www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 12 December 2017, Page No. 22250-22254
Index Copernicus value (2015): 58.10, 76.25 (2016) DOI: 10.18535/ijecs/v6i12.02

Ranking Radically Association Among Users On Web Forum

Ms. Ghule Kiran Bharat [1], Prof. R. L. Paikrao[2] (HOD)

[1,2]Computer Department
Amrutvahini College of Engineering,
Sangamner, A. Nagar

Abstract
In the recent past, it has been found that the web is used as a tool by radical or extremist groups and users to perform several kinds of mischievous acts with concealed agendas and promote their ideologies in a sophisticated manner. Some of the web forums are specially being used for open discussions on critical issues influenced by radical thoughts.
We propose 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. The experiments are conducted on a standard data set to find radical and infectious threads, members, postings, ideas, and ideologies. Proposed system to rank the user on text and image based similarity measures.
We make the following key contributions in proposed system: An application of analyze the data it may be text data or image data. If it is text data it will go through preprocessing stages like stop word removal, suffix removal, then by cosine similarity function it check the similarity with threat list then decide whether that user is radical or not. If it image data, if it contain text data then it separate text from image by OCR technique. Send that text to text analysis and image goes through image preprocessing like image filtering, EHD to take aggregate features, using similarity measures it check similarity with training data set. Finally after measures of radicalness of user, it ranks the users by PageRank algorithm.

Keywords—Terms— Security informatics, Extremist group, Radical user identification, Users collocation analysis, Social media analysis

1. Introduction

In the recent past, it has been found that the web is used by extremist groups, hate groups, racial supremacy groups, and terrorist organizations on the web with numbers of multimedia websites, online chat room and web forums in posing grievous threats to our societies as well as the national security. The multimedia websites promote psychological warfare, whereas chat room and web forum promote their strategies and ideologies through discussions with naïve users. Public discussions among differently minded extremist groups lead to irascible talks accompanied with abusive languages, and promote online hate and violence.

Now in society web forum is used as the most active medium being used for this purpose [2]. Research on identifying radical and infectious threads, members, posting 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.

Dark Web [3] is portion of the web circumscribing the sinister objectives of extremist groups and specially the web forums with substantial prevalence of activities which supporting extremism. Another class called *Gray Web Forums* in which the discussions focus on topics that might potentially encourage offensive and disruptive behaviors and may disturb the society or threaten public safety. They include topics like pirated CDs, gambling, spiritualism, bullying and online-pedophilia.

There are many global extremist groups which perform radical activities like Islamic military groups, have created thousands of websites that support psychological warfare, fund raising, recruitment and distribution of propaganda materials [1]. They are tried to keep their agenda alive and attract more supporters, they always maintain certain level of publicity [5].

2. Related Work

Previous work is based on the web content analysis as well as identification of radical users.

1. Radical user identification

Previously studied works on the problem of radical user identification have been done in a business intelligence orientation for marketing product through targeted influential users. Ghosh and Lerman[6] work on the dynamics of voting on digging posts to rank radical users. They defined an empirical measure of influence based on number of in-network votes which post by receiver. Richardson and Domingos[8] worked on the social network formed from collaborative ratings and modeled it as markov random fields, considering each customer's product buying probability as a function of both is intrinsic desirability for the customer and the influence of others. Kempe et al [7] work on a greedy approach based discrete-optimization model to maximize the spread of influence through a social network. Kimura[9] et al. found that the computation cost of conventional greedy approach to identify influential nodes in a network is very high and consequently they proposed a method based on graph theory. Hill et al.[10] performed a statistical analysis on email network based marketing and established a hypothesis for a direct affect of network linkages on product/service adoption.

Java et al.[11] applied the influence models proposed by applying algorithms like PageRank, in blogosphere.

Agarwal et al. define a comprehensive definition of influential bloggers and the challenges associated with their identification.. Zhang et al[12] proposed expertise rank to rank the java expertise using forum threads and posts in the popular java forum. Tang and Yang contributed towards online health social networks, specifically the swine flu online forum which is based on the concept if page rank algorithm. They proposed UserRank to identify the influential users using content similarity and response immediacy it is shown as out-performing PageRank, in-degree and out-degree ranking. Tang and Yang [14] showed the application of UserRank algorithm in the domain of Dark Web forums.

2. Research on Dark web forum

A previous work [1] described how all major extremist organizations in the word, like Islamic military groups, show their presence on the Internet. They also performed a multi-region study on these organizations' Internet presence. In 1995 by Don Black, work on the Storm front, a white nationalist and supremacist neo-nazi Web forum was identified as the first hate side on the web [15]. Al Lab of the University of Arizona started to automatize the complete monitoring system and came up with their Dark Web Portal with different functionalities for data collection as well as analysis. The research on the dark web starts from the automatic accumulation of extremist websites and all related web data in a repository on which the data mining techniques are applied. It includes content analysis and user interaction analysis [13] as the main research area to analyze the sentimental and affects on the whole community. Ranging from automatic to semi automatic processes, several attempts have been made in the past for crawling and downloading of web pages from the surface web as well as hidden web.. Abbasi et a. differentiate affect analysis from sentiment analysis by characterizing it as assigning text with emotive intensities across a set of manually inclusive and possibly correlated affect classes. Skillicorn, work on a content analysis of Ansar forum for topic –based ranking of posts. Clustering of posts and threads has also been attempted in several studies to get communities with overlapping interest. Kramer analyzed Ansar forum for clustering based unsupervised anomaly detection with an objective to provide a robust, focus--of- attention mechanism to identify emerging threats in time dependent, unlabeled datasets. Huillier et al. consider a Dark Web forum as virtual communities of interests (VCoI) and performed a topic-based social network analysis of the Ansar community with an objective to discover key members. Based on the concept of page rank algorithm [13], devised the UserRank algorithm to rank influential users using content similarity and response immediacy.

3. Radically Influential Users

Radical user is the people whose thoughts are beyond the norm to an extreme antagonistic political,religious, racial, nationalist or any other ideology. This people do not have personal values for ethics and ratinalism, and are characterized by the term *radical*. This kind of thooughts arouse in minds when they feel of some unjust or discrimination happened with them either directly or indirectly, thogh it actually may be false. These thoughts are sometimes triggered by their personal involvement (e.g., death of a close relative or friend), political involvement (e.g., being a follower of a political or religious belief), and social involvement (e.g., racism, nationalism). values for ethics and ratinalism, and are characterized by the term *radical.*

4. Proposed System

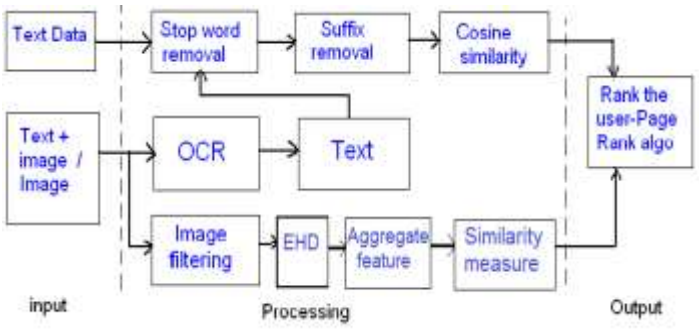

Fig. Architecture of proposed system

We propose an application of collocation theory to identify radical 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. The experiments are conducted on a standard data set to find radical and infectious threads, members, postings, ideas, and ideologies.

Proposed system to rank the user on text and image based similarity measures.

We make the following key contributions in proposed system:

- An application of analyze the data it may be text data or image data.
- If it is text data it will go through preprocessing stages like stop word removal, suffix removal, then by cosine function it check the similarity then decide whether that user is radical or not.
- If it image data, if it contain text data then it separate text from image by OCR technique. Send that text to text analysis and image goes through image preprocessing like image filtering, EHD it gives aggregate features, by similarity measures check similarity with training data set.
- Finally after measures of radicalness of user, it ranks the users by PageRank algorithm.

V. Implementation

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical and important stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

Text Analysis:

Stop word removal algorithm:

In computing, **stop words** are words which are filtered out before or after processing of natural language data (text). Though **stop words** usually refer to the most common words in a language, there is no single universal list of stop words used by all processing of natural language tools, and indeed not all tools even use such a list. Some tools specifically avoid removing these **stop words** to support phrase search.

Any group of words can be chosen as the stop words for a given purpose. For some search engines, these are some of the most common, short function words, such as *the*, *is*, *at*, *which*,

and *on*. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as "The Who", "Take that". Other search engines remove some of the most common words—including lexical words, such as "want"—from a query in order to improve performance.

Suffix removal algorithm

1) Suffix-stripping algorithms

Suffix stripping algorithms do not rely on a lookup table that consists of inflected forms and root form relations. Instead, a typically smaller list of "rules" is stored which provides a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include:

- if the word ends in 'ed', remove the 'ed'
- if the word ends in 'ing', remove the 'ing'
- if the word ends in 'ly', remove the 'ly'

Suffix stripping approaches enjoy the benefit of being much simpler to maintain than brute force algorithms, assuming the maintainer is sufficiently knowledgeable in the challenges of linguistics and morphology and encoding suffix stripping rules. Suffix stripping algorithms are sometimes regarded as crude given the poor performance when dealing with exceptional relations (like 'ran' and 'run'). The solutions produced by suffix stripping algorithms are limited to those lexical categories which have well known suffixes with few exceptions. This, however, is a problem, as not all parts of speech have such a well formulated set of rules. Lemmatization attempts to improve upon this challenge. Prefix stripping may also be implemented. Of course, not all languages use prefixing or suffixing.

2) Stochastic algorithms

Stochastic algorithms involve using probability to identify the root form of a word. Stochastic algorithms are trained (they "learn") on a table of root form to inflected form relations to develop a probabilistic model. This model is typically expressed in the form of complex linguistic rules, similar in nature to those in suffix stripping or lemmatization. Stemming is performed by inputting an inflected form to the trained model and having the model produce the root form according to its internal rule set, which again is similar to suffix stripping and lemmatization, except that the decisions involved in applying the most appropriate rule, or whether or not to stem the word and just return the same word, or whether to apply two different rules sequentially, are applied on the grounds that the output word will have the highest probability of being correct (which is to say, the smallest probability of being incorrect, which is how it is typically measured).

Some lemmatization algorithms are stochastic in that, given a word which may belong to multiple parts of speech, a probability is assigned to each possible part. This may take into account the surrounding words, called the context, or not. Context-free grammars do not take into account any additional information. In either case, after assigning the probabilities to each possible part of speech, the most likely part of speech is chosen, and from there the appropriate normalization rules are applied to the input word to produce the normalized (root) form.

Cosine algorithm

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1].

Note that these bounds apply for any number of dimensions, and cosine similarity is most commonly used in high-dimensional positive spaces. For example, in information retrieval and text mining, each term is notionally assigned a different dimension and a document is characterized by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter.

One of the reasons for the popularity of cosine similarity is that it is very efficient to evaluate, especially for sparse vectors, as only the non-zero dimensions need to be considered.

Image and Text analysis

a)OCR (Optical Character Recognition):

OCR is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into edible and searchable data. Imagine you've got a paper document - for example, magazine article, brochure, or PDF contract your partner sent to you by email. Obviously, a scanner is not enough to make this information available for editing, say in Microsoft Word. All a scanner can do is create an image or a snapshot of the document that is nothing more than a collection of black and white or colour dots, known as a raster image. In order to extract and repurpose data from scanned documents, camera images or image-only PDFs, you need an OCR software that would single out letters on the image, put them into words and then - words into sentences, thus enabling you to access and edit the content of the original document.

Fig : OCR Technique

b)Filtering an image

Image filtering is useful for many applications, including smoothing, sharpening, removing noise, and edge detection. A filter is defined by a kernel, which is a small array applied to each pixel and its neighbors within an image. In most applications, the center of the kernel is aligned with the current pixel, and is a square with an odd

number (3, 5, 7, etc.) of elements in each dimension. The process used to apply filters to an image is known as convolution, and may be applied in either the spatial or frequency domain.

Frequency Domain = Input pixel * Filter Function

Within the spatial domain, the first part of the convolution process multiplies the elements of the kernel by the matching pixel values when the kernel is centered over a pixel. The elements of the resulting array (which is the same size as the kernel) are averaged, and the original pixel value is replaced with this result. The CONVOL function performs this convolution process for an entire image.

Within the frequency domain, convolution can be performed by multiplying the FFT (Fast Fourier Transform) of the image by the FFT of the kernel, and then transforming back into the spatial domain. The kernel is padded with zero values to enlarge it to the same size as the image before the forward FFT is applied. These types of filters are usually specified within the frequency domain and do not need to be transformed. IDL's DIST and HANNING functions are examples of filters already transformed into the frequency domain. The following examples in this section will focus on some of the basic filters applied within the spatial domain using the CONVOL function:
Low Pass Filtering, High Pass Filtering, Directional Filtering, Laplacian Filtering

Since filters are the building blocks of many image processing methods, these examples merely show how to apply filters, as opposed to showing how a specific filter may be used to enhance a specific image or extract a specific shape. This basic introduction provides the information necessary to accomplish more advanced image-specific processing.

b) EHD (Edge Histogram Descriptor)

The edge histogram descriptor (EHD) in MPEG-7, generate an extra histogram bin from the 5-bin local edge histogram of each 4 × 4 sub-image. This extra histogram bin nothing but the ratio of the non-edge area (i.e., monotonous region) in the sub-image. Forming a feature vector with 6 edge/non-edge types, we can generate 33 different feature vectors (or 33 × 6 = 198 feature elements) including 16 vectors from 4×4 sub-images, 1 vector from a global histogram, 13 vectors from semi-global histograms, 1 vector from entropy, and 2 vectors from centers of gravity. A statistical hypothesis testing is employed to see which feature vectors/elements are most informative to differentiate different image classes. Experimental results show that non-edge and entropy features are the most informative features among all 33/198 feature vectors/elements.

c) Aggregate Feature

- Aggregate feature extraction nothing but feature extraction. In machine learning, pattern recognition and in image processing, **feature extraction** starts from an initial set of measured data and builds feature values intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.
- When large input data need to give to algorithm , then it must be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector).feature selection is nothing but subset selection from initial values. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

Feature extraction involves reducing the amount of resources required to describe a large set of data. When performing analysis of large and complex data, require number of variables. Analysis with a large number of variables generally requires a large amount of memory and computation power, also it may cause a classification algorithm to over fit to training samples and generalize poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

PageRank Algorithm:

Finally, using PageRank algorithm to rank the radical users. It identifies a ranked list of radically influential users in web forum. It gives result based on association and radical measures parameter.

$$\mathbf{PR}(p\,j) = (1-d) + d \times \forall pi : li\,j \in L\ \mathbf{prob}(p\,j\,|pi\,) \times PR(pi\,)$$

Where (p j) is small value of page rank score and linkages (L) among them are iteratively used to compute their new page rank score (PR (PJ)) using above equation. d[0..1] is damling factor typically set to 0.85. prob(Pi/Pj) is hyperlink from page Pi to Pj. The iterative process is continued untill a convergence ia achieved and the score ata that instance are accepted as their final page rank score.

6. Conclusion and Future Work

In this paper, we build an application of collocation theory to identify radically influential users in web forum. The radicalness of user is captured by a measure based on the degree of match of the commented posts with threat list, here we have considered text as well as image data. There are different collocation metrics are formulated to identify the association among users and they are finally embedded in a customized PageRank algorithm to generate a ranked list of radically influential users. The experiments are conducted on a standard data set which we preprocesses first then to find radical and infectious threads, members, postings, ideas and ideologies. Application system to rank the user on text and image based similarity measures.

An application of analyzed the data it may be text or image data. In this application we preprocesses the data then implemented so that gives correct result in PageRank algorithm.

References

5. J. Qin, Y. Zhou, and H. Chen, "A multi-region empirical study on the Internet presence of global extremist organizations," *Inf. Sys.*
T. Anwar and M. Abulaish, "Identifying cliques in

Dark Web forums—An agglomerative clustering approach," in *Proc. IEEE ISI*, Jun. 2012.
6. 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*. 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.
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. R. Ghosh and K. Lerman. (2010). "Predicting influential users in online social networks." [Online]. Available: http://arxiv.org/abs/1005.4882
9. D. Kempe, J. Kleinberg, and E. Tardos, "Influential nodes in a diffusion model for social networks," in *Proc. 32nd Int. Conf. ICALP*, 2005, pp. 1127–1138.
10. M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc.8th ACM SIGKDD Int. Conf. KDD*, 2002.
11. M. Kimura, K. Saito, and R. Nakano, "Extracting influential nodes for information diffusion on a social network," in *Proc. 22nd Nat. Conf. AAAI*, 2007.
12. S. Hill, F. Provost, and C. Volinsky, "Network-based marketing: Identifying likely adopters via consumer networks," pp. 256–276, 2006.