IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 11, November 2016
ISSN (Online) 2348 – 7968 | Impact Factor (2015) - 4.332
www.ijiset.com

A Survey on Ranking Radically Influential Web Forum Users on Social Media

Samiksha R. Tadas[1], Prof. K. B. Bijwe [2]

1. P. G. Student, P. R . Pote (Patil) Welfare & Education Trust's college of Engineering & Management, Amravati Department of Computer Science & Engineering, Email- Samikshatadas@gmail.com
2. Assistant Professor, P. R . Pote (Patil) Welfare & Education Trust's college of Engineering & Management,Amravati Department of Computer Science & Engineering, Email- komalbijwe@gmail.com

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. Some forum are used for a open discussions on a critical topics influenced by radical thoughts. The influential users dominate the mind of naive users using their radical thoughts. Influential users complete the naive users to do wrong things. The main intension of this topic is to understand blocking of malicious links on social media post using content mining and text comparison with existing item sets which will predicate the category of the post and help to stop violent data on social media.

KEYWORDS— Social media analysis, security informatics, radical user identification, user collocation analysis.

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[1][2]. 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 forms 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 INFLUENIAL 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[1].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[2].

B. Our Contribution: The influential web forum user based on customized page ranking logic.We has 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. LITERETURE REVIEW

Matthew Richardson and Pedro Domingos in year 2012 [3]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.

J. Qin, Y. Zhou, H. Chen in year 2011 [4]works on dark web collection building process. 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.

M. Chau and H. Chen in 2011 [5]describes as the Web continues to grow, it has become increasingly difficult to search for relevant information using traditional search engines. Topic-specific search engines provide an alternative way to support efficient information retrieval on the Web by providing more precise and customized searching in various domains. However, developers of topic-specific search engines need to address two issues: how to locate relevant documents (URLs) on the Web and how to filter out irrelevant documents from a set of documents collected from the Web.

R.J. Mooney and L. Roy in 2012 [6]describe Recommender systems improve access to relevant products and information by making personalized suggestions. Most existing recommender systems use social filtering methods that base recommendations on other users' preferences. By contrast, content-based methods use information about an item itself to make suggestions. This approach has the advantage of being able to recommend previously unrated items to users with unique interests and to provide explanations for its recommendations. We describe a content-based book recommending system that utilizes information extraction and a machine-learning algorithm for text categorization. Initial experimental results demonstrate that this approach can produce accurate recommendations.

F.Sebastiani in 2013 [7]describes the automated categorization of texts into predefined categories has witnessed a booming interest in the last ten years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of pre-classified documents, the characteristics of the categories.

M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari in 2012 [8]proposes a system enforcing content-based message filtering for On-line Social Networks (OSNs). The system allows OSN users to have a direct control on the messages posted on their walls. This is achieved through a flexible rule-based system, that allows a user to customize the filtering criteria to be applied to their walls, and a Machine

Learning based soft classifier automatically labeling messages in support of content-based filtering.

Summary: in this dissertation various techniques related to the literature been discussed and below table represent various advantages and disadvantage comparison in between discussed technique.

Authors	Years	Technique	Disadvantages	Results
Matthew Richardson and Pedro Domingos	2013	Marketing and social network sites technique	Marketing plan when the structure of network is unknown or partially known.	Works on different way of marketing and social network sites by mining knowledge sharing.
J. Qin, Y. Zhou, H. Chen	2011	Dark web collection building process	Dark web attributes system which is used to analyze and compare extremist organizations.	Works on dark web collection building process. This system use a systematic content analysis tool called the dark web attribute system.
M. Chau and H. Chen	2011	Machine combines Web content analysis and Web structure analysis.	System can be applied only topic-specific search engine Development and Web applications.	Describes as the Web continues to grow, it has become increasingly difficult to search for relevant information using traditional search engines.
R.J. Mooney and L. Roy	2012	Content-based book recommending system	To provide explanations for its recommendations.	Describe Recommender systems improve access to relevant products and information by making personalized suggestions.
F.Sebastiani	2013	Machine learning techniques	Increased availability of documents in digital form and the ensuing need to organize	Describes the automated categorization of texts into predefined categories
M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari	2012	On-line Social Networks (OSNs)	OSN users direct control on the messages posted on their walls.	Proposes a system enforcing content-based message filtering for On-line Social Networks (OSNs)

4. RANKING

A Ranking is simply defined as an association, between a set of items on repository or web pages such that, for any two items, the first is either ‘ranked higher than’, ‘ranked lower than’ or ‘ranked equal to’ the second. In mathematics, this is known as a weak order or total preorder of objects. It is not necessarily a total order of objects because two different objects or

items can have the equivalent ranking[10][12]. The rankings as we know are totally in ordered manner.

To bring down detailed measures to an arrangement of more than two things in a successive order, rankings make it possible to measure complex of heterogeneous information according to certain specific criterion. For example, the Internet search engine like; Google, may rank the pages it according to their relevance, making it conceivable for the end user rapidly to select the pages they are likely wish to retrieve. The ranking method is more effective for end-users is that all the similar terms or keyword in the query are used for retrieval, which result is ranked based on co-occurrence of query terms, as modified by statistical term-weighting. The ranking approach also works well for phrases or for the complex queries that may be difficult for end-users. For example, "human factors, human emotions, performance of the system in medical, military, university, organization data-space" is difficult for end-users to express in Boolean logic.

Ranking algorithms form an essential part of any search engine and a large amount of research has been done on them because they determine the quality of a search engine from the user's perspective. A manual ranking scheme could have been sufficient, if the Web had been composed of a few hundred pages, as was the case during the initial years of the Web with search engines. However, with the dynamic explosion of information over web it could no longer be practical to rank millions of pages and automated means had to be developed in the form of ranking algorithms. The paper gives an overview of the various ranking algorithms that can be used to enhance the search experience of the technical users over the blogs .

5. METHODOLOGY

A) Data collection: Data collection is the initial step of web usage mining, the data authenticity and the integrality will directly affect to the following works smoothly carrying on and the final recommendation of characteristics service's quality. Therefore it must utilize scientific, reasonable and advanced technology to gather many data. At present, towards web usage mining technologies, the main data origin has three kinds: server data, client data and the middle data.

B) Data preprocessing: Some database is In - sufficient, inconsistent and including noise. The data pretreatment is to carry on a unifications transformation to those databases. The result is that the database will to become integrated and consistent.

C) Knowledge Discovery: Use statistical method to carry on the analysis and mine the pretreated data. We might discover the user or the user community's interests then construct interest model. At present the generally used machine learning methods mainly have clustering, classifying, the relation discovery and order model discovery[10]. Each method has its own excellences and shortcomings, but the quite effective method mainly is classifying and clustering at the presents.

D) Pattern analysis: Challenges of Pattern Analysis are to filter uninteresting information and to visualized and interpret the interesting pattern to the users. Initial delete the less significance rules or models from the interested model storehouse; Next utilize technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered information or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

E) Focused Crawling: A focused web crawler takes a set of well-selected web pages exemplifying the user interest. The focused crawler start from the given page and recursively explores the linked web pages. While the crawlers performs a breadth-first explore of the complete web, a focused crawler explores only a small portion of the web using a best-first search guided by the user interests. crawling for retrieving multimedia content in the web, instead of plain HTML documents.

F) 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 web pages, which are then passed to the data pre-processing module.

G) 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(pi | pj) =1\out- Degree(pj)

This equation is the transition probability from webpage pj to webpage pi. The transition process is continued until a convergence is achieved and the scores at that instance are accepted as their final PageRank scores[1].

H) 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.

I) 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)

6. PROPOSED WORK

The popularity of online social media is spread day by day for various online community purposes. A social networking service is an online platform that is used by people to build social networking or social relation. This things also done in web forum, chat room etc. some forum are used for open discussions on a critical topics influenced by radical thoughts. The influential users dominate the mind of naive users using their radical thoughts. Influential users compel the naïve users to do wrong things. We are proposed a system which will check for the user post links post by them and their content in post as well as link and determining the abusing of that which help to make social media to make more secured from spreading such type of volatile thoughts on web.

7. Conclusion

In Ranking Radically Influential Web Forum Users it will study the advantages and feature of the system. Analyzing the affect of radical influence on the forum community is also a promising research direction to study the radicalness propagation in different extremist. This system is useful for removing the radical influential users from the web forum.

REERNCES

[1] Tarique Anwar and Muhammad Abulaish, "Ranking Radically Influential Web Forum Users", IEEE Transaction on information forensics and security, Vol. 10, No. 6, June 2015

[2] Priyanka B. Mane, Sonali S. Rathid, Deepali D. Sanap & Bhavana S.Shirude, "Ranking Influential Users", Imperial Journal of Interdisciplinary Research (IJIR) Vol-2, Issue-5, 2016

[3] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing", In Proc. 8th ACM SIGKDD Int. Conf. KDD, pp. 61–70 ,2012.

[4] 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.

[5] M. Chau and H. Chen, "A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis,", In Proc. Int. Conf. Adv. Soc. Newt. Anal. Mining (ASONAM), pp. 281–285, Aug. 2011.

[6] R.J. Mooney and L. Roy, "Content-Based Book Recommending Using Learning for Text Categorization,", J. Marketing Res, vol. 47, no. 4, pp. 643–658, Aug. 2012

[7]] F. Sebastiani, "Machine Learning in Automated Text Categorization,",In Proc. IIITD PhD Comprehensive Report, Vol.no 34, No. 1, November 2013

[8] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, "Content-Based Filtering in On-Line Social Networks", In Proc. IEEE ISI , pp. 171–173,June 2012.

[9] 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

[10] X. Tang and C. C. Yang, "Identifing influential users in an online healthcare social network", in Proc. IEEE Int. Conf. ISI, pp no43–48 ,May 2010.

[11] 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.

[12] T. Anwar and M. Abulaish, "Modeling a Web forum ecosystem into an enriched social graph,", Berlin, Germany: Springer-Verlag, pp. 152–172, 2013.

[13] T. Anwar and M. Abulaish, "Identifying cliques in Dark Web forums An agglomerative clustering approach", in Proc. IEEE ISI, pp. 171–173,June 2012.

[14] 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.

[15] 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.

[16] J. Qin, Y. Zhou, E. Reid, G. Lai, and H. Chen, "Analyzing terror campaigns on the Internet social media: Technical sophistication, content richness, and Web interactivity", Int. J. Human-Comput. Stud., vol. 65, no. 1, pp. 71–84, 2007.

[17] K.Babu , P.Charles," A System to Filter Unwanted Words Using Blacklists In Social Networks", International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 1748-1753,2014.