forum. Two different algorithms are used to find the user rank. They are Page rank algorithm and MRR (Mean reciprocal rank) algorithm.

6.2 ARCHITECTURE



## 6.3 ALGORITHMS AND TECHNIQUES

The user ranking is done by using two different kinds of Algorithms and Techniques known as Page rank algorithm and MRR (Mean reciprocal rank) algorithm.

### 6.3.1 Page Rank Algorithm

The algorithm determines the radicalness of a user and the one with more radical posts is given the highest rank. The algorithm can be formulated as,

$$PR(A) = (1-d)+d(PR(T_1)/C(T_1)+...+PR(T_n)/C(T_n))$$

Where,

PR(A) is the PageRank of page A,

$PR(T_n)$ is the PageRank of page $T_n$,

$C(T_n)$ is the number of outbound links on page $T_n$ and

d is a damping factor.

The page rank is calculated based on the inbound links and the outbound links for each page of an web forum. When a page is having many inbound links from other pages then it is considered as the important page and hence ranked as the most important page.

### 6.3.2 Mean Reciprocal Rank Algorithm

The Mean Reciprocal Rank Algorithm or the MRR algorithm performs evaluation on various processes and produces the possible responses for an query. The responses are evaluated based on probability measures for correctness. The response obtained through the algorithm is used to rank users based on their posts. The posts are evaluated by probability means and hence the ranking

## 6.4 APPLICATIONS

The Final output will be who is an influential user in the web forum along with the detection on the trustfulness of the messages posted by the users. It requires the development of two applications, which is given below. They are, Web application and Standalone application.

### 6.4.1 Web Application

Web application will be the kind of forum discussion application like Java-Ranch, java-forums, stack-overflow. This application will run in tomcat server and it should have login and registration pages. When ever user wants to post the question or post the answer, the user should login into the application.

### 6.4.2 Stand Alone Application

Stand Alone application is used for to identify the influential user based on the data set from the web forum.Whenever identify the user rank, we have to match with the list of existing horsing words and if we found any match against user post, then we have to ignore that post thread and remove owner of those post from the rank. The existing data set as below. Finally apply the MRR/Page rank algorithm, to find the output. The final output will be in the integer points. Based on those points, we have to declare the list of users name in the screen on rank wise.

Active/dead user: As per our process, first of all we have to identify the Active/dead user based on their post and timestamp of the user login. If user 'long time no login' means, we have to group them and treated as in-active/dead user. After that, we have to apply the collocation theory in the user post/comments.

## 7 CONCLUSION

A proposed model to identify the radical posts and then performing user ranks based on their posts for a web forum was dealt. The measure of radicalness was identified with the help of a created database and the most effective contingency coefficient measure. The probability of a word in the post to be one in the created threat list database is evaluated with the help of the Mean Reciprocal Rank. The user with high radical posts is ranked as the top and then followed by the others in an web forum. Collocation theory is a measure which groups all the threads on an user as a whole.

The system to identify for the truthfulness of the post posted by an user is in process. A plan for enabling the model with a larger database with more influential words in usage. Getting a real time database from online websites is a difficult job. A future process of getting a real time database from a real functioning blog for viewing their posts and identifying and ranking them has to be done. The ranking of the web pages is also to be done.

## ACKNOWLEDGEMENT

I would like to thank my project guide Ms. P.Sathiyapriya, Assistant Professor, CSE Dept., she served as a sounding board for both contents and programming work. Her valuable and skillful guidance, assessment and suggestions from time to time improved the quality of work in all respects.

## REFERENCES

1. Allodi (2015), "Then and Now: On the Maturity of Cybercrime Markets: The lesson that black-hat marketers learned" DOI 10.1109/TETC.2397395.

2. Jialun Qin (2010), "A multi-region empirical study on the internet presence of global extremist organizations" Volume #SpringerMedia, LLC 2010.

3. Jingtian Jian (2013), "FOCUS: Learning to Crawl Web Forums" Volume 25 Number 6.

4. Kanagavalli (2014), "Analysing User Posts for Web Forum using K-maens Clustering" Volume 4 Issue 5.

5. Larry (2015), "A Review Paper on Page Rank Algorithm" Volume 4 Issue 6.

6.Michael Trusov (2010), "Determining Influential Users in Internet Social Networks" Volume XLV II.

7. Nitin (2008), "Identifying the Influential Bloggers in a Community" ACM 978-1-59593-927-9080002.

9. Steve Kramer (www.pargonscience.com) , "Anamoly detection in extremists web forum using dynamical systems approach" Austin.

10. Tarique Anwar (2013), "Modeling a Wen Forum Ecosystem into an Enriched Social Graph" LNCS Volume. 8329, Springer, PP.152-152.