International Journal on Recent Researches In Science, Engineering & Technology
(Division of Computer Science and Engineering)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. It is an absolutely free (No processing charges, No publishing charges etc) Journal Indexed in JIR, DIIF and SJIF.
Research Paper
Available online at: www.jrrset.com

ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105

Volume 5, Issue 7, July 2017.

JIR IF : 2.54
DIIF IF : 1.46
SJIF IF : 1.329

Ranking of Radical Association amongst Users on Web Forum

R. Sathiya Priya, M. Edwin Jayasingh, K. Lalitha
Dhanalakshmi Srinivasan Institute of Technology, Samayapuram, Tamil Nadu, India

Abstract

In the current past, it has been found that the web is used as a device through radical or extremist organizations and users to perform countless types of mischievous acts with hid agendas and promote their ideologies in a sophisticated manner. Some of the net boards are in particular being used for open discussions on critical troubles influenced by using radical thoughts. We advocate software of collocation principle to pick out radically influential customers in net forums. The radicalness of a person is captured by using a measure based on the diploma of healthy of the commented posts with a threat list. The experiments are performed on a well-known statistics set to discover radical and infectious threads, members, postings, ideas, and ideologies. Proposed machine to rank the person on text and image-based similarity measures. We make the following key contributions in the proposed system: software of examining the statistics it may also be text facts or photo data. If it is textual content records it will go through preprocessing tiers like cease word removal, suffix removal, then by using a cosine similarity function, it tests the similarity with threat list then determine whether that person is radical or not. If its photo data, if it carries textual content statistics then it separates textual content from the image through OCR technique. Send that textual content to text analysis and image goes via image preprocessing like photo filtering, EHD to take aggregate features, using similarity measures it tests similarity with training records set. Finally, after measures of radicalness of user, it ranks the customers with the aid of Page Rank algorithm.

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

1. Introduction

In the recent past, it has been found that the internet is used by means of extremist groups, hate groups, racial supremacy groups, and terrorist companies on the net with numbers of multimedia websites, online chat room and net boards in posing grievous threats to our societies as properly as the countrywide security. The multimedia web sites promote psychological warfare, whereas chat room and net discussion board promote their techniques and ideologies thru discussions with naïve users. Public discussions among otherwise minded extremist companies lead to irascible talks accompanied by using abusive languages and promote online hate and violence. Now in a society internet forum is used as the most active medium being used for this motive [2]. Research on identifying radical and infectious threads, members, posting ideas and ideologies in internet forums for tracking the grievous threats posed via the energetic extremist and hate companies has received tremendous interest of the research community. Dark Web [3] is a component of the internet circumscribing the sinister objectives of extremist organizations and particularly the web boards with the substantial occurrence of activities which aiding extremism. Another category referred to as Gray Web Forums in which the discussions focus on topics that would possibly probably encourage offensive and disruptive behaviors and might also disturb the society or threaten public safety. They include matters like pirated CDs, gambling, spiritualism, bullying, and on line pedophilia. There are many global extremist groups which perform radical things to do like Islamic army groups, have created thousands of websites that assist psychological warfare, fundraising, recruitment and distribution of propaganda substances [1]. They are tried to keep their agenda alive and attract extra supporters, they always hold a certain degree of publicity [5].

2. Related Work

Previous work is based on the web content material evaluation as nicely as the identification of radical users.

1. Radical person identification

Earlier studied works on the problem of radical person identification have been carried out in a commercial enterprise brain orientation for advertising product thru centered influential users. Ghosh and Lerman[6] work on the dynamics of vote casting on digging posts to rank radical users. They described an empirical measure of affect based totally on a quantity of in-network votes which post by the receiver. Richardson and Domingos[8] worked on the social network shaped from collaborative scores and modeled it as Markov random fields, considering each customer's product buying chance as a characteristic of both is intrinsic desirability for the purchaser and the affect of others. Kempe et al [7] work on a grasping method based totally discrete-optimization mannequin to maximize the unfold of impact thru a social network. Kimura [9] et al. observed that the computation value of conventional greedy approach to perceive influential nodes in a network is very high and subsequently they proposed a method primarily based on format theory. Hill et al.[10] performed a statistical analysis on email network-based advertising and marketing and set up a speculation for a direct impact of network linkages on product/service adoption. Java et al.[11] applied the impact fashions proposed by means of making use of algorithms like PageRank, in the blogosphere. Agarwal et al. define a complete definition of influential bloggers and the challenges related with their identification... Zhang et al[12] proposed know-how rank to rank the java information using forum threads and posts in the popular java forum. Tang and Yang contributed closer to on-line health social networks, in particular the swine flu online discussion board which is based totally on the thought if the web page rank algorithm. They proposed UserRank to become aware of the influential users the use of content material similarity and response immediacy it is proven as outperforming PageRank, in-degree and out-degree ranking. Tang and Yang [14] showed the application of UserRank algorithm in the area of Dark Web forums.

2. Research on Dark internet forum

A preceding work [1] described how all fundamental extremist organizations in the word, like Islamic army groups, show their presence on the Internet. They also carried out a multi-region find out about on these organizations' Internet presence. In 1995 by means of Don Black, work on the Storm front, a white nationalist and supremacist neo-nazi Web forum have been recognized as the first hate website on the net [15]. Al Lab of the University of Arizona started out to automatize the whole monitoring gadget and came up with their Dark Web Portal with one-of-a-kind functionalities for information collection as well as analysis. The lookup on the dark net begins from the computerized accumulation of extremist web sites and all associated internet statistics in a repository on which the statistics mining methods are applied. It consists of content evaluation and person interplay analysis [13] as the predominant lookup region to analyze the sentimental and effects on the complete community. Ranging from automated to semi-automatic processes, several tries have been made in the previous for crawling and downloading of internet pages from the floor internet as properly as hidden web. Abbasi et a. differentiate affect evaluation from sentiment evaluation by using characterizing it as assigning text with emotive intensities across a set of manually inclusive and maybe correlated affect classes. Skillicorn, work on a content evaluation of Ansar discussion board for topic-based ranking of posts. Clustering of posts and threads has also been attempted in numerous studies to get communities with overlapping interest. Kramer analyzed Ansar discussion board for clustering primarily 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. reflect on consideration on a Dark Web forum as virtual communities of pastimes (VCoI) and carried out a topic-based social network analysis of the Ansar community with an objective to find out key members. Based on the idea of page rank algorithm [13], devised the UserRank algorithm to rank influential users the use of content material similarity and response immediacy.

3. Radically Influential Users

The radical consumer is the human beings whose thoughts are past the norm to an intense opposed political, religious, racial, nationalist or any different ideology. This people do now not have non-public values for ethics and rationalism and are characterized by means of the time period radical. This variety of ideas arose in minds when they feel of some unjust or discrimination came about with them either directly or indirectly, although it surely may be false. These thoughts are on occasion precipitated by means of their personal involvement (e.g., the loss of life of a shut relative or friend), political involvement (e.g., being a follower of a political or non secular belief), and social involvement (e.g., racism, nationalism). Values for ethics and rationalism, and are characterized with the aid of the term radical.

4. Proposed System

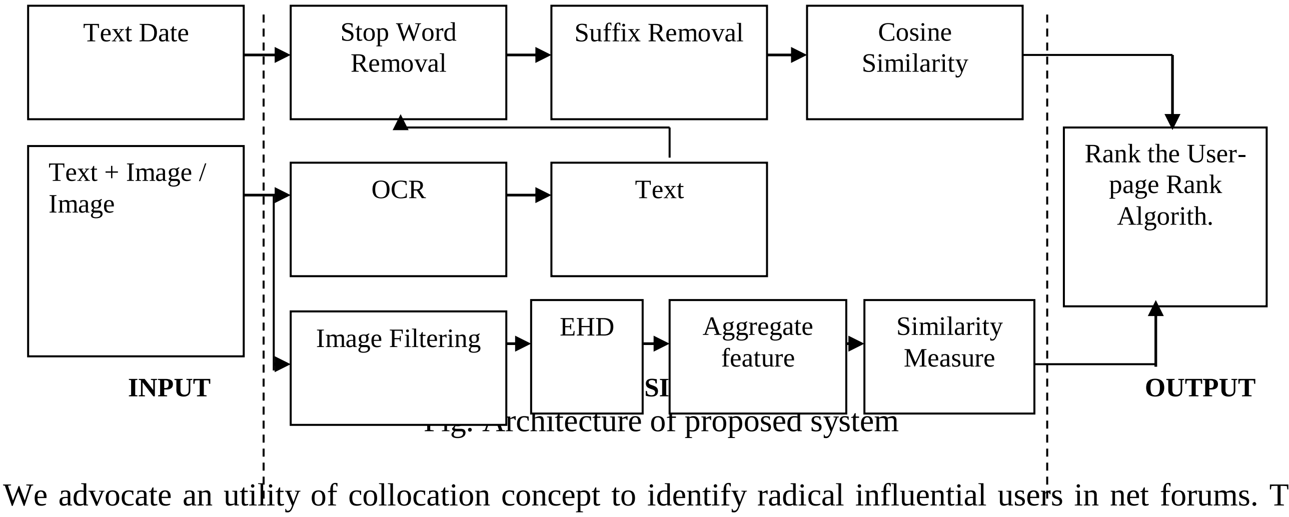

Fig. Architecture of proposed system

We advocate an utility of collocation concept to identify radical influential users in net forums. The radicalness of a person is captured through a measure based on the degree of suit of the commented posts with a danger list. The experiments are performed on a fashionable facts set to locate radical and infectious threads, members, postings, ideas, and ideologies. Proposed gadget to rank the user on textual content and image-based similarity measures.

We make the following key contributions in the proposed system:

An software of analyzing the information might also be textual content statistics or photo data. If it is text data it will go via preprocessing levels like cease word removal, suffix removal, then with the aid of a cosine function, it exams the similarity then determine whether that person is radical or not. If its photograph data, if it consists of text records then it separates text from the photograph by OCR technique. Send that text to textual content evaluation and photograph goes via photograph preprocessing like photograph filtering, EHD it gives combination features, with the aid of similarity measures take a look at similarity with coaching records set. Finally, after measures of radicalness of user, it ranks the customers by way of Page Rank algorithm.

Implementation

Implementation is the stage of the mission when the theoretical graph is turned out into a working system. Thus it can be viewed to be the most indispensable and necessary stage in attaining a successful new system and in giving the user, self belief that the new gadget will work and be effective. The implementation stage involves cautious planning, investigation of the current machine and it's constraints on implementation, designing of methods to acquire changeover and evaluation of changeover methods.

Text Analysis:

Stop phrase elimination algorithm:

In computing, quit phrases are words which are filtered out earlier than or after processing of natural language statistics (text). Though end phrases commonly refer to the most frequent phrases in a language, there is no single ordinary listing of give up words used by way of all processing of herbal language tools, and certainly now not all equipment even use such a list. Some equipment particularly avoid removing these stop phrases to assist phrase search. Any team of phrases can be chosen as the quit words for a given purpose., these are some of the most common, short feature words, such as the, is, at, which, and on. In this case, stop words can reason issues when searching for phrases that consist of them, specifically in names such as "The Who", "Take that". Other search engines get rid of some of the most frequent 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 no longer remember on a lookup table that consists of inflected forms and root structure relations. Instead, a normally smaller listing of "rules" is stored which gives a route for the algorithm, given an enter word form, to discover its root form. Some examples of the rules include:

- if the word ends in 'ed', put off the 'ed'
- if the phrase ends in 'ing', remove the 'ing'
- if the phrase ends in 'ly', cast off the 'ly'

Suffix stripping methods revel in the advantage of being a great deal easier 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 occasionally considered as crude given the terrible overall performance when dealing with extraordinary family members (like 'ran' and 'run'). The solutions produced with the aid of suffix stripping algorithms are constrained to those lexical classes which have nicely recognized suffixes with few exceptions. This, however, is a problem, as now not all parts of speech have such a nicely formulated set of rules. Lemmatization attempts to improve upon this challenge. Prefix stripping may additionally be implemented. Of course, not all languages use prefixing or suffixing.

2) Stochastic algorithms

Stochastic algorithms involve the use of likelihood to pick out the root form of a word. Stochastic algorithms are skilled (they "learn") on a desk of root structure to inflected shape relations to increase a probabilistic model. This mannequin is typically expressed in the structure of complicated linguistic rules, comparable in nature to these in suffix stripping or lemmatization. Stemming is carried out via inputting an inflected form to the skilled model and having the mannequin produce the root structure in accordance to its inner rule set, which again is similar to suffix stripping and lemmatization, without that the decisions worried in applying the most splendid rule, or whether or not or no longer to stem the phrase and just return the same word, or whether to apply two unique guidelines sequentially, are applied on the grounds that the output phrase will have the perfect likelihood of being correct (which is to say, the smallest probability of being incorrect, which is how it is normally measured). Some lemmatization algorithms are stochastic in that, given a word which can also belong to multiple parts of speech, a probability is assigned to every viable part. This may additionally take into account the surrounding words, referred to as the context, or not. Context-free grammars do not take into account any extra information. In either case, after assigning the probabilities to every viable section of speech, the most probably part of speech is chosen, and from there the excellent normalization regulations are applied to the input word to produce the normalized (root) form.

Cosine algorithm

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

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the attitude between them. The cosine of 0° is 1, and it is less than 1 for any different angle. It is therefore 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 adverse have a similarity of - 1, unbiased of their magnitude. Cosine similarity is especially used in fantastic space, the place the effect is neatly bounded in [0, 1]. Note that these bounds apply for any quantity of dimensions, and cosine similarity is most frequently used in high-dimensional tremendous spaces. For example, in records retrieval and text mining, every term is notionally assigned a distinctive dimension and a record is characterized by way of a vector the place the value of each dimension corresponds to the wide variety of times that time period appears in the document. Cosine similarity then gives a beneficial measure of how similar two archives are possibly to be in terms of their subject matter. One of the motives for the reputation of cosine similarity is that it is very efficient to evaluate, especially for sparse vectors, as solely the non-zero dimensions need to be considered.

5. Image and Text analysis

a) OCR (Optical Character Recognition): OCR is a technological know-how that enables you to convert exceptional sorts of documents, such as scanned paper documents, PDF files or photographs captured by using a digital digi cam into fit to be eaten and searchable data. Imagine you’ve obtained a paper file - for example, journal article, brochure, or PDF contract your accomplice sent to you through email. Obviously, a scanner is not enough to make this statistics handy for editing, say in Microsoft Word. All a scanner can do is create a photograph or a photograph of the record that is nothing extra than a series of black and white or coloration dots, known as a raster image. In order to extract and repurpose statistics from scanned documents, digi cam pictures or image-only PDFs, you need OCR software that would single out letters on the image, put them into words and then - phrases into sentences, as a result enabling you to get right of entry to and edit the content of the unique document.

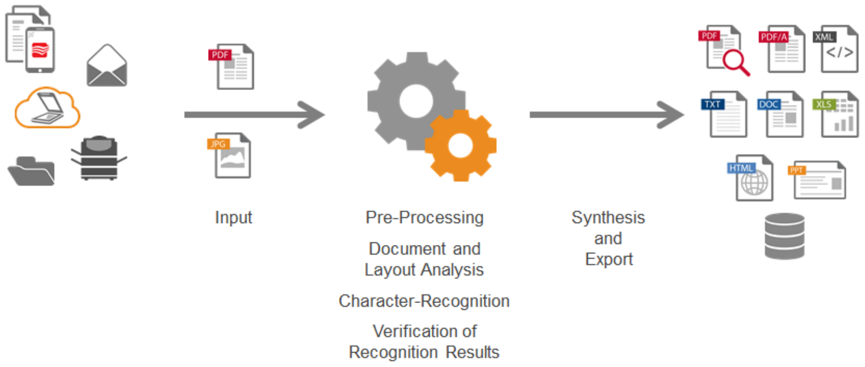

Fig : OCR Technique

Filtering an photograph Image filtering is useful for many applications, which includes smoothing, sharpening, eliminating noise, and aspect detection. A filter is described via a kernel, which is a small array utilized to each pixel and its neighbors inside an image. In most applications, the middle of the kernel is aligned with the present day pixel, and is a square with an atypical range (3, 5, 7, etc.) of factors in every dimension. The process used to apply filters to an picture is regarded as convolution, and may be utilized in both the spatial or frequency domain.

Frequency Domain = Input pixel * Filter Function Within the spatial domain, the first part of the convolution method multiplies the factors of the kernel through the matching pixel values when the kernel is centered over a pixel. The elements of the ensuing array (which is the equal measurement as the kernel) are averaged, and the authentic pixel fee is changed with this result. The CONVOL function performs this convolution manner for an whole image. Within the frequency domain, convolution can be performed with the aid of multiplying the FFT (Fast Fourier Transform) of the picture by means of the FFT of the kernel, and then remodeling again into the spatial domain. The kernel is padded with zero values to expand it to the same dimension as the photograph earlier than the forward FFT is applied. These sorts of filters are usually special inside the frequency domain and do no longer want to be transformed. IDL's DIST and HANNING features are examples of filters already converted into the frequency domain. The following examples in this section will focal point 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 constructing blocks of many image processing methods, these examples in simple terms exhibit how to practice filters, as adversarial to displaying how a unique filter may additionally be used to beautify a particular photo or extract a specific shape. This basic introduction offers the statistics quintessential to accomplish greater superior image-specific processing.

EHD (Edge Histogram Descriptor) The aspect histogram descriptor (EHD) in MPEG-7, generate an extra histogram bin from the 5-bin local aspect histogram of every 4 × four sub-image. This greater histogram bin nothing however the ratio of the non-edge place (i.e., monotonous region) in the sub-image. Forming a feature vector with 6 edge/non-edge types, we can generate 33 one-of-a-kind function vectors (or 33 × 6 = 198 function elements) including 16 vectors from 4×4 sub-images, 1 vector from a world histogram, thirteen vectors from semi-global histograms, 1 vector from entropy, and 2 vectors from centers of gravity. A statistical hypothesis trying out is employed to see which characteristic vectors/elements are most informative to differentiate special picture classes. Experimental effects exhibit that non edge and entropy elements are the most informative facets amongst all 33/198 function vectors/elements. c) Aggregate Feature

Aggregate characteristic extraction nothing but feature extraction. In computer learning, pattern attention and in photograph processing, feature extraction starts from an preliminary set of measured records and builds characteristic values meant to be informative and non-redundant, facilitating the subsequent getting to know and generalization steps, and in some instances main to better human interpretations. Feature extraction is associated to dimensionality reduction.

When large enter records want to give to algorithm, then it have to be processed and it is suspected to be redundant (e.g. the equal dimension in each toes and meters, or the repetitiveness of pics introduced as pixels), then it can be modified into a reduced set of aspects (also named a feature vector).feature decision is nothing however subset selection from initial values. The chosen elements are expected to contain the relevant statistics from the enter data, so that the favored task can be performed via the usage of this decreased illustration rather of the entire preliminary data. Feature extraction involves lowering the amount of assets required to describe a giant set of data. When performing analysis of massive and complex data, require range of variables. Analysis with a large range of variables normally requires a giant quantity of reminiscence and computation power; additionally it might also purpose a classification algorithm to over match to training samples and generalize poorly to new samples. Feature extraction is a everyday term for techniques of setting up mixtures of the variables to get around these problems while nonetheless describing the information with adequate accuracy.

Page Rank Algorithm:

Finally, the usage of PageRank algorithm to rank the radical users. It identifies a ranked listing of radically influential customers in internet forum. It offers end 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 fee of page rank rating and linkages (L) amongst them are iteratively used to compute their new web page rank score (PR (PJ)) the use of above equation. d[0..1] is damping aspect typically set to 0.85. prob(Pi/Pj) is hyperlink from web page Pi to Pj. The iterative process is endured until a convergence is finished and the score at that occasion are usual as their last page rank score.

6. Conclusion and Future Work

In this paper, we build a software of collocation principle to identify radically influential customers in net forum. The radicalness of user is captured with the aid of a measure based on the degree of fit of the commented posts with danger list; here we have considered text as nicely as image data. There are extraordinary collocation metrics are formulated to discover the association amongst customers and they are subsequently embedded in a personalized PageRank algorithm to generate a ranked list of radically influential users. The experiments are conducted on general facts set which we preprocesses first then to discover radical and infectious threads, members, postings, ideas and ideologies. Application system to rank the user on text and photo based similarity measures. An application of analyzed the data it may be textual content or picture data. In this software we preprocesses the facts then applied so that gives right end result in PageRank algorithm.

References

[1] Anwar, Tarique, and Muhammad Abulaish. "Ranking radically influential web forum users." IEEE Transactions on Information Forensics and Security 10.6 (2015): 1289-1298.

[2] Zhang, Yu, Zhaoqing Wang, and Chaolun Xia. "Identifying key users for targeted marketing by mining online social network." 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops. IEEE, 2010.

[3] Baig, Shahbaz S., and Kishor P. Wagh. "User dominance measure in online community Forum." Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2016 2nd International Conference on. IEEE, 2016.

[4] Tadas, Samiksha R., and K. B. Bijwe. "Ranking Radically Influential Web Forum Users on Social Media." (2017).

[5] Qin, Jialun, Yilu Zhou, and Hsinchun Chen. "A multi-region empirical study on the internet presence of global extremist organizations." Information Systems Frontiers 13.1 (2011): 75-88.

[6] Chen, Hsinchun, et al. "Uncovering the dark Web: A case study of Jihad on the Web." Journal of the American Society for Information Science and Technology 59.8 (2008): 1347-1359.

[7] Wang, Jau-Hwang, et al. "A framework for exploring gray web forums: analysis of forum-based communities in Taiwan." International Conference on Intelligence and Security Informatics. Springer, Berlin, Heidelberg, 2006.

[8] Qin, Jialun, et al. "Analyzing terror campaigns on the internet: Technical sophistication, content richness, and Web interactivity." International Journal of Human-Computer Studies 65.1 (2007): 71-84.

[9] Ghosh, Rumi, and Kristina Lerman. "Predicting influential users in online social networks." arXiv preprint arXiv:1005.4882 (2010).

[10] Kempe, David, Jon Kleinberg, and Éva Tardos. "Influential nodes in a diffusion model for social networks." International Colloquium on Automata, Languages, and Programming. Springer, Berlin, Heidelberg, 2005.

[11] Kimura, Masahiro, Kazumi Saito, and Ryohei Nakano. "Extracting influential nodes for information diffusion on a social network." AAAI. Vol. 7. 2007.

[12] Hill, Shawndra, Foster Provost, and Chris Volinsky. "Network-based marketing: Identifying likely adopters via consumer networks." Statistical Science 21.2 (2006): 256-276.