

CONTINGENCY COEFFICIENT MEASURES FOR WEB FORUM USERS AND SENTIMENT ANALYSIS

Aswin Kumar.P,
Student,

Department of Computer Science and Engineering,
Velammal Institute of Technology, Chennai, India.

Murali Krishna Kumar.U,
Student,

Department of Computer science and engineering,
Velammal Institute of Technology, Chennai, India.

Bhagavan Iyyappan.J,
Student,

Department of Computer science and engineering,
Velammal Institute of Technology, Chennai, India.

Chitra.R,

Assistant Professor,
Department of Computer science and engineering,
Velammal Institute of Technology, Chennai, India.

Abstract: The online social network plays a vital role in the day to day life for much purpose. It has been found that the web is used as a tool by many hate groups or extremist groups to promote various kinds of ideologies and practice several kinds of acts. Some of the web forums are being used for open discussions on critical issues influenced by various radical thoughts. In the web forums the influential users have control and dominate the newly joined innocent users through their radical thoughts. In this paper we have proposed an approach to identify the ranked list of influential users and designed an algorithm based on page rank to rank the radically influential users. The Contingency Coefficient measure is found as the most promising measure, when embedded in the customized Page Rank algorithm along with the radicands measure. The collocation measure deals with such ranking problem more effectively than textual and temporal similarity based measures. We have formulated a new measure Sentiment Analysis. It is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker to eliminate the radical activities.

Keywords: Social Media, Web Forum Discussions, Influential users, Ranking, Sentiment Analysis.

I. INTRODUCTION

Now days, Web is being used as a tool to practice several kinds of mischievous acts with concealed agendas and promote ideologies in a sophisticated manner. Infiltration of extremist groups, hate groups, racial supremacy groups, and terrorist organizations on the Web with hundreds of multimedia websites, online chat rooms and Web forums is posing grievous threats to our societies as well as the national security. The multimedia websites provide support for their psychological warfare, fund-raising, recruitment, and propagation of their agendas whereas chat rooms and Web forums promote their strategies and ideologies through discussions with naive users. Often the public discussions among differently minded extremist groups lead to irascible talks accompanied with abusive languages, and promote online hate and violence.

Web forums are recognized for their exhaustive, vivid and non-spontaneous nature of discussions that are achieved for later reference. Previous studies have found Web forums as the most active medium being used for this purpose. Research on identifying radical and infectious threads, members, postings, ideas and ideologies in Web forums for tracking the

grievous threats posed by the active extremist and hate groups has gained considerable attention of the research community. The portion of the Web circumscribing the sinister objectives of extremists group is said as Dark Web, and specifically the Web forums with substantial prevalence of activities supporting extremism are said as Dark Web forums. Another class called Gray Web forums refer to the forums in which the discussions focus on topics that might potentially encourage biased, offensive, or disruptive behaviors and may disturb the society or threaten public safety. They include topics like pirated CDs, gambling, spiritualism, bullying, and online-pedophilia.

A. Influential User Identification : Influential user identification have been done in a business intelligence orientation for marketing products through targeted influential users .Some other objectives are information dissemination community leader identification and expertise discovery. An empirical measure is done of influence based on the number of in-network votes that the post of a user receives to rank using content similarity and response immediacy. It is shown as out-performing Page Rank, in-degree and out-degree rankings

helps in the identification of the user and the application of User Rank algorithm in the domain of Dark Web forums.

B. Our Contribution: Proposed system of this paper is to identify the influential web forum user based on customized page ranking logic. We have to implement the customized page ranking algorithm to categories the web forum users. The proposed method starts with crawling and pre-processing the forum data, followed by user radicalness identification, user collocation identification, and also the trust worthy information of users and finally ranking the users based on a customized Page Rank algorithm and Sentiment analysis which helps to determine whether the user's opinion meaning is positive, negative or neutral so that the Radical user's comments can be viewed by the members.

II. RANKING MODULES

A. Forum Crawling and Parsing : The process starts with a data crawling and preprocessing step in which the URL of the forum home page is passed to the forum crawler, which crawls all relevant web pages and eliminates the duplicates heuristically. A platform- specific parser module is employed to extract the meaningful snippets from the crawled WebPages, which are then passed to the data preprocessing module.

B. Data Pre-Processing : The metadata extraction task works in close coordination with the parser module to extract all relevant metadata. The obtained data is organized as a collection of threads having a unique id and title; each thread containing one or more posts having a post id, time-stamp, body text, author, and quotations. The body text is additionally processed through some cleaning and chunking mechanisms to remove the noise and crystallize into individual meaningful pieces of information

C. User Radical Identification: The foundation of their automatic radical identification process is laid on a set of manually crafted list of threat words that are typically found in radical texts. The forum is delivered by many people as representing radical ideology. We noticed that the threat list is quite long, and most of the words in the list are also used in general situations. For example, honor, hard, puppet, and movement are general terms and these are very likely to mark a non-radical message as a radical. Because the list is manually crafted, there needs to be strong rationality to use the words for characterizing radicalness. We reduced the list to a set of highly focused words based on our observation and perception. Also the terms could be acronyms or synonyms or in different languages. To handle these real scenarios, the list

needs to be updated regularly with time. Shorter lists may give some radical members a chance to evade, whereas longer lists may mark even innocents as radicals. Therefore one needs to be extreme careful while preparing or updating the threat list.

D. User Collocation Identification: 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.

E. Ranking: Once collocation identified, our system is ready to discover or rank the user against WEB Forum and then customized Page rank algorithm to identify the posts.

F. Sentiment Analysis: Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. This analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation affective state (that is to say, the emotional state of the author when writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader).

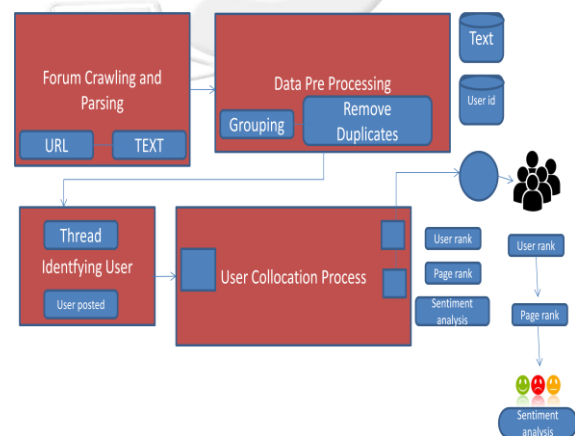


Figure 1. Work-flow of the Module

III. IMPLEMENTATION

Final Decision will be who is an influential user in the web forum. Actually we have to develop two applications. They are, Web application [1], Standalone application [2].

A. 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. Whenever user wants to post the question or post the answer, the user should login into the application. 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

B. Stand Alone Application: Stand Alone application is used for to identify the influential user based on the data set from the above application. 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 Page rank algorithm and the Sentiment analysis, to find the output. The final output will be in the list displaying the analysis of the words.

C. Algorithms and Techniques: Page rank algorithm and Sentiment analysis is a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries, ordered by probability of correctness. The Sentiment analysis helps to determine the opinion meaning and aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.

IV. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an approach to identify a ranked list of radically influential users in Web forums. We have formulated a radicalness measure and a variety of collocation-based association measures, and designed an algorithm based on Page Rank to rank the radically influential users and Sentiment Analysis to determine the opinion of the users and differentiate between users taking positive and negative. This work opens several promising directions for future research. Considering social relations in addition to the threaded interactions, exploring semantic factors like discussion context and topic drift for radicalness identification, and applying Analysis of the users comments word by word further to deeply eradicate the radicalness of the users.

V. REFERENCES

- [1] Tarique Anwar and Muhammad Abulaish, "Ranking Radically Influential Web Forum Users", IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 10, NO. 6, JUNE 2015
- [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] T. Anwar and M. Abulaish, "Modeling a Web forum ecosystem into an enriched social graph," in Ubiquitous Social Media Analysis. Berlin, Germany: Springer-Verlag, 2013, pp. 152–172.
- [4] T. Anwar and M. Abulaish, "Identifying cliques in Dark Web forums— An agglomerative clustering approach," in Proc. IEEE ISI, Jun. 2012, pp. 171–173.
- [5] 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.
- [6] J.-H. Wang, T. Fu, H.-M. Lin, and H. Chen, "A framework for exploring Gray Web forums: Analysis of forum-based communities in Taiwan," in Proc. IEEE Int. Conf. ISI, May 2006, pp. 498–503.
- [7] J. Qin, Y. Zhou, E. Reid, G. Lai, and H. Chen, "Analyzing terror campaigns on the Internet: Technical sophistication, content richness, and Web interactivity," Int. J. Human-Comput. Stud., vol. 65, no. 1, pp. 71–84, 2007.
- [8] J. Glaser, J. Dixit, and D. P. Green, "Studying hate crime with the Internet: What makes racists advocate racial violence?" J. Soc. Issues, vol. 58, no. 1, pp. 177–193, 2002.
- [9] 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.
- [10] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in Proc. Int. Conf. WSDM, 2008, pp. 207–218.