© 2018 IJSRSET | Volume 4 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Security Problem Detection of Hidden Data in Unstructured Log Messages With a Novel Text Mining Technique

Paridha Oza, Mr. Premkumar

Computer Engineering Department, Silver Oak College of Engineering & Technology, Ahmedabad, Gujarat, India

ABSTRACT

Text mining is an area where ample possibilites of research is being opened because of large data being shared everyday with the use of applicaitions and social media, both online and offline. Security here is a concern while passing textual infromation in such ways. Further more the logs generated with these data has all the important information in it. In text mining the wide area of text analysis contains machine learning which is further inherited from artificial intelligence. In text mining applications, information from Document is present in the form of Text along with Side Information or Metadata. XML documents found as RSS feed and other links generated for web pages have title of the document, author name or date of Publication which are present in the text document. Such metadata may possess a lot of information for the clustering purposes. Here there are possibilities of delievering unclear data. Using metadata for analysing information without filtering it, can result to lower security as well as data leakage. To improvise the results, we have used an efficient Feature Selection method to perform the mining process to select the metadata or xml tags which is useful for Clustering so as to maximize the advantages from using it. In my research, I have added process to find words that are not commonly being processed. I am expecting to get better results than the earlier researches by modifying the process.

Keywords : Security, Text Mining, Unstructured data, IARPA Dataset, clusters, Threat Detection

I. INTRODUCTION

With the development of information technology and the extensive application of network,the internet has become an essential part of people's life.Web pages and social network sites will generate large amounts of unstructured text data such as blogs,forum posts,technical documentation etc.These data contains a lot of information which is very difficult to deal with because of the alot and different forms.But the need of analyzing text data is rising.Therefore,how to develop the information,people need from large numbers of structured text data becomes the research point in the field of data mining and information[8].

Here in text mining,security is concern while generating textual information and passing these information in application and social media.

For experimented purposes,we selected the SKAION 2006 IARPA Dataset ,text mining techniques like clustering,feature extraction,classification and text mining preprocessing techniques like Tokenization,stemming,stop word removal etc.

The organization of this document is as follows. In Section 2 (**Methods and Material**), I'll give detail of any modifications to equipment or equipment constructed specifically for the study and, if pertinent, provide illustrations of the modifications. In Section 3 (**Result and Discussion**), present your research findings and your analysis of those findings. Discussed

in Section 4(**Conclusion**) a conclusion is the last part of something, its end or result.

II. METHODS AND MATERIAL

A. Literature Review

Candace Suh-Lee,Ju-YeonJo and Yoohwan Kim explained Text mining for security Threat detection.In this paper,the extracted features from log messages are used to run number of experiments on the packet clearing House SKAION 2006 IARPA Dataset[6,7]. and prediction capability is evaluated.These experiments are conducted using the extracted features only,both extracted features and message together[2].

Widodo Wahyu,Catur Wibowo explained improving classification performance by extending document terms.This paper proposes a method to improve the performance of text classification.In this study,they used TFIDF model,Hidden Markov Model,k-means clustering and Latent semantic Indexing for expanding documents.After expanding documents,documents are classified[9].

Jinju Job P.,Jyothi Korra proposed one technique which uses the auxiliary information that is present inside the text documents to improve the mining. This auxiliary information can be a description to the content,which can be either useful or completely useless for mining.In this paper,to mine the datasets,a combination of classical clustering algorithms is used.The propposed technique is aimed at improved results for document clustering[3].

Ramya Elizabeth Thomas,Shamsuddin S Khan proposed improved clustering technique using metadata for text mining. In many text mining applications, information from Document is present in the form of Text along with Side Information or Metadata. Such metadata may possess a lot of information for the clustering purposes. But this Side information may be sometimes noisy.In this paper,they make use of Co-Clustering with the help of TFIDF and Gini Index to elimination of noise,Cosine Similarity and K-means Algorithm for creating Clusters[4].

Garima,Hina Gulati and P.K.Singh explained comparison of different clustering techniques in Data mining.Clustering techniques in Data mining. Clustering plays an important role in the field of data mining due to the large amount of data sets.This paper reviews the various clustering algorithms available for data mining and provides a comparative analysis of the various clustering algorithms like DBSCAN, CLARA, CURE, CLARANS, Kmeans etc[5].

B. Methodology

Text Mining is the process seeking or extracting the useful information from the textual data.It is an exciting research area as it tries to discover knowledge from unstructured texts.

In this research to obtain more precise results we divided our work into the following stages.

1)Tokenization,Stop word removal,Stemming:- Text mining pre-processing stage starts with tokenization process.This method is used to breaking a stream of text up into words called tokens. second stage of pre-processing is stop word removal.In this process,removing of HTML,XML tags from web pages and the process of removal of stop words like “a”,”of” etc are performed.stop words are removed from documents because those words are not measured as kewords in text mining applications.Third stage of Pre-processing is stemming.This technique is used to find out the root or stem of a word.It is the process to converting the word to their stem or root.There are stemming algorithms like HMM,N-gram etc[8].

2)DBScan:-DBScan is Density based clustering algorithm.Clustering is a technique in which a given data set is divided into groups called clusters in such a manner that the data points that are similar lie together in one

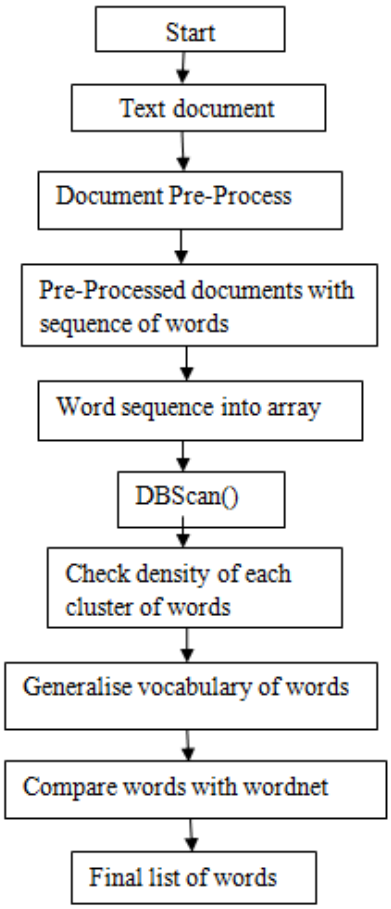
cluster.Density based clustering techniques are very useful in mining large datasets because they can easily identify noise and can deal with clusters of arbitrary shape[5].

III. RESULTS AND DISCUSSION

A.Proposed Algorithm

Here input is in text form.

First step is read the text document and preprocess it using text mining preprocessing steps.Preprocess the documents that are collected from social media or application.In preprocess stage,first step is tokenization,second step is stop word removal and third one is stemming.The output of this stage is sequence of words.



Fig(a)

In next step,put word sequence to an array for further work.Now this word array is input in DBScan algorithm.Here DBScan is used to make clusters of the words.DBScan will give cluster of words with density.

In next step,check density of each cluster of words and generalise vocabulary of words.In this step to generalise vocabulary we distribute words using density of each cluster of words.Now compare these words with wordnet dictionary to find meaning of each word.Now this is final list of words.

B. Performance Analysis

This section shows experiment results.We have collected Packet Clearing House SKAION 2006 IARPA Dataset.Testing is conducted using a novel algorithm.We measure the performance of algorithm by measuring its accuracy,precision and recall.

TP	FP
0.32	0.1
0.42	0.2
0.66	0.3

[a] Message only

TP	FP
0.24	0.1
0.58	0.2
0.75	0.3

[b] Feature only

TP	FP
0.52	0.1
0.68	0.2
0.77	0.3

[c] Both

The algorithm is implemented into a system which takes the input as the dataset and the output would be in three stages as we evaluate performance in three steps.First evaluate performance for message only.second evaluate performance for feature only and third evaluate for both feature and message .

IV. CONCLUSION

During this work, I analyzed the clustering of documents and implementation with the aim of

finding threats and by using the hybrind algorithm by adding db scan algorithm with text mining method.

For the security norms, the advantage is only the improvement in earlier result, where as in accuracy in results tend to increase. Algorithms are used to numeric and text documents. The system revealed me relevant results in my experiments. As I explained the implementation of hybrid algorithm for log data set of IARPA dataset. However, when it is implemented for much larger log documents, it will show the sufficient results. The comparative analysis of algorithms on basis of above given criteria demonstrates that the proposed algorithm yield a better results than existing one.

V. REFERENCES

[1]. P. J. Joby and J. Korra, "Accessing Accurate Documents by Mining Auxiliary Document Information," 2015 Second International Conference on Advances in Computing and Communication Engineering, Dehradun, 2015, pp. 634-638.

[2]. C. Suh-Lee, Ju-Yeon Jo and Yoohwan Kim, "Text mining for security threat detection discovering hidden information in unstructured log messages," 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, 2016, pp. 252-260.

[3]. P J. Joby and J. Korra, "Accessing Accurate Documents by Mining Auxiliary Document Information," 2015 Second International Conference on Advances in Computing and Communication Engineering, Dehradun, 2015, pp. 634-638.

[4]. R E. Thomas and S. S. Khan, "Improved clustering technique using metadata for text mining," 2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2016, pp. 1-5.

[5]. Garima, H. Gulati and P. K. Singh, "Clustering techniques in data mining: A comparison," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 410-415.

[6]. https://www.predict.org

[7]. Packet Clearing House, SKAION 2006 IARPA Dataset. http://pch.net.

[8]. Dr.S.Vijayarani et al , International Journal of Computer Science & Communication Networks,Vol 5(1),7-16.

[9]. Widodo and W. C. Wibowo, "Improving classification performance by extending documents terms," 2014 International Conference on Data and Software Engineering (ICODSE), Bandung, 2014, pp. 1-5.

[10]. N. Kanya, S. Geetha, "Information Extraction -a text mining approach" Information and Communication Technology in Electrical Sciences (ICTES 2007), 2007. ICTES. IET-UK International Conference .

[11]. Text Mining with R: A Tidy Approach by Julia singe, David Robinson.