INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 1)
Available online at: www.ijariit.com

Detection of online spread of terrorism using web data mining

Naseema Begum A.
naseemabegum1122@gmail.com
KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu

Hanu Rakavi S.
hanurakavi20@gmail.com
KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu

Mohanambal R.
mohanakani346@gmail.com
KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu

Aswathy R. H.
rhaswathy@gmail.com
KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu

ABSTRACT

Terrorist growth has increased in certain parts of the world. Terrorist groups use Facebook, WhatsApp, messages to spread their information on the social network. It is essential to detect terrorism and prevent its spreading before a certain time. The basic idea of this project is to reduce or stop spreading of terrorism and to remove all these accounts. A terrorist is spreading their terrorism activities using the internet by speech, text, videos. Terrorist groups are utilizing the internet as a medium to convince innocent people to take part in terrorist activities by infuriating the people through web pages that inspire disenchanted individuals to take part in the terrorist organization. This needs a lot of human effort to implement this project that will collect the information and find the terrorist groups. To reduce the human effort, we implement the system which detects terrorist groups in social media.

Keywords— *Terrorism, World, Data mining, Clustering, Online*

1. INTRODUCTION

All terrorist activities are taking place through the web. Their infrastructure is based on the web for multiple purposes. The web data mining is used track the terror activities which is focused on tracking the abnormal contents transferring through online such as the sites generated by terrorist which may include examining the information that is utilized by the users of the web. This terrorist tracking system has two modes of operations: 1.the mode of training 2.the mode of detection. The training mode in the terrorist tracking system is used to determine the likes and interests of a typical terrorist group. It is performed by webpage processing utilized by the terrorist in the due course of time. The mode of detection in the terrorist tracking system is used to perform real-time analyzing the traffic on the web. It is obtained by the group monitoring the content that is available on the web. This, in turn, raises the alarm if the information utilized is related to the terrorist behaviour and it is not similar to the behaviour and interest of the ordinary user. The version of the terrorist tracking system experiments is implemented on the environment of the local network and it is evaluated. This ingenious knowledge-based technique is used for tracking the terrorist by analyzing the web traffic data as per the information was given by the audit. This project works by learning the basic behaviour and interest of the terrorists by using the web data mining algorithm to extract the textual content of terrorist websites. The typical behaviour of the terrorist is obtained by the precious analyses which are utilized to perform the real-time detection of terrorist from the group of normal users.

2. MINING STRUCTURE

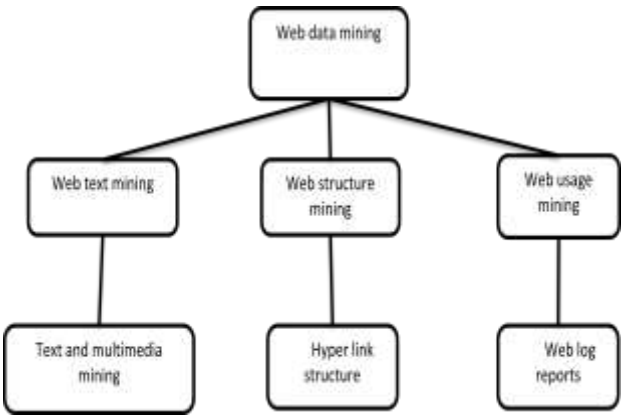


Fig. 1: Mining Structure

2.1 Web data mining

Mining of the web is similar to data mining and text mining but varies in certain aspects. This is similar to data mining since most of the data mining mechanisms are used in web data mining. This is also similar to text mining since most of the contents of the web are text. Moreover, this is not exactly data mining since web data mining includes all types of data like semi-structured, unstructured data, while the data mining

technique works only with the structured data primarily. The mining of web data is not similar to text mining since it has the nature of the semi-structured on the web. Mostly the text mining process aims at the unstructured text. The web data mining needs innovative applications that are to be used in data mining and also in the text mining methods that possess its own identical methodologies. In recent years, there is tremendous growth in the activities taking place in web data mining.

2.2 Web mining system structure

Web mining is also a type of Data Mining. This data mining methodology involves the extraction of the data with meaningful insights and also valuable content is to be collected from the huge volume of the data, Web mining technique involves the mining of the characteristics and typical behaviour of the web users through the web applications proposed. The information that had been extracted is to be used for a wide variety of the methods it includes there comes of the proposed web application, which may also determine fraudulent information and elements available etc. Web mining is moreover often determined as the part and parcel of the business intelligence system in very important organizations inferior to its technical aspects. This is solely utilised for determining the business strategies and taking various decisions by the efficient and also the effective use of web applications. The major purpose of this project is the system is very helpful for determining the terrorism-related activities. By using this project all the users of this project may be able to conclude the access by the suspicious users by checking their search list and conversation more over the system is able to detect the major source and initiators of the terrorism-related activities and the project is very helpful to minimize and even stop the terrorism-related activities. The project very said to possess a very wide range of scope in the national security and protecting the civilians of the nation. The project is said to be helpful to detect terror-related activities. The project has a wide range of scope among the Military and also the CBI officers. By utilising the project mentioned the counter-terrorism organisations may come to identify the plans of the terrorist and this takes the necessary steps to avoid the accident and spread of terrorism. Blogs, websites, social networking sites, and other sources of web used and hosted for carrying out such activities by the terrorist groups is to be avoided by the use of web mining.

3. CLEANING PROCESS

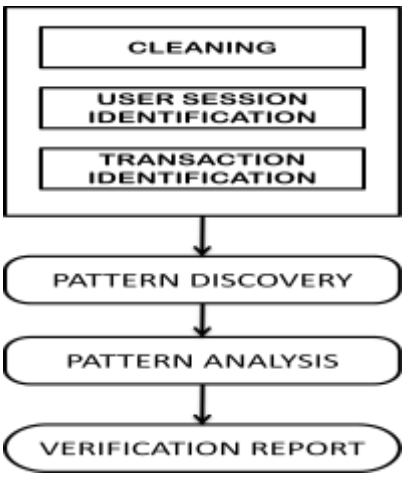


Fig. 2: Cleaning Process

3.1 Data Storage

The preprocessing results from the desired logs of the web server area to be stored and maintained in the relational database this in turns helps for the easy storage, retrieval and also the analysis. Preprocessing of the data: Preprocessing of the data converter is used to convert the raw data available according to the needs which will be helpful for the discovery of the pattern. The sole purpose of the preprocessing of the available data is to make the data quality better and there comes the need to increase the accuracy of the mining. Preprocessing task is consists of the following task suck as the extraction of the field, cleaning the processed data. This is the ideological phase is the most hectic, complex and also the ungrateful step involved in the overall processing process. The proposed system misty be used to describe the things in a brief manner and also determine that the task involved is to clean the raw data from the log files of the web and also insert the data processed into the relational database system this in turn orders to apply the appropriate data mining techniques seconds is applied in the second phase of the processing process. since the major steps of the phase included are: initially extract the logs from the web so that includes a collection of the data from the web server. Secondly, it includes the cleaning of the logs of web obtained and this, in turn, removes the redundant and also the repeated information. Finally, the data is parsed and the parsed data is to be given in a relational database system otherwise data warehouse system is used and also the data is reduced by frequency analysis in order to create reports in summary.

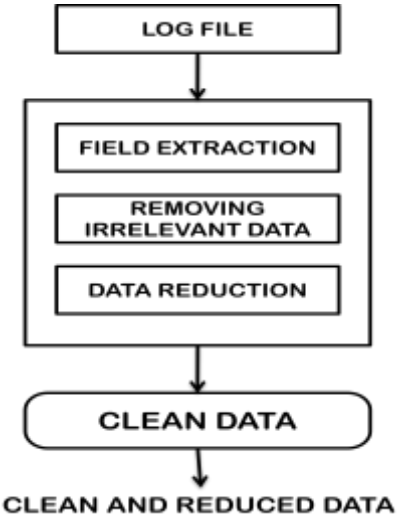


Fig. 3: Data storage

4. SCOPE OF RESEARCH WORK

The objective of this project is to find the typical behaviour of the terrorist in terms of their activities. The user can able to get knowledge about the distrustful conservation so that the user can able to detect the terrorism source. This will minimize the propagation of terrorist activities. This project will have a wide scope for national assets security. Terrorism activities can be easily found in this project. This methodology has the main scope in military purpose as well as for the Central Bureau of Investigation agents. By this project terrorist communication can be avoided. Terrorist provocation their activities mainly by blogs, social media by means of this project terrorist activities can be avoided.

5. LITERATURE SURVEY

The basis of study idea: The fastest growing area in the research is web data mining. It consists of web data mining, Web structure mining, and Web content mining. Web data mining is the process of discovering the users accessing pattern from the web data logs. The web structure mining is useful in discovering the knowledge from the structure of the hyperlinks.

Problem Statement: The proposed system is used for the patterns detection, detection of the keywords and detecting the relevant information relevant from the unstructured data available in the web page is used to mining of the web data and also the data mining [2]. The proposed system is used for the mining of the web pages by using the algorithm for web mining for mining the textual content in the available web pages from the internet and also detection of the web pages which are suspected to be in relevant with the terrorism. Web mining is used along with the data mining at most of the times for the accurate results.

5.1 Existing Solutions

Actually, there are none of the available systems is ready to look into the eye in the various available websites and also detecting the suspicious and the terrorist-related words that are available online [4]. The counter-terrorism organizations are not able to trace such terrorism-related websites and also the web pages or considering any of the user searching for the suspicious contents and also the terrorism-related terminologies. The terrorism is in the high peak and the ratio of terrorist activities are heavily growing in the world using the internet. Then there is a need to trace those systems and track all the websites for spreading of the terrorism using the internet and the ultimate aim of the system is to make the terrorism ratio very less and also stopping the online spread of the terrorism.

5.2 Clustering technique comparison:

Clustering is to be divided into two major subgroups broadly:

- **Hard clustering:** In this type of clustering, every data in the cluster should be in a cluster entirely or it should not be in that cluster. Example consider the individual customer in any of the 10 clusters.
- **Soft Clustering**: In this type of clustering, not every data is put in a separate in the cluster instead there comes a probability, any other consideration to be kept in a group or likelihood of the data to be assigned to the clusters. Example consider each of the customer to be assigned to either in any of the ten clusters available or the customer.
- **Connectivity models:** According to the name suggests, the concerned models of this type is based on the fact that the data that are close in the data space is said to express a high similarity to each other and not similar to data points outside of the cluster. The above-mentioned models are said to follow one of the two approaches. The first approach is starting firstly with the classification of all the data available into clusters separately & aggregation of the data points decreases the distance. Secondly the approach, the data points are said to be classified as the single and complete cluster as per the partition the distance among them is increased.
- **Centroid models:** This type of clustering is the iterative form of clustering algorithms which is based on the fact that the similarity of the data points is based on the centroid it is derived based on the closeness of the cluster. K-Means clustering algorithm is one among the popular clustering algorithms this k means algorithm comes under this category.
- **Distribution models:** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.
- **Density Models:** These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assigns the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS

5.3 Opinion definition and opinion summarization

Automatic summarization is the challenge due to the extensive information. The essential techniques that are needed for the recognizing of opinion are the opinion definition and the opinion summarization. Opinion definition of the data is done based on the text, sentence that are available as a document; this is also obtained from the complete document, summarization of the opinions are extensively important because of the need to analyse the normal behaviour of the user [17]. The aforementioned work is based on the intermediate sentences. Subsequent tracking requires the opinion mining of the individuals'.The opinion and the behaviour of the people are analysed through the text, comments and the search history available. After collecting the opinion and behaviour of the user there comes a necessary to summarize opinion and the typical behaviour since all the opinion and behaviour of the user is not expressed and described as a document. Opinion summarization is extremely important for the government to analyse the behaviour and the opinion of the people [16], [18]. This step helps in analysing the users to detect the terrorist from normal users. The opinion mining is important to classify the people into the normal, legitimate user and the terrorist groups are detected. There comes the need for the opinion mining of the people to determine the terrorist groups from normal users. Based on the opinion mining the characteristics of the normal user is detected from which the deviations are said to be terrorist.

5.4 Basic Clustering Techniques

K-means clustering and the hierarchical clustering that includes the agglomerative clustering analyzed used to analyze the text documents that are smaller in size analyze small text. The k means clustering method is based on the distance and the clustering is performed on basis of the mean value there occurred the continuous need of the iteration to determine the exact clustering and those deviates from this normal cluster are assumed to be a terrorist [13]. The strongest clusters obtained through the continuous iteration which is iterative in nature the continuous iteration is based on the value of the mean distance which is taken as random value initially. The values of the mean distance are recomputed regularly to determine the exact clusters with greater accuracy is to be ensured. Moreover, a lot of iterations are required for the stable cluster formation and then the output is made as final positive or negative.

5.5 Support Vector Machine

Support Vector Machine (SVM) is the clustering analysis technique. This technique is the supervised learning that is kernel based technique. Support vector machine is said to a research problem that is unsolved so far but it is the fast resolution technique. The SVM technique is extensively used in sentiment analysis. It is a popular technique because of its non-linear nature it is assumed to be simple for evaluation in aspects of both theoretically and computationally easy. SVM technique is the model the input-output relationship is efficient based on the output variable.SVM technique has much of the efficiency that is greater in the traditional level of categorization of the text this is compared with the other classification techniques such as the Naïve Bayes and also the Maximum Entropy [20].

5.6 Naïve Bayes

Naïve Bayesian classification algorithm is the conditional probabilities based algorithm. It is done by counting and

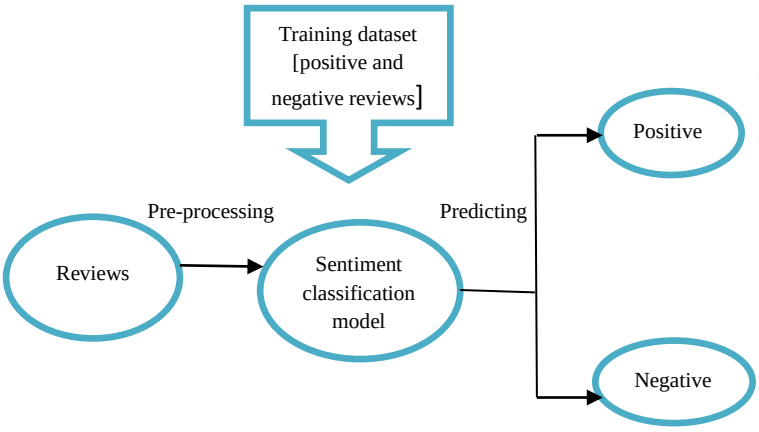
computing values based on the occurrence and the values are made into combinations based on the historical data. Therefore this classification is called the probabilistic method. Naïve Bayes is known to be the best-known algorithm based on probability values. Naïve Bayes is the supervised learning used for the accurate classification [14], [21], [19]. [20] Naïve Bayesian algorithm, as well as the binary keyword, are used together in order to produce a single dimensional degree analysis.

5.7 Neural Network

Neural Network technique is not similar to the support vector machine. It is a technique that is non-linear in its aspects this makes this not well known and popularly used the technique for data mining to use in the analysis. Associated set back of the non-linear techniques is considered as the following such as problem arises in the theoretical analysis and also problem that is associated with deciphering the text and the computation required (Zhang, 2001). Moreover is the technique that is kernel-based technique, the prediction as well as the classification strength associated non-linear techniques is made very much effective that is determined in consideration with the linear techniques.

5.8 K-Nearest Neighbor

It had been concluded based on the researches that have occurred till now but k nearest neighbour algorithm is the non-parametric method so that they were used in the regression and also the classification. The input this method consists of the k neighbour training data set to classify the terrorist and the typical normal behaviour and the behaviour of the terrorist is fed to the system and more over the output of the system is determined based on the input that has been given to the system and the output is based on the class membership property .if the object is more similar work the data point and if the value of the k is 1 then the object is assigned to the cluster if not kept under another cluster. This algorithm is not highly grasped the attention in many of the text analysis when compared to the support vector machine. The Naïve Bayesian classification and the specified Maximum Entropy classification are said to explored widely in many numbers of the experiments in analyzing the behavior of the pattern based on analyzing their sentiments and views such is performed by the above-mentioned algorithms. The typical behavior that includes their character analysis through their analyzing their so far liked and the disliked items and the typical behavior of the user is analyzed crucially. The performance of the clustering technique is based on the attribute selection.



5.9 Decision Tree

As the name implies the decision tee is the standard too used for the decision making. It is the tree-based model determining the various conditions and the resulted consequences are also explained. It is the flow chart like structure in which the internal nodes are considered as the test nodes and the child nodes are considered to be the resultant nodes. The classification from the internal node consider the main root node to the child node that is class labelled node is mentioned as the classification path and each of the internal nodes is taken as the test condition for the classification. The linear representation of the decision tree is the decision rule. But this method has certain difficulty as the determination of the test condition is very difficult. If the determination of the test condition goes in a bad way then the entire classification is collapsed. Hence determining the decision rule is difficult without analysing and class labelling a very huge number of the population.

6. METHODOLOGY

Web mining algorithms are used for mining the textual information that is available in the web pages after collecting the textual information the words relevant to the terrorism is detected. [12] Websites that have been created using different platforms, different algorithm and different programming languages are tracked. The proposed system is able to check whether the websites and the contents on the internet are promoting and spreading the activities related to terrorism and the terrorism-related propaganda is checked and then detected. The proposed system is used to detect and analyze the websites and also classify them accordingly as the terrorism-related and the normal legitimate users and sort them as the normal user or the terrorist. Data mining and web data mining are the two features that are to be used together for this detection process. Data mining technique is utilized to determine and define the pattern from the available collection of the websites and the data from the websites mined are the huge volume of data sources, the results obtained are widely used [4]. Web mining is also similar to data mining since it involves the text mining methods that are used to scan the data and also extracting the useful pattern from unstructured data. The proposed system is widely used by the government for anti-terrorism organizations. The proposed system aims to help such organizations for tracking the terrorist.

Websites must have characteristics that are mentioned:
Load Balancing: This is required by the websites since the admin can access the system by the available logs of the admin based on the load that is given to the server is to be limited and set control over it. This is done for the access of the admin.
Accessibility: For the easy accessibility of the records and store the other available information for easy access respectively.
The system must be user-friendly: The Website that is given to the public must be user-friendly for easy usage. The system must be efficient and also reliable. Easy maintenance of websites is essential for web data mining. These characteristics are said to be followed by websites for efficient mining of the websites.

7. APPLICATIONS

- Web mining algorithms are used for mining the textual information that is available in the web pages and also the relevancy of the websites to the terrorism is detected.
- Websites that are developed in the various platforms is also traced by the application without any difficulty. The proposed system will analyze the entire websites for the contents related to terrorism.
- The proposed system will categorize the websites as the normal website and the terrorism-related websites and also the users are classified as the normal user or the terrorist is determined.

- Data mining mythologies to use to derive a pattern from the websites and the available data. The pattern is used for the classification of the users and the websites.
- Web mining technologies are used along with the data mining methods to be utilized for the data in the websites to be and scanned for extracting the useful contents that are present in the structured, semi-structured as well as the from the unstructured data. Both the methods used simultaneously for the efficient detection of the terrorist.
- The application is used to block the terrorist once they are suspected and detected to be a terrorist.

8. CONCLUSION

The application that is to be developed will stop the spread of terrorism by preventing the radicalization of the people through online websites and the other media available on the internet. People are denied access to these websites and these media because the terrorist organizations are using the internet as the media to promote and spread terrorism and hence the terrorism is avoided.

9. REFERENCES

[1] Theint Theint Aye, "Web Log Cleaning for Mining of Web Usage Patterns" 2011 IEEE.

[2] Shaily Langhnoja, Mehul Barot, Darshak Mehta," Pre-Processing: Procedure on Web Log File for Web Usage Mining" International Journal of Emerging Technology and Advanced Engineering December 2012.

[3] L.K.JoshilaGrace, V.Maheswari, Dhinaharan Nagamalai," Analysis of Web Logs and Web User In Web Mining" International Journal of Network Security & Its Applications (IJNSA), Vol.12, No.1, January 2011.

[4] J.Vellingiri, S. Chenthur Pandian, "A Survey on Web Usage Mining" Volume 11 Issue 4 Version 1.0 March 2011.

[5] Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist group Web forum messages. IEEE Intelligent Systems, Special Issue on Artificial Intelligence for National and Homeland Security, 20(5),[67–75].

[6] Dietmar Jannach and Simon Fischer, Recommendation-based Modeling Support for Data Mining Processes, Germany and Simon

[7] Jiawei Han Micheline Kamber Jian Pei, Data Mining: Concepts and Techniques 12rd Edition(22nd June 2011)

[8] T. Sunil Kumar, Dr. K. Suvarchala, "A Study: Web Data Mining Challenges and Applications for Information Extraction",

[9] Syed Ahsan, Abad Shah," Data Mining, Semantic Web and Advanced Information Technologies for Fighting Terrorism"

[10] Robert Grossman, Simon Kasif, Reagan Moore, David Rocke, and Jeff Ullman," Data

[11] Mining Research: Opportunities and Challenges", A Report of three NSF Workshops on Mining Large, Massive, and Distributed Data, January 21, 1998 (Draft 8.4.5)

[12] Aggarwal, N., Liu, H.: Blogosphere: Research Issues, Tools, Applications. ACM SIGKDD Explorations. Vol. 10, issue 1, 20, 2008.

[13] Aggarwal, C.: An introduction to social network data analytics. Springer US, 2011

[14] Boiy, E., Moens, M.: A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. Information Retrieval, 12(5): 526-558, 2009.

[15] Chakrabarti, S.: Data Mining for Hypertext: A Tutorial Survey. ACM SIGKDD Explorations, 1(2):1-11. 2000.

[16] Dave, K., Lawrence, S., Pennock, D.: Mining the peanut gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: Proceedings of WWW 519-528, 20012.

[17] Ku, L.-W., Liang, Y.-T., Chen, H.-H.: Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In Proc. of the AAAI-CAAW'06, 2006.

[18] Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T.: Mining Product Reputations on the Web. ACM SIGKDD 2002, 1241-1249, 2002.

[19] Pang, B. and L. Lee, Vaithyanathan, S.: Thumbs up? Sentiment Classification Using Machine Learning Techniques. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, 79 - 86. Association for Computational Linguistics, 2002.

[20] Shi, H-X., and Li, X-J.: A Sentiment Analysis Model for Hotel Reviews Based on Supervised Learning. In: Proceedings of 2011 International Conference on Machine Learning and Cybernetics, Guilin, 2011.

[21] Tan, S., Zhang, J.: An Empirical Study of Sentiment Analysis for Chinese Documents. International J. Expert System with Applications, 124(4): 159-165, 2008.

