International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Identify Ranking Position with Web Forum Users

Kavitha .D[1], P. Sathyapriya[2]

[1]PG student Sir Isaac Newton College of Engineering Technology, Pappakovil, Nagapattinam-611001, India

[2]Assistant Professor Sir Isaac Newton College of Engineering Technology, Pappakovil, Nagapattinam-611001, India

Abstract: ***In the recent past, it has been found that the web is also being used as a tool by radical or extremist groups and users 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. This paper presents an application of collocation theory to identify radically influential users in web forums. The radicalness of a user is captured by a measure based on the degree of match of the commented posts with a threat list. Eleven different collocation metrics are formulated to identify the association among users, and they are finally embedded in 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 studied earlier.***

Keywords : Social Media, Web Forum Discussions, Influential users, Ranking

1. Introduction

Now a 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 recognised for their exhaustive, vivid and non-spontaneous nature of discussions that are archieved 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 behaviours and may disturb the society or threaten public safety. They include topics like pirated CDs, gambling, spiritualism, bullying, and online-pedophilia.

2. Role of Influential Users

Due to enormous and rapid growth of user-generated content on social media sites, and users generally avoid going through every comment posted by others. There always exist some users who develop some relationship of trust with other members by their activeness and quality of comments, and their comments always receive significant attention of a large community. These are the *influential users,* whose activities and comments greatly affect the society. Influential users find it very easy to convince the other users with their ideologies. Recent studies have found it to be an important issue and a challenging task to identify such influential users.

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. sing content similarity and response immediacy. It is shown as out-performing PageRank, in-degree and out-degree rankings helps in the identification of the user,and the application of UserRank 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.

3. 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 crystalize 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 associvity of different users, and estimate their influence while propagating an ideology through their interactions.

E. User Ranking

Once collocation identified, our system is ready to discover or rank the user against WEB Forum.We are using two different algorithms to find the user rank. They are Page rank and MRR (Mean reciprocal rank)

4. Architecture

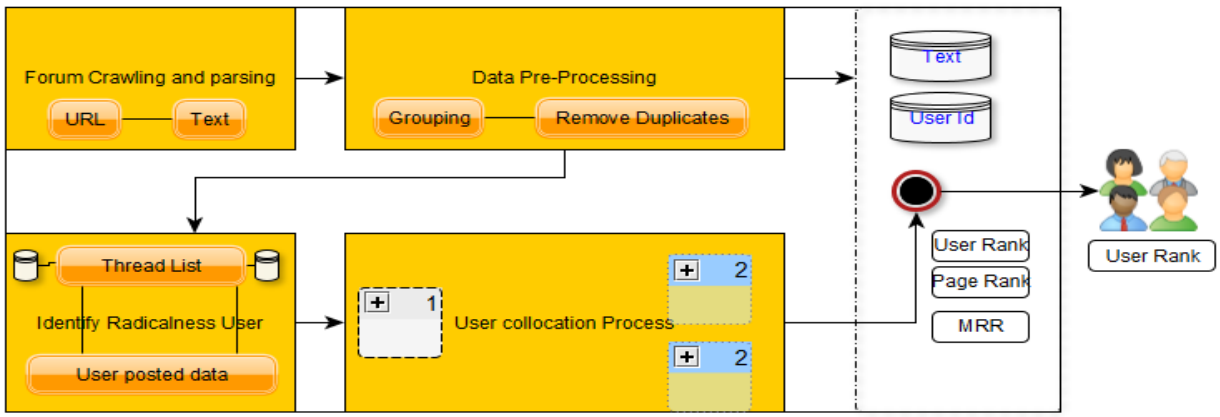

Figure: Worflow of the Modules

5. 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].

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. 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.

2. 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 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.

3. Algorithms and Techniques

Page rank algorithm and MRR (Mean reciprocal rank) 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 reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of

the reciprocal ranks of results for a sample of queries. The reciprocal value of the mean reciprocal rank corresponds to the harmonic mean of the ranks.

6. Conclusion and Future Work

More Emphasis is given on the understanding of the Forum crawling of the Pre processing of the data, our future work is in the identification of the user radicalness and user collocation by the information from them and rank them based as radical users and trust worthy users. This ranking is based on a database which created to have the radical words in it when compared with that posted by the user.

References

[1] 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.
[2] T. Anwar and M. Abulaish, "Modeling a Web forum *ecosystem* intoan enriched social graph," in *Ubiquitous Social Media Analysis*. Berlin,Germany: Springer-Verlag, 2013, pp. 152–172.
[3] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.*
[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst., vol. 30, nos. 1–7,* pp. 107–117, Apr. 1998.
[5] E. M. Voorhees, "The TREC-8 question answering track report," in *Proc. TREC, 1999, pp. 77–82.*
[6] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top *k lists," in Proc. 4th Annu. ACM-SIAM SODA, 2003, pp. 28–36.*

Author Profile

Mrs. Kavitha. D has completed her B.E from S.R.M Engineering College and currently doing M.E in Computer Science & Engineering from Sir Issac Newton College of Engineering and Technology, Nagapattinam.

Prof. P. Sathyapriya has completed her B.E and M.E. in Computer Science & Engineering, and currently working with Sir Issac Newton College of Engineering and Technology, Nagapattinam.