I.J. Intelligent Systems and Applications, **2014, 05, 44-54**
Published Online April 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2014.05.04

Discovering Hidden Networks in On-line Social Networks

Pooja Wadhwa, M.P.S Bhatia
Netaji Subhas Institute of Technology, Sector 3, Dwarka, New Delhi, India
E-mail:pooja.chopra82@gmail.com, bhatia.mps@gmail.com

***Abstract*—** Rapid developments in information technology and Web 2.0 have provided a platform for the evolution of terrorist organizations, extremists from a traditional pyramidal structure to a technology enabled networked structure. Growing presence of these subversive groups on social networking sites has emerged as one of the prominent threats to the society, governments and law enforcement agencies across the world. Identifying messages relevant to the domain of security can serve as a stepping stone in criminal network analysis. In this paper, we deploy a rule based approach for classifying messages in Twitter which can also successfully reveal overlapping clusters. The approach incorporates dictionaries of enriched themes where each theme is categorized by semantically related words. The message is vectorized according to the security dictionaries and is termed as 'Security Vector'. The documents are classified in categories on the basis of security associations. Further, the approach can also be used along the temporal dimension for classifying messages into topics and rank the most prominent topics of conversation at a particular instance of time. We further employ social network analysis techniques to visualize the hidden network at a particular time. Some of the results of our approach obtained through experiment with information network of Twitter are also discussed.

***Index Terms*—** Rule based Classification, Data Mining, Topic based Social Network Analysis, Security Mining

I. Introduction

With the growth of Web and underlying Information Communication Technologies (ICT), cyberspace has emerged as an endless repository of knowledge base which has a lot of potential for exploration. Today, Internet offers unparalleled opportunities for facilitating the exchange of information across the 'Global Village', thus stimulating and sustaining liveliness on web. Growth of online social networks can be attributed to the inherent benefits which they provide: - ease of access, little or no government control, potentially huge audience, rapid flow of information, the inexpensive development, maintenance and the anonymity of communication. Furthermore, these inherent benefits associated with the use of online social networks, have also paved way for a wave of threats arising out of the use of social networks by extremists and have motivated a great deal of research in the field of application of advanced information technology in the analysis, anticipation and countering of these threats.

According to a majority of experts in the field of terrorism and counterterrorism, the presence of subversive groups and organizations has been increasing at an alarming rate [1]. Cyber extremism, cyber hate propaganda (also called online radicalization) have emerged as prominent threats to the society, governments and law enforcement agencies across the world. There have been many instances on the web highlighting the use of internet by extremist groups [1][2]. Modern means of communication, such as e-mail, forum, Blogs, chatrooms, websites and social networking sites such as Twitter, Facebook have provided a platform to these subversive groups where they can meet, exchange ideas, plan, coordinate and propagate their ideas. Further, the use of social media by these groups has fuelled the motivation towards the analysis of rich information available on the web.

Investigative data mining techniques incorporating the use of data mining and social network analysis have been widely used by law enforcement agencies in criminal network analysis [3][4][5]. Since the platform is not monitored by the government / agency, makes itself more vulnerable to its use by these elements as it provides them to connect with their groups/leaders online at any time, post messages and ideologies. Another important aspect of consideration while carrying out the analysis of these groups is that pure link analysis may lead to inconsistent results as the importance of a member in a group cannot be understood by analyzing its connections to other members at any time without taking into consideration the topic of conversation [6]. A situation may occur, where a member might be posting messages relating to different agendas at different instance of time [6]. The content analysis of the message may therefore reveal the presence of a hidden group which is related to a specific topic.

Thus, the detection of subversive groups or radical groups in online social networks poses two major challenges:-

1. Capturing and processing large amount of data to filter out the relevant informative content.
2. To uncover/identify the hidden groups prevailing in the online social networks related by the common topic of interest.

Social Network Analysis (SNA) provides tools to examine relationships between people. Text Mining (TM) allows filtering the relevant text produced by users of Web 2.0 applications, however it neglects their social structure. This paper applies an approach to combine the two methods named “content-based SNA” and aims to address the above two issues by incorporating the use of a rule based classification approach to classify the messages in categories which are relevant to the domain of security and then applying techniques of social network analysis to view the network of these subversive groups which are revealed by topic based categories. The approach incorporates dictionaries of enriched themes where each theme is categorized by semantically related words. The document/ message is vectorized according to the security dictionaries and is termed as ‘Security Vector’. The documents are classified in categories on the basis of security associations. Further, the approach can also be used along the temporal dimension for clustering messages into topics and rank most prominent topics of conversation at a particular time.

The paper is organized as follows: Section II presents an overview of the existing literature on the subject. Section III presents the broad overview of the approach. The approach for uncovering the hidden subversive groups in online social networks is explained in Section IV. The results and interesting findings observed out of our experiment with information network of Twitter are discussed in Section V. Section VI concludes by presenting future directions of our research.

II. Related Work

Data mining has emerged as a powerful tool enabling crime investigators to explore large databases quickly and efficiently. Many methods and tools have emerged to assist investigators in data analysis for crime investigation. Today law enforcement agencies across the world are facing challenge in analyzing plethora of data present on the web. Intelligence agencies face significant challenge in pre-processing and analyzing data [7]. It has also been proven that social network analysis can play an important role in discovering criminal communities from a well structured database. Social Network Analysis (SNA) techniques are designed to discover patterns of interactions between social actors in social networks [8], they have been found to be exceptionally useful for studying criminal networks [9][10]. In a criminal network, traditional data mining techniques such as cluster analysis can be applied to detect underlying hidden groupings followed by SNA to identify the patterns of interactions between subgroups [11].

Yang and Ng. [12] have presented a method to extract criminal networks from websites that provide blogging services by using a topic-specific exploration mechanism. Chen et al. [13] have successfully employed data mining techniques to extract criminal relations from a large volume of a police department’s incident summaries. Furthermore, Yong et al. [14] [15] have shown that text information on terroristic activities from various network media can be effectively processed by text data mining and visual network models. A general framework for crime data mining is proposed incorporating the data mining techniques such as entity extraction, association, prediction and pattern visualization [4]. Chau & Jennifer Xu [5] have applied techniques such as pattern matching and rule-based algorithms for extracting useful information related to online hate groups in blogs. The study of extremist groups and their interactions have always attracted the interest of security community.

L’Huiller et al. [16] have employed topic-based social network analysis for identifying key community members and Latent Dirichlet Allocation model (LDA) has been used for identifying topics from dark websites. A multi-stage clustering algorithm to identify named-video clusters in YouTube online video community has been proposed by Gargi et al. [17].

Analysis of communication over social networks has been a hot topic over the last several years. Klerks [18] has examined the underlying graph structure in social networks to understand the dynamics of how an information propagates in large networks. This is extremely important as it can be used to understand the influence of a topic. Further, the study of dynamics of social network is also an interesting area of research and has motivated a number of researchers on the subject. According to Hu, D., [19] dynamic social network analysis involves the study of three major issues:- network recovery, network measurement and statistical analysis. Network recovery refers to a process by which multiple snapshots of a network are constructed from the longitudinal data so as to model the evolution of a network. Constructing network at discrete time intervals consists of taking cross-sectional snapshots of time. Hence, longitudinal analysis focuses on the change from one network state to another without any (explicit) reference to the sequence of changes between the intervals of time. Due to the cost and practical design implications in research, most longitudinal network studies use discrete time. Network measurement deals with measures which are used over time to describe the changes in network. These include measures for static network analysis like degree, closeness, betweenness etc. Changes in the above values with time reveal information about network evolution. Apart from these probabilistic measures like degree distribution and clustering coefficient are also used to explain the dynamic process of growth [20]. Statistical analysis procedures for networks are initially used to explain the emergence of network topologies. The process of detection and analysis of hidden subversive groups in online social networks requires that the temporal dimension be incorporated so as to closely understand the dynamics of such group with time.

III. Research Contribution

The explosive rate at which data is growing on web poses limitations in its analysis which is predominantly governed by the underlying techniques used and the analyst's capabilities. Gathering data from large social networks such as Facebook and Twitter has become increasingly more difficult due to the external pressure to improve privacy and internal pressure to generate revenue [21]. We examine the case of Twitter, a micro-blogging social network in our paper. We propose the following in this paper which is an extension of our previous work [22]:-

(i) Creation of Security Dictionaries as per topics related to security. We will have an exhaustive dictionary for each topic consisting of semantically related keywords.

(ii) Document Vectorization incorporating Security related features: The approach caters to the fact that the presence of terms relevant to security is more important that their frequency of occurrence.

(iii) Determination of Security Association rules and Rule Pruning: This step takes into account the security associations among two or more terms to reveal final new categories of topics. It can also yield one to many class mapping i.e. we can classify a message in more than one class. Once the security association rules are identified corresponding to each category, they are subsequently pruned so as to cater only the relevant rules.

(iv) Message Classification according to security association rules.

(v) Creation of Topic based social network.

(vi) Temporal Category Rank: Simplicity of the technique can be extended to compute topic rank along temporal dimension.

IV. Research Contribution

There has been a steady rise in the extremism on the internet on many social networking sites [2]. Our research will address the problems of detecting hidden subversive groups with the use of investigative data mining techniques incorporating Web mining for information Extraction, Traditional data mining for data pre-processing, filtering, a variant of rule based approach for message classification and social network analysis for community tracking and analysis. The approach at present will focus only on the textual content present in the message as we feel that the presence of specific words serve as important indicators and hence we will ignore the hyperlinks posted about the pictures and other websites. Our approach considers the fact that pure link analysis may lead to inconsistent results as the importance of a member in a group cannot be understood by analyzing its connections to other members at any time without taking into account the context of conversation. This becomes a prime concern when we are dealing with information network like Twitter. Keeping this in view, we present a process flow chart of our approach in Fig. 1, and is explained as follows:-

A. **Gathering data** from large social networks such as Facebook and Twitter has become harder than it used to be several years ago due to external pressure to improve privacy and internal pressure to generate revenue [21]. We have chosen Twitter for our case study, since it provides APIs which allow us to access data. The data will be captured corresponding to specific hashtags through our customized crawler for a week. This will provide the first level of filtering, as we are able to fetch messages of our interest. Messages/ tweets captured will have additional details like sender information, receiver information, date, time, language, whether it is a reply_to message along with the text. Thus, each tweet will have a format comprising of the above mentioned fields along with text. The messages captured will be stored in an excel database from where they can be further processed. We have fixed the maximum length of each excel file to be comprising of ten thousand tweets per file.

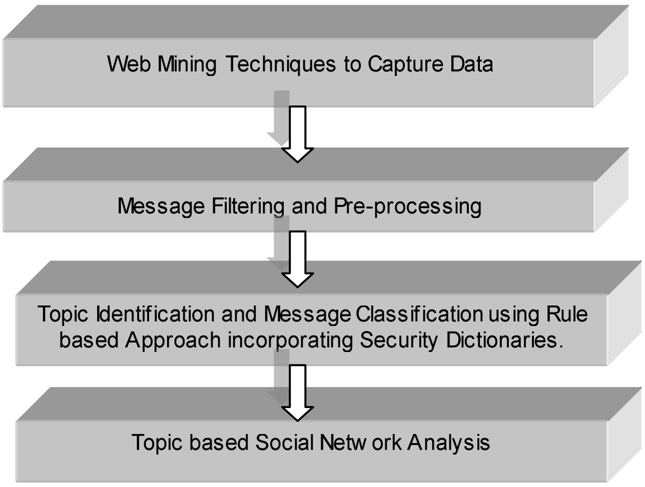
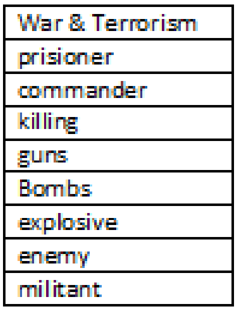

Web Mining Techniques to Capture Data
Message Filtering and Pre-processing
Topic Identification and Message Classification using Rule based Approach incorporating Security Dictionaries.
Topic based Social Network Analysis

Fig. 1: Broad Approach

B. **Message Filtering and Pre-processing**:-After the messages have been captured, each excel file will be pre-processed so as to remove unwanted words. This will help us to reduce the unnecessary processing overhead. Firstly, the messages are filtered for English language keeping in view that messages corresponding to other languages will require additional language processing capabilities which are at present not available with us. In order to apply text mining capabilities on message, a message will be tokenized where a token will correspond to a word. Our approach relies of the fact that the presence of specific words in the message provide indication about the context, hence we remove the hyperlinks present in the message. Furthermore, the messages will be stripped off the stopwords like ‘of’, ‘on’, ‘for’, ‘at’. We deploy NLTK library, which has a rich list of stopwords to generate an initial list of stopwords. Additionally, we also create a ‘waste-list’ comprising of words which are irrelevant to the domain of security and whose presence is not required in the message. We have created an exhaustive waste-list for our experiment. After the unwanted words are removed, the size of the message will reduce significantly. Finally, we perform the task of stemming where all the words which are extended from root words are replaced by root words. We initially employed stemming module from NLTK library to perform the stemming. However, to improvise our results we created our own stemming module so as to stem certain words which are relevant in terms of security like ‘jihadology’, ‘jihadist’, ‘jihadi’ were mapped to ‘Jihad’. Thus, Filtering and pre-processing formed a very crucial step in order to successfully apply the approach.

C. **Topic Identification and Message Classification using Security Associations**:-The explosive rate at which data is growing on web poses limitations in its analysis which is predominantly governed by the underlying techniques used and the analyst’s capabilities. So this brings to light an important factor that the inclusion of domain security features in message/data classification can improve results significantly. This approach takes into account the knowledge of domain security while classifying messages. The main steps of the approach [23] are as follows:-

(i) Creation of Security Dictionaries. A Security dictionary is an exhaustive dictionary for each topic consisting of semantically related keywords. Creation of Security dictionaries is an essential and time consuming step. Firstly, topics are identified in consultation with domain experts and then dictionaries are created so as to augment with semantically related words and synonyms. For each topic a separate dictionary exists which serves as a baseline for revealing topics. This was a very effective step as identification of words relevant to categories which are normally used online can be used to create security dictionaries which can be enriched from time to time. Further, considering the application of the approach in security domain where obtaining security dictionary is hard, creation of dictionaries with temporally enhanced words related to topics can serve as a baseline for all future security endeavors. A partial security dictionary for topic ‘War-Terrorism’ is shown in Fig. 2.



War & Terrorism
prisioner
commander
killing
guns
Bombs
explosive
enemy
militant

Fig. 2: A Partial Security Dictionary for ‘War-Terrorism’

(ii) Document Vectorization: Once subcategories have been decided, document vector is represented as V, where V= [v_1,v_2, ...v_n] where 1..n refer to subcategories which are enriched security dictionaries of themes of interest to the security communities. Value in any column vector is 1, if the document contains words relevant to that category. For example, if we have chosen three categories for our messages say V= [‘Jihad’, ‘Terrorism’, ‘Country’] then D_1 refers to document vector for document 1. D_1= [1, 0, 1] means that document contains words relevant for categories ‘Jihad’ and ‘Country’. Thus, the document is vectorized not according to the frequency of terms but rather on the basis of presence of security related keywords. This has relevance keeping in view if a person is talking about ‘Alqaeda’, then the number of times he talks about it is less important than the fact that keyword ‘Alqaeda’ interests him.

(iii) Security Associations refer to rules which help us to predict final topics. They take into account the presence of one or more words relevant to predefined categories and deduce final categories of topics for classification. Practical experiments reveal that many times two or more subtopics combine to reveal more appealing topics. For example we may have topics like ‘Jihad’ and ‘Country’ but when two or more related topics occur together in a message we can have a more relevant topic like ‘Global Jihad’ which refers to Jihad present in many countries. Such types of security associations can be mapped to visual form such as ‘Topic Hierarchy’. Topic Hierarchy for ‘Global Jihad’ is shown in Fig. 3. It can be seen that if topics corresponding to ‘Jihad’ and ‘Country’ come together with topic of ‘Media’, we can infer that it must be referring to ‘Global Jihad’ as usage of ‘Media’ with ‘Jihad’ and ‘Country’ may either reveal media coverage of jihad in specific country or media of a specific country talking about ‘Jihad’, both of which raise a global alarm for ‘Global Jihad’. The corresponding security association rule is shown in Fig. 4.

However, many times a situation may occur where we may find that presence of two or more topics can lead to discovery of more than one class of topics. For example in Fig. 3, we may observe that presence of words related to ‘Country’, ‘Jihad’ and ‘Operations’ may reveal new classes of topics such as ‘Global Jihad’ and ‘Global Operations’. Here, learning of security association rules is achieved by observing training message burst along with security domain expert who reveal coexistence of topics which are of interest to the security community but may occur rarely. Hence, such rules are also incorporated in our engine.

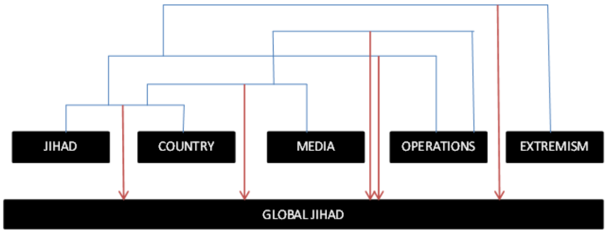


Fig. 3: Topic Hierarchy for ‘Global Jihad’

{Jihad, Country} → *{Global Jihad}*

{Jihad, Country, Media} → *{Global Jihad}*

{Jihad, Country, Operations} → *{Global Jihad, Global- Operations}*

Fig. 4: Security Association rules

(iv) Pruning of Security Association Rules: Though the rule creation step mentioned above requires manual effort, to compare a document vector against all the possible rules to identify category will consume lot of time and memory. To overcome this, we propose a rule pruning step which will automatically identify the most appropriate rule from rule base corresponding to a category. The rule pruning phase is explained as follows:-

(a) List all possible rules against each category. This leads to the listing of all rules against each category which need to be optimized.

(b) Compute Support for each predefined category subtopic among rules corresponding to a category. The support Sup (X) of an itemset ‘X’ is defined as the proportion of transactions in the data set which contain the itemset. It is a measure in association rule mining, intended to identify strong rules discovered in databases. In our case, the notion is based on the fact that subtopics which have a support value of ‘1’ seem to be essentially present in a document vector for a document/ message to be classified in that category. Similarly, subtopics which have a support count of ‘0’ serve as indicators that absence of subtopics of a specific category, also make document, a viable candidate for classification in relevant category. Thus the rule pruning will involve words with support value ‘1’ combined with words having support value ‘0’. Whereas, support count >0 and support count <1 for certain subtopics indicate that presence/ absence of such words have limited impact on classification in the final category. Thus, a rule will be of form (1) where R_c refers to pruned rule for category ‘C’ , $Subtopic_i$ refers to subtopic with support value ‘1’ for category ‘C’ and $Subtopic_j$ refers to subtopic with support value of ‘0’ for category ‘C ‘ respectively.

Thus, Pruned Rule for each Category:-

$$R_c = \sum_{i=1}^{i=N} Subtopic_i \bigcup \sum_{j=1}^{j=N} Subtopic_j \qquad (1)$$

D. Topic Based Social Network Analysis

As observed, data related to the activity of the subversive groups is dynamic [7] and requires time based collection, analysis and response, but due to the limitation of scalability of human skills and abilities, a gap is bound to exist. We consider the fact that pure link analysis without taking into account the context of conversation may lead to inconsistent results as it might not provide any insight into the topic of conversation, behaviour of member of such groups in a network as the Topic changes.

In order to carry out social network analysis and evolution of subgroups, we first need to determine the number of communities. In a topic based social network analysis, the number of communities will correspond to the number of categories identified. Once the number of communities ‘k’ has been decided we will model the behaviour of nodes which correspond to the member of groups with time. Hence, dynamic social network analysis and visualization will serve as a crucial step in understanding the complex process of evolution of members in these groups. After the messages are grouped into various subtopics/categories, the information will be mapped to a graph which we refer to as ‘Conversation Cluster’ which will have a directed edge from node ‘A’ to node ‘B’ if there is any communication among the two within that specific amount of sampled time interval (which in our case is 60 minutes). It is also possible that at any instance of time, a member is posting messages related to a specific topic, but not directed to any specific individual. In that case, we will group the nodes in the same cluster but with self-directed edges only. This will yield us a forest within a topic cluster which will gradually evolve with time. In an evolving network, group metric like cohesion, which measures the extent to which members within a group are related, find little or no relevance in initial stages but once connections start emerging, metric provides a useful insight. However, we can apply group metric like 'Group Stability' [24] to measure stability of group membership with time.

In order to measure the influence of a topic at an instance of time, we propose a 'Topic Entity map' which is a star network containing topic as the central hub and nodes are the entities/ users who posting messages relevant to the topic. Thus, at each instance of sampled time, we will have new nodes being added to star network, the visualization will result in active nodes in community. As shown in Fig. 5. We can see that nodes 'A', 'B', 'C', 'D' and'E' are talking about a subtopic. Thus, the number of active nodes gives the degree of 'Topic Influence' at any instance of time. Topic clusters may also be arranged according to the number of active nodes at any instance of time to reveal more talked about topic.

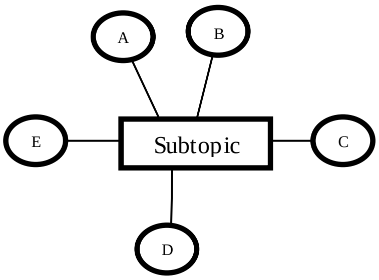


Fig. 5: Topic-Entity Map

As, it is also possible to get a message mapped to more than one category, thereby revealing overlapping clusters, we can compute the degree of overlap by counting the number of nodes which fall in overlap.

'Topic-Entity map' at any instance provides an insight into the most active topic of conversation along with the most active users. However, if one wants to find out how the users are communicating among each other within topic-based network, we draw a directed graph with an edge between users if they are communicating between themselves relating to the topic of interest. This leads to a topic-based network which we refer to as 'Conversation Cluster' which will be a forest with more of user nodes and less number of edges at any instance of time which might evolve into a dense forest with time. A 'Conversation Cluster' is of much importance in our analysis since it provides a useful insight into the most prominent user or reveals an important 'Bridging node' which will be a user who is active in two or more topics of communication at any instance of time. Identification of such 'information rich nodes' will play an important role in spreading information or receiving an important information.

E. Temporal Category Rank

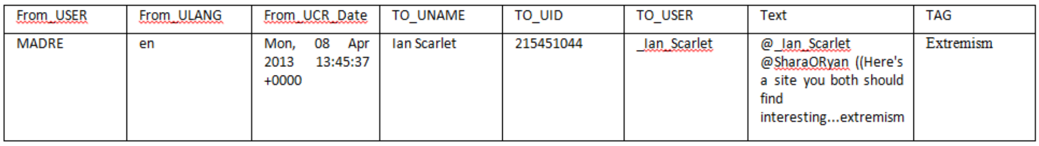
Simplicity of the technique can be extended to compute topic rank along temporal dimension. At any instance of time we can compute the size of a topic cluster and rank topics according to the size of cluster. Here, the size of cluster reveals the number of users talking about a particular topic at any instance of time. This helps us to identify the top topics according to the number of users involved in them. Thus, a topic may be termed as more 'popular' if it occupies a high rank at an instance of time or maintains high rank over a period of time.

V. Experimental Results

We carried some experiments with our approach on Twitter and designed a customized crawler in python for capturing data corresponding to hashtag 'Alqaeda' for testing the accuracy of our approach. The data was captured for a week, however to validate our approach we carried out hourly analysis of data. The results of our experiment are discussed below:-

A. Web Mining Techniques to Capture Data:

With the gradual move of subversive groups towards social networking sites, we conducted experiments with our approach on the social networking site due to its popularity, general open nature and availability of limited APIs for extracting data. Twitter, launched in 2006 has several hundred million users and is primarily categorized as a micro-blogging social networking site whose posts (tweets) can have a maximum of 140 characters. We designed our customized crawler in python to capture tweets according to hashtag 'Alqaeda'. The tweets were initially filtered for English language and stored in the form of excel files. General format of tweet is shown in Fig. 6. Each of our captured tweet contains information about sender (represented by From_User), receiver (To_UNAME, TO_USER, TO_UID) if any, text (represented by Text), Date of posting (From_UCR_Date) , Tag (represented by TAG) and language (From_ULANG) as shown in Fig. 6. respectively. Each excel file stored 10,000 tweets and a new file was created for additional tweets.

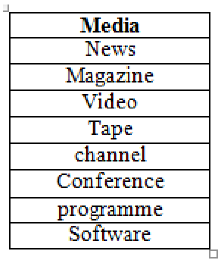


From_USER	From_ULANG	From_UCR_Date	TO_UNAME	TO_UID	TO_USER	Text	TAG
MADRE	en	Mon, 08 Apr 2013 13:45:37 +0000	Ian Scarlet	215451044	_Ian_Scarlet	@_Ian_Scarlet @SharaORyan ((Here's a site you both should find interesting...extremism	Extremism

Fig. 6: General tweet format as captured through crawler

B. Message Filtering and Pre-processing:

The tweets captured were pre-processed by tokenizing, stopwords removal, removal of unwanted words according to 'wastelist' and then by finally performing stemming. This was a major step as it reduced the size of messages significantly.



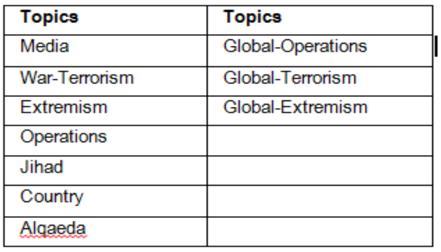
Media
News
Magazine
Video
Tape
channel
Conference
programme
Software

Fig. 7: Partial Security Dictionary for 'Media'

C.Topic Identification and Message Classification using security associations: After pre-processing the messages, they were classified into topics on the basis of rule based approach augmented by security dictionaries and support based pruning of rules. Initially security dictionaries were created for topics like 'Media', 'Operations', 'War-Terrorism', 'Extremism', 'Jihad', 'Country' and 'Alqaeda'. We created extensive dictionaries for each category. These categories were chosen with a view that they are the topics which are of prime concern to law enforcement agencies. They provided an initial set of topics in hashtag based conversation. Part of the Security dictionaries corresponding to topics like 'Media' and 'Operations' are shown in Fig. 7 and Fig. 9. After the creation of security dictionaries, topic hierarchies were created for all possible combination of topics which may occur together to yield a new topic. This resulted in the identification of new security association rules. These were then pruned for each category. The pruned rules for some of the categories are shown in Table 1. On the basis of security associations, new topics were also identified on the basis of co-occurrence on one or more topics. The final list of topics is shown in Fig. 8. This lead to the identification of security association rules. These were then pruned for each category. The pruned rules for some of the categories are shown in Table Each message after initial pre-processing was converted to document vector D=[V_1, V_2, V_3, V_4, V_5, V_6, V_7] for categories ['Media', 'War-Terrorism', 'Extremism', 'Operations', 'Jihad', 'Country', 'Alqaeda'] in the order.

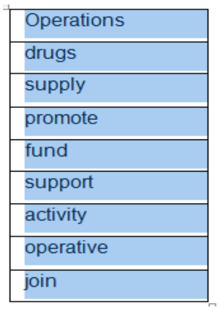
Table 1: Pruned Rules for Some Categories

Category	Pruned Rules
'Media'	If ('Media') v (~ 'War-Terrorism') v (~ 'Extremism') v (~ 'Operations') v (~ 'Jihad') v (~ 'Country')
'War-Terrorism'	If ('War-Terrorism') v (~ 'Country') v (~ 'Jihad')
'Global- Terrorism'	If ('War-Terrorism') v ('Country') v (~ 'Jihad')
'Jihad'	If ('Jihad') v (~ 'War-Terrorism') v (~ 'Country')
'Global-Terrorism' and 'Global-Jihad'	If ('Jihad') v ('War-Terrorism') v ('Country')



Topics	Topics
Media	Global-Operations
War-Terrorism	Global-Terrorism
Extremism	Global-Extremism
Operations	
Jihad	
Country	
Alqaeda	

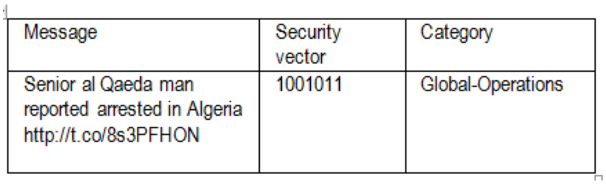
Fig. 8: Final categories of Topics



Operations
drugs
supply
promote
fund
support
activity
operative
join

Fig. 9: Partial Security Dictionaries for 'Operations'

An example of a classified message represented as a document vector is shown in Fig. 10. Category in which a message falls is shown in rightmost column.



Message	Security vector	Category
Senior al Qaeda man reported arrested in Algeria http://t.co/8s3PFHON	1001011	Global-Operations

Fig. 10: Example of Classified Message Vector

Using the above mentioned approach, we were able to classify messages in twitter into relevant categories successfully.

D. Topic based Social Network Analysis

In this paper, we use content-based SNA, an approach to combine SNA and Text Mining (TM) in order to find out the topic of communication among users thereby revealing a content-based Social Network. Content-based SNA consists of extracting overlapping topic related sub-networks from the entire communication network. In our case, we were able to discover subgroups of users related together by topic of conversation at any instance of time. Once the messages were classified, we retrieved the users who had posted the messages corresponding to categories and constructed a network which can be analysed using techniques of social networks analysis.

Keeping in view, the importance of topic of conversation in information network like Twitter, we classified the messages into the topic of conversation and created 'Topic-Entity' maps along with topic based social network corresponding to topics at different instances of time and observed the changes in the user's interests to get an insight into most active topic of conversation which could be of importance to the law enforcement agencies. The two Topic-Entity maps sampled at successive interval of 60 minutes are shown in Fig. 11 and Fig. 12. respectively. It can be seen that during the first interval, users are mostly talking about 'War-Terrorism', 'Global-Terrorism' and 'Alqaeda' (shown in square nodes in Fig. 11), users are shown as nodes which are connected directly to the topic (centroid of the graph). Topic-Entity map also provides an insight into the most talked about topic at any instance of time as we can see in Fig. 11, 'Alqaeda' and 'War-Terrorism' are the most talked about topics with maximum number of users related to them.

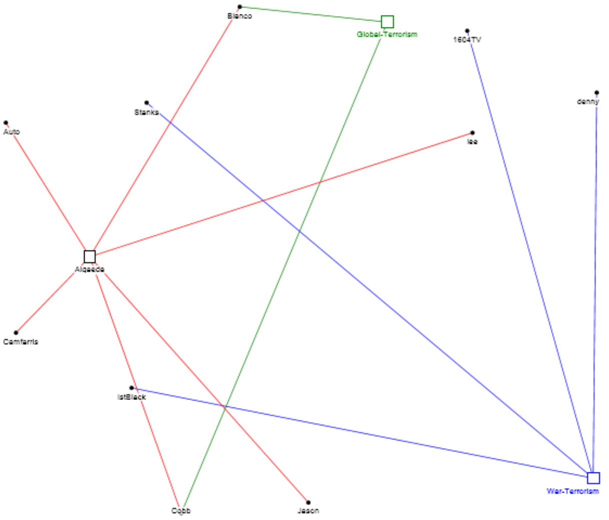


Fig. 11: Topic-Entity Map for hashtag 'Alqaeda' for a duration of 60 minutes

Similarly, the 'Topic-Entity' map in Fig. 12 provided some more useful insight for the next successive interval. We could see many new topics arising for discussion, some of the prominent ones being 'Country', 'War-Terrorism', 'Alqaeda', 'Global-Jihad'.

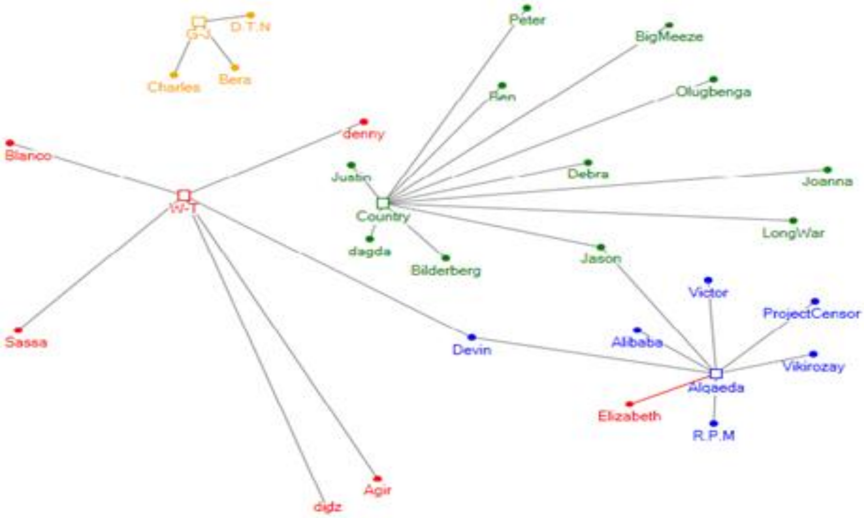


Fig. 12: Topic Entity Map for next successive Interval

In order to understand the level of conversation among members on a topic in our twitter network, we draw a directed graph called 'Conversation Cluster' which was actually a forest where few or more will be talking among themselves. A 'Conversation Cluster' corresponding to Fig. 11 is shown in Fig. 13. A careful insight into the graph provides information about three communities with respective users. It can be seen that user 'Cobb' and 'Blanco' (for simplicity we have replicated them as two entities in each of the sub-community as 'Cobb1' and 'Cobb2', 'Blanco1' and 'Blanco2' respectively. These users are actively posting tweets related to topics 'Global-Terrorism' and 'Alqaeda' and seem to be very important from the point of information flow and are also posting direct messages to each other thereby facilitating information exchange. The graph at this instance seems to be evolving and can be expected to mature with time. Such a 'Conversation Cluster' can be very useful as the number of users converse more among themselves, a conversation is likely to continue for a long time.

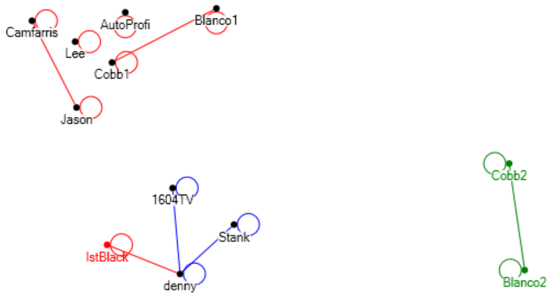


Fig. 13: Conversation Cluster for Fig. 11

A more useful insight is provided by 'Conversation Cluster' of next successive interval as it reveals more topics of interest and involved user community. It can be seen from Fig. 14, user named 'Blanco' is still active in community 'War-Terrorism' and users 'Devin' and 'Jason' being bridge node in communities 'War-Terrorism' and 'Alqaeda' respectively. A careful insight into each community lists nodes of high degree values highlighting the fact that they are more active nodes in communication at that time. Another hidden information which can be revealed by looking into second 'Conversation Cluster' is the hidden information flow link which can be deduced from our previous knowledge is that user 'Blanco' had been active in community 'Alqaeda' whereas at present he is actively involved in communities 'War-Terrorism' and 'Country' revealing a direct information flow between three communities. Such an inference makes 'Blanco' the most influential user and is of high significance in our research.

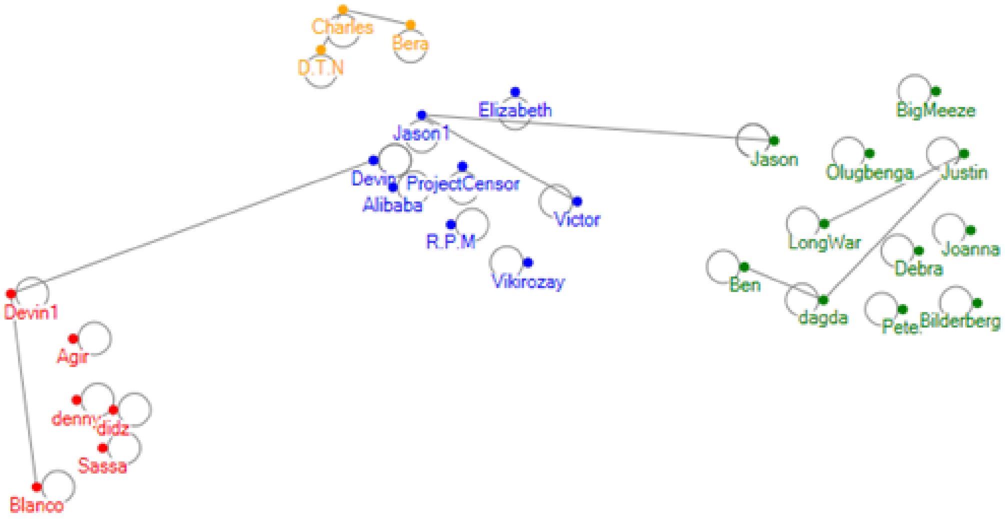


Fig. 14: Conversation Cluster for Fig. 12

Further, it can be observed that our analysis is focused at a specific time interval though it involves knowledge flow from previous time interval, but does not predicts the change in behavior of communities with time, thus we call our analysis a static one, but a dynamic outlook to such scenarios is highly desirable and will be addressed in near future.

E. Temporal Category Rank

By visualizing and analyzing the conversation clusters of Fig. 11 and Fig. 13, we can see that 'Alqaeda' is the most talked about topic in Fig. 11 whereas 'Country' specific talks rule the message burst for the next successive duration.

V. Conclusion and Future Research

In this paper, we use content-based Social Network Analysis (SNA), an approach to combine SNA and Text Mining (TM) in order to find out the hidden groups of users governed by Security topics of interest. Content-based SNA consists of extracting overlapping topic related sub-networks from the entire communication network.

We conducted an experiment on a live data sample collected from Twitter corresponding to hashtag 'Alqaeda' for a week. However, in this paper, we have presented our results of hourly analysis to gain useful insight into the information networks of these subgroups. We were able to successfully classify messages according to topics of conversation using the proposed approach and were also able to visualize those clusters along with identification of top topics of conversation at any specific instance of time. Currently, our results are limited to static network analysis over an interval of 60 minutes. However, in future we would like to conduct more extensive Social Network Analysis employing more of dynamic network analysis techniques to find out the change in topic influence with new users joining the network and identification of factors influencing change in user behavior with time.

References

[1] "Framework for Understanding Terrorist Use of the Internet". Technical Report, Canadian Centre for Intelligence and Security Studies, Trends in Terrorism Series, ITAC, Volume 2006-2.

[2] Declan Mc.. 'White House: need to monitor online 'extremism''. 2011. http://news.cnet.com/8301-31921_3-20087677-281/white-house-need-to-monitor-online-extremism/

[3] Muhammad A.S and Wang J. "Investigative Data Mining: Identifying Key Nodes in Terrorist Networks", 2006, IEEE.

[4] Nasrullah M. and Abdul Q.R. "Investigative data mining and its Application in Counterterrorism". In theProceedings of the 5th WSEAS International Conference on Applied Informatics and Communications, 2005.

[5] Chau M. and Xu. J. "Mining communities and their relationships in blogs: A study of online hate groups", In the International Journal of Human-Computer Studies 65 (2007) 57-70, Elsevier.

[6] Wadhwa P. and Bhatia M.P.S. "Tracking On-line Radicalization Using Investigative Data Mining", National Conference on Communications (NCC), pages 1-5, 2013, New Delhi, India.

[7] Roberts N.C. "Tracking and disrupting dark networks: Challenges of data collection and analysis", Information Systems Frontier (2011) 13:5-19. DOI 10.1007/s10796-010-9271-z.

[8] Wasserman S. & Faust K. Social Network Analysis: methods and applications. 1994, Cambridge University Press, Cambridge.

[9] Sparrow M.K. "The application of network analysis to criminal intelligence: An assessment of the prospects", Social Networks, Volume13, Issue 3, Pages 251-254, 1991, Elsevier.

[10] Xu J.J and Chen H. "CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery", ACM Transactions on Information Systems, Vol. 23, No.2, April 2005, Pages 201-226, Broadway, New York, NY 10036, USA.

[11] Xu J. and Chen H. "Criminal Network Analysis and Visualization", Communications of the ACM, Vol. 48, No.6, 2005.

[12] Yang C.C, Ng. T.D. "Terrorism and Crime Related Weblog Social Network: Link, Content Analysis and Information Visualization", In the proceedings of Intelligence and Security Informatics, pages 55-58, IEEE, 2007.

[13] Chen H., Chung W.,Xu J., Qin, Y.W.G and Chau M. "Crime Data Mining: A General Framework and Some Examples", IEEE Computer Society, 2004.

[14] Duo-Yong S., Shu-Quan G., Hai Z. and Ben-Xian L. "Study on Covert Networks of Terroristic Organizations Based on Text Analysis", 2011, IEEE.

[15] Duo-Yong S., Shu-Quan G., Ben-Xian L. and Xiao-Peng L.. "Study on Covert Networks of Terrorists Based on Interactive Relationship Hypothesis", 2011, IEEE.

[16] L'Huillier G., Alvarez H., Rios A.S. and Aguilera F. "Topic- Based Social Network Analysis for Virtual Communities of Interests in the Dark Web". In the proceedings of ISI-KDD 2010, Washington, D.C., USA.

[17] Gargi U., Lu W., Mirrokni V. and Yoon S. "Large-Scale Community Detection on YouTube for Topic Discovery and Exploration", Association for the Advancement of Artificial Intelligence (www.aaai.org), 2011.

[18] Klerks P., (2001), "The Network Paradigm Applied to Criminal Organisations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands", Connections 24(3): 53-65, 2001, INSNA.

[19] Hu D., Kaza S. and Chen H.. "Identifying significant facilitators of dark network evolution", Journal of the American Society for Information Science & Technology, Vol. 60, Issue 4, April 2009, Pages 655-665, John Wiley & Sons, Inc. New York, NY, USA.

[20] Barabasi A.L, Jeong H., Zeda Z., Ravasz E., Schubert A. and Vicsek T. "evolution of the Social Network of Scientific Collaborations", Physics A, pp 590-64,2002.

[21] Cogan P., Andrews M., Bradonjic M., Tucci G., Kennedy S. W. and Sala A. "Reconstruction and Analysis of Twitter Conversation Graphs", Proceedings of the first ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research (HotSocial'12), 2012, pages 25-31, ACM.

[22] Wadhwa P. and Bhatia M.P.S. "Analyzing Radicalization on Twitter", Second Student Research Symposium (SRS), International Conference on Advances in Computing, Communications and Informatics (ICACCI'13), 22 – 25 August 2013, Mysore, India in Press.

[23] Wadhwa P. and Bhatia M.P.S. "Classification of Radical Messages in Twitter using Security Associations". Accepted in Book on Case Studies in Secure/ Intelligent Computing- Achievement and Trends, To be Published by Taylor and Francis.

[24] Xu J., Marshall B., Kaza S. and Chen H. "Analyzing and Visualizing Criminal Network Dynamics: A Case Study", Intelligence and

Copyright © 2014 MECS

Security Informatics, lecture Notes in Computer Science Volume 3073, pp 359-377, 2004, Springer.

Authors' Profiles

Pooja Wadhwa received her B.Tech and M.Tech degrees from Guru Gobind Singh Indraprastha University in 2003 and 2005 respectively. She has been working with Government of India since 2005 as a Scientist in the area of Cyber Security. She is currently pursuing her P.h.D. in Computer Engineering Department at Netaji Subhas Institute of Technology, New Delhi, India. Her research interests include cyber security, data mining, social network analysis, malware analysis and computer networks.

M.P.S Bhatia received his Ph.D in Computer Science from University of Delhi. He has been working as a professor in the computer engineering department of Netaji Subhas Institute of Technology, New Delhi. He has guided many M.Tech and Ph.D students. His research interests include cyber security, data mining, semantic web, machine learning, software engineering and social network analysis.

How to cite this paper: Pooja Wadhwa, M.P.S Bhatia,"Discovering Hidden Networks in On-line Social Networks", International Journal of Intelligent Systems and Applications(IJISA), vol.6, no.5, pp.44-54, 2014. DOI: 10.5815/ijisa.2014.05.04