ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 9, September 2014

Data Leakage Prevention by Using Word Gram Based Classification and Clustering

Rohan Vadsola[1], Dev Desai[2], Mihir Brahmbhatt[3], Alpesh Patanwadia[4]

Student, Information Technology, Sigma Institute of Technology, Vadodara, India[1]
Student, Information Technology, Sigma Institute of Technology, Vadodara, India[2]
Student, Information Technology, Sigma Institute of Technology, Vadodara, India[3]
Assistant Professor, Information Technology, Sigma Institute of Technology, Vadodara, India[4]

Abstract: Nowadays keeping one's data safe and secure has become very difficult task. The theft and misuse of the data has become a reason of concern ranging from small kids to adults, from small scale industries to multi-national industries. So to prevent this data leakage is extremely necessary. There were many techniques invented like firewalls, anti-viruses, watermarking etc. to prevent this leakage but eventually all tend to fail. Furthermore another techniques were invented known as the Data Leakage Prevention Software (DLPs) which keep the track of data while they are being transferred in the network. We have done research on one such technique known as the word gram. In this paper, the different uses and methods of using the word gram has been shown. This technique classifies the data packets before sending them in the network and makes clusters so while receiving the clusters can be compared with the sent clusters and we can be sure that the data hasn't been altered or interfered with during the transfer.

Keywords: Word Gram, Data Leakage and Prevention, Clustering, Frequency Matching, N-gram

I. INTRODUCTION

The leakage of the important data can be caused by an insider willingly or erroneously. The entity who causes the leak can be from the company itself or it may be some outsider. This leakage can cause great damage to an individual as well as an entire company. The data that has been leaked can be of an individual, a group of individuals or it can be the data of the company of their shares, assets, their financial information etc. So a small leak in the data can bring the company straight from top straight to the bottom. So it is necessary to stop such data leaks.

For instance there was a large data hack which occurred in the U.S. The hackers stole 160 million credit and debit card numbers and targeted 800,000 bank accounts [2]. This was considered as one of the largest hacking scheme posted in the U.S. This incident started in 2007, when a malware was inserted in the network of NASDQ's computer network which resulted in the theft of log-in credentials and it caused the leak.

The data can be possibly leaked at 3 instances. Firstly, when the data is at-rest (before the data is being is transmitted). Secondly, when the data is in-motion (when the data is being transmitted). And finally, after the data has reached the destination (after the data has been transmitted).

The DLPs now analyse the text to find the sensitive information. Word gram is one such DLP technique, on which we have done our research. It is used to analyse the documents and cluster the according to sensitivity. This is used widely now-a-days as the other classification and clustering DLP techniques has many limitations such as hash code generation. Meaning that a small change in the text can cause large amount of change in the hash codes. Which causes the data to leak without detection. The word gram technique eliminates this limitation to a great extent.

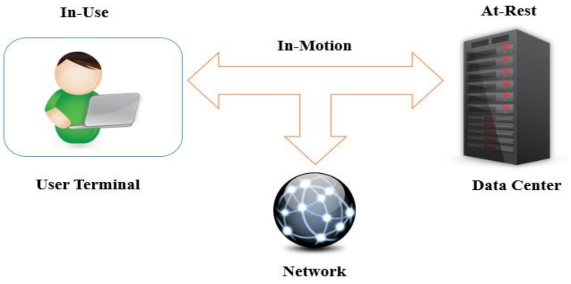

Fig. 1 Positioning of Data

So at the end we are try to accomplish following task in our research paper. Firstly, our research beings with the research of previous information on word gram. Secondly, we will find how the word gram eliminates the limitations regarding the hash codes. And last but not the least, we will try finding out the effectiveness of the word gram in data leakage prevention.

The research paper is further divided into following sections: Section II Related work, Section III Classification, Section IV Analysis, Section V Conclusion, and References.

II. RELATED WORK

Word gram was used in a variety of ways according to the studies. For example, as proposed by Cavnar, and Trenkle in [1] the entire document was broken into small characters known as n-grams. Then based on that a profile was created which was known as the n-gram profile. Then the profile that was created was compared with the existing n-gram profiles. Then the document is classified with smallest measure distance. Then all the "out of places" values are added for each n-gram and the distance was calculated.

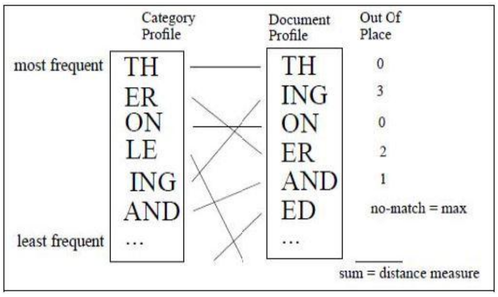

Fig. 2 Out Of Place Distance Calculation [1]

The method used here is based on the Zipf's Law or Zipf's Distribution [3]. The law suggests that the words are ranked are based on their occurrences, meaning higher the occurrence higher its rank. This law applies according to the language that is being used. This law was applied even for complex languages like Chinese, Arabic etc. and the words were classified in those languages based on this law.

III. CLASSIFICATION

In our practical, we include words which are used in N-gram sizes and original document's sizes. This process is been used for sensitive data testing. Some basic categories which are include in any type of topic that we want to protect. Mostly Antivirus are used to protect the system and documents against viruses and malwares. So we classify Antivirus and denote it as "A".

Now in our experiments we created the following categories:

Antivirus (A),
Firewall (F),
Data leak prevention (D),
Intrusion detection (I),
Virtual private network (V). [4] These categories come under security area. There are two things that should be taken into account. First, the N-gram documents with is specific to N-gram size. And Second, a classification level. The classification levels may include commonly used levels such as top-secret, secret, and confidential [5].

The practical starts by specifying the N-gram size according to the size of category profile. N-grams are shorted according to the frequency so as to start the distance measure phase. Then the smallest distance were tested against multiple categories of the documents N-gram profiles. At the end a category with the smallest distance was selected.

IV. ANALYSIS

To use the word grams other techniques should be learnt first, such as the Zipf's law. And word gram classification is often used to protect the data while in-motion. Word gram is usually used for small documents which contains mostly characters.

As shown in our paper there are many advantages of word gram over the previous traditional methods but with the advantages there are a few disadvantages that can cause trouble while using the N-gram technique.

As we all know that encryption is a basic technique used for data leakage prevention. But we can't use word gram and encryption techniques simultaneously as when we use the word gram it will not be able to detect the frequency of the encryption document as so it can't be compared with the word frequencies in the English language.

Also we shall include the documents with special characters and large documents in these limitations. Because it is not possible to calculate the frequency of the word gram of special characters, as a result they can't be ultimately compared with the word gram frequency of English language. And if we start analysing the long documents it will almost take forever to make the word grams of such long documents which adds it to the limitations.

V. CONCLUSION AND FUTURE WORKS

Word gram is almost inexpensive and it gives highly effective results when used for classifying the documents. In this paper we mentioned how the word gram is used. All the laws related to it and the techniques which were used to apply the word gram based classification. The correct number of classification is stated to be 85%, and there is a high chance that the data will be kept safe and secure using this technique. Also the advantages along with the limitations were stated in this paper.

We have studied in our research that since the results of modifying the DLP category profile were encouraging [4], an automated category profile update was proposed.

As the system doesn't include the special characters, in the future works, special characters may be added in the word grams. In addition to this we can give a predefined set of word grams which may be useful when we are using this technique with long documents.

ACKNOWLEDGMENT

We would like to extend our gratitude to our respected teacher – Mr. Alpesh Patanwadia, for his constant support and for imparting us with the knowledge and helping us throughout the paper.

REFERENCES

[1] W. B. Cavnar and J. M. Trenkle, "*N-gram-based text categorization*," Ann Arbor MI, vol. 48113, pp. 161-175, 1994.

[2] http://www.nydailynews.com/news/national/russians-ukrainian-charged-largest-hacking-spree-u-s-history-article-1.1408948

[3] G. K. Zipf, *Human behavior and the principle of least effort. Massachusetts*: Addison Wesley, 1949.

[4] Sultan Alneyadi, Elankayer Sithirasenan, Vallipuram Muthukkumarasamy" *Word N-gram Based Classification for Data Leakage Prevention*" Faculty of Science, Environment, Engineering and Technology Griffith University Gold Coast Campus, Australia.

[5] C. E. Landwehr, C. L. Heitmeyer, and J. McLean, "*A security model for military message systems*," ACM Transactions on Computer Systems (TOCS), vol. 2, pp. 198-222, 1984.