International Journal of Engineering &Technology, 7 (1.7) (2018) 149-151

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



A mechanism for identifying the guilt agent in a network using vector quantization and skew Gaussian distribution

S. Praveen Kumar[1*], Y. Srinivas[2], M. Vamsi Krishna[3]

[1]Research Scholar, Dept. Of CSE, Centurion University, Paralakhemundi, Orissa
[2]Dept. Of IT, GIT GITAM, Visakhapatnam, India
[3]Dept. Of CSE, Centurion University, Paralakhemundi, Orissa
[*]Corresponding author E-mail: spkmtech@gmail.com

Abstract

The recent state of art technologies has facilitated towards the ease of transfer of information from source to destination in a most comfortable and optimized medium. The channel transmitting discrepancies are rotted out by the sophisticated mechanisms that are being used currently using fiber optical cables etc… However sophisticated transmission mechanisms are available, the data transmission is always under threat due to intruder's hackers and guilt agents, who try to either steal the information, override the information or share the information to miscreants. Therefore techniques are to be developed to identify these guilt agents so that the data transmission can be transformed in a more secure way without any pitfalls. This paper addresses in this direction by proposing models based on vector quantization and statistical models.

Keywords: *Vector Quantization, Data Leakage, Intruder, Attacker, Transmission, Statistical Models.*

1. Introduction

Data leakages are one of the main concerned areas that take place during the transmission of vital information from source to destination. This data transfer is mostly of primary concerned in areas like military applications, forensic data and in some particular cases of medical data pertaining to an individual. To address these issues many models are developed in the literature and several methodologies and theories have been derived and deployed for safeguarding the information and transferring the information in a most secured manner. However, in most of the techniques, mathematical modeling like cryptography, visual cryptography, hash function based are utilized for ensuring the safety of the data during the transmission processes. Many sophisticated tools are therefore deployed and developed for this very purpose. At the same time, the lawbreakers ensure that effective breaking mechanisms are enforced to the systems using methodologies like keyloggers software,sending a virus through push messages and also trying to break the cryptographic algorithms using brute force technologies by creating a mess of parallel computing techniques and applying all possibilities and feasibilities to break the confidentiality. There are of late new technologies deployed for the purpose which are mostly based on the latent semantic analysis, N-Gram analysis (S.Praveen et al 2018) methodologies based on dimensionality reduction techniques using clustering techniques and machine learning techniques for annotating the text etc [1], [2], [3], [4], [5], [6].However, each of these techniques has their own limitations such as identifying the relevant text, correlating the text, interpreting the inherent meaning etc. Most of the technologies, therefore, are developed by associating a tag to every transmitted data and trying to identify the tag at the time of decryption. However, due to the technological developments, these tags can be breached or reconfigured [7], [8], [9]. Therefore, researchers came with strategies for associating a reference tag to each of these transmitting text and this mechanism also could not yield the expected results[10], [11]. Therefore to overrule these disadvantages one-time graphical password systems, captchas are also used as an interface. These methodologies also have also underlined some of them, therefore, to be customary to develop methodologies that can interpret and identify the guilt agent without breaking the associated tag.

With this objective, the present article is structured with vector quantization coupled with a statistical modeling approach based on Skew Gaussian distribution [12], [13]. The rest of the paper is oriented as follows, Section-2 of the paper deals with the vector quantization algorithm. The Section-3 deals with the probability density. The Section-4 of the dataset considered is presented together with an insight into a methodology. The Section-5 proposes the experimentation framework and the performance evaluation carried out by the model developed are highlighted in the corresponding Section-6. The final Section-7 summarizes the article.

2. Vector Quantization

One of the key features of Vector Quantization (VQ) is that it results into lossless compression technique and hence it reduces the dimensionality of the data by retaining the core significant features, because of this ability, it is widely used in many applications ranging speech signals to voice recognition, data reduction etc.

It is a process of clustering the data into partitions called chunks. Each of the chunks is called a cluster. To each of the cluster; the cluster centroids are estimated, termed as a codeword. All these code words are grouped into an array or book, named codebook. These processes of maintaining the codebook were first developed by Linda Buzo Grey (LBG) and hence this vector quantization is also called as LBG quantization. In this methodology, the chunked data is considered and the appropriate code word is identified. Generally, these code words can be the adjectives of relevant words and

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

all these adjectives of a given string are polled together thereby formulating a codebook.

2.1 LBG Algorithm

The Linde, Buzo and Gray, algorithm or VQ algorithm is presented in the following algorithm.

1. Construct a 1-D codebook by considering the code words, for the entire dataset.
2. Enhance the codebook size to by isolating each fresh code word Xn, using the formula Xn+ = Xn (1+α) and Yn+ = Yn (1-α), 1 <n<to size of the book
3. Find the centroid against each of the codebooks
4. Reiterate steps 2 & 3 for each and every fresh code word
5. Enhance the centroids for every newly formulated chunk. Using the formula $X_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$
   i represents the factor of each vector (x, y, z, ... directions), m is the number of vectors in the cluster.
6. Repeat steps 2 and 3 until no new codewords are formulated.
7. Calculate the number of codewords, N.
8. The major codebook is formulated by choosing the N codewords at random.
9. Using the Euclidean distance measure, to estimate the distance between the code words.

3. Skew Gaussian distribution

The main purpose of Skew Gaussian distribution is that it helps to identify the data which is both symmetric and asymmetric in nature i.e. it can identify the data which is homogeneous and non-homogeneous along with the capability to interpret the data of varying length i.e. having long duration samples or small duration samples in the present context the code words generated will be interpreted and these code words are given as input these model to formulate the PDF.

The general form of the Skew Gaussian Distribution is given by a formula.

$$f(z) = 2.\emptyset(z).\Phi(\propto z); -\infty < z < \infty \quad (1)$$

Where, $\Phi(\propto z) = \int_{-\infty}^{\propto z} \emptyset(t)dt.$

And, $$\emptyset(z) = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} \quad (2)$$

Let, $$y = \mu + \sigma z \quad (3)$$

$$z = \frac{y-\mu}{\sigma} \quad (4)$$

Substituting the equations 2,3,4 in 1, we have

$$f(z) = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} \left[\int_{-\infty}^{\alpha\left(\frac{y-\mu}{\sigma}\right)} \frac{e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}} dt\right]$$

4. Dataset

In order to propose the correct methodology, we have considered a data set of emails from ENRON dataset of emails. These emails were from the organization, ENRON having 425 employees. These emails consist of both official and personal emails there are around 11000 emails in this data set. These emails are considered for the purpose of experimentation

The sample ENRON dataset is given in the following Figure1

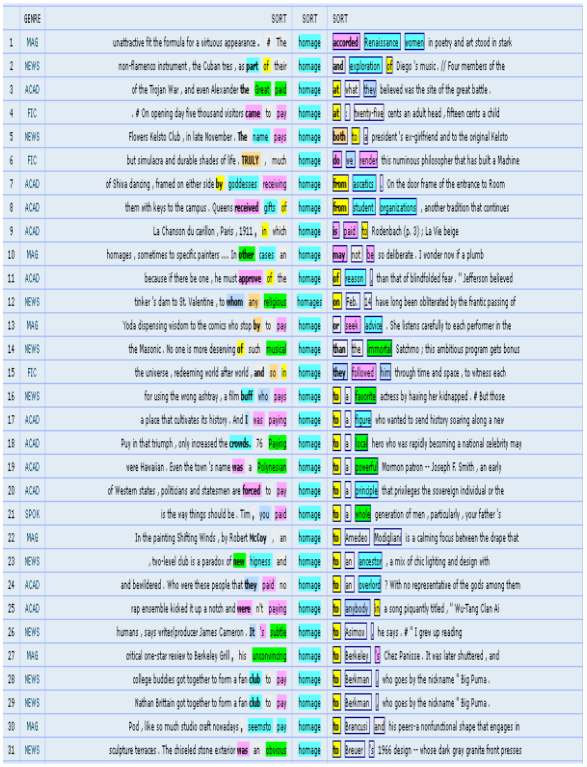


Fig: 1 ENRON Dataset

5. Experimentation

The experimentation is carried out with above data set as mentioned in Section-4 of the article and the input of this data is clustered to identify the code words and the code words are given as input to the Skew Gaussian distribution model mentioned in Section-3 of the article. Against each of these codewords, the corresponding probability obtained is noted and the frequencies of each of the contents of the emails are identified.

While transferring the data from source to destination in order to ensure the security and overcome attacks, the code words of the employees who are pipelined for receiving are scrutinized and a relative codebook is generated. Upon the authenticity, the data with code words will be sent to the respective emails where the inverse of PDF is considered to have the appropriate text. This methodology ensures total confidentiality of the total confidentiality rate. The architecture of the proposed model is presented below in fig:

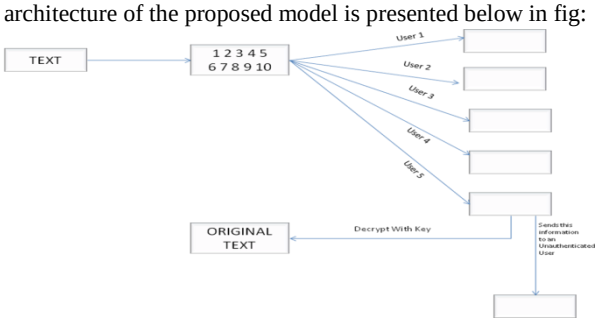


Fig: 2 Architecture Model

Here each of the emails of the authenticated users is employees within a group to whom the data needs to be transferred are collected. Each of these users is shared with a key and the data is transferred across by encrypting the text and ensuring that the authenticated users have the flexibility of decrypting. During this process, all the anonymizer attacks are taken care since the text is shared only with those authenticated users. Each of the data being transmitted also consists of an embedded key variant for a group of codewords of the similar adjective are considered and it is stored as watermark and the adjectives are based on the location, such that each

location has a designated code word. All these code words are maintained in a codebook and the administrator manages the total transmission process. In between if there is any data leakage from these intended users, the key that is decrypted will be changed automatically such that the decryption process becomes next to impossible. In order to deactivate the text, each of the data in the text is converted into PDF was the ASCII of each word is considered for the purpose. Each of these data in the numeric form is then subjected to the skew distribution underlined in Section-3 of the article such that the PDF is generated.

In order to retrieve the original text, these PDF should be inverted for which KL divergence is KL transmissions(kauhn loullie transformation) is used, and these process of generating the inverse PDF is taken care by the administrator after confirming the email address. However, there is every possibility that a particular user may share his email address and password to the third party and in order to curtail such activities the secret key attached to each of these emails gets tampered and helping the administrator to identify the guilt agent.

6. Performance Evaluation

In order to propose the efficiency of the methodology, a performance evaluation is carried out by using the metrics like miss-classification rate, false acceptance rate (FAR) and false rejection rate(FRR) the formulas for calculating each of these metrics are given by

MCR = ((Total number of Unauthenticated received users) / (Total number of users)) * 100

FAR = ((Total number of miscreants for whom the data is transferred) / (Total number of genuine users))*100

FRR= ((Total number of Unauthorized users for whom the data is transferred by the administrator)/ (Total no of intended users))*100.The data is evaluated and is presented in the following table1 of the article.

Table1: Data Evaluated

Metric	Efficiency	Total No. Of users
Miss Classification Rate(MCR)	76%	175
False Acceptance Rate(FAR)	82%	222
False Rejection Rate(FRR)	12%	200

7. Conclusion

In this article, a new proposal for identifying the guilt agent is proposed by considering a statistical mixture model together with the concepts of vector quantization. A skew Gaussian mixture model is considered because of the fact that it can interpret the data transferred across in a distributed environment, both in symmetric and asymmetric nature. This Skew Gaussian mixture models helps to even handle the data with long tails and hence can construct different patterns of text communicated across. The vector quantization ensures dimensionality reduction and clustering the data into groups without loss of dimensionality. This proposal helps to have a more précised precision accuracy rate in identifying the guilt agents.

References

[1] Poomagal, Shanmugam, Palanisamy Visalakshi, and Thiagarajan Hamsapriya.

[2] "A novel method for clustering tweets in Twitter." International Journal of Web Based Communities 11.2 (2015):170-187.http://dx.doi.org/10.1504/IJWBC.2015.068540

[3] Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34.1 (2002): 1-47.

[4] Khorsheed, Mohammad S., and Abdulmohsen O. Al-Thubaity. "Comparative evaluation of text classification techniques using a large diverse Arabic dataset." Language resources and evaluation 47.2 (2013):513-538. http://dx.doi.org/10.1007/s10579-013-9221-8

[5] Ataa Allah, Fadoua. "Information retrieval: applications to English and Arabic documents." (2008).

[6] Ghwanmeh, Sameh H. "Applying Clustering of hierarchical K-means like Algorithm on Arabic Language." International Journal of Information Technology3.3 (2005).

[7] Beil, Florian, Martin Ester, and Xiaowei Xu. "Frequent term-based text Clustering." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002. http://dx.doi.org/10.1145/775047.775110

[8] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze.Introduction to information retrieval. Vol.1.Cambridge: Cambridge university press, 2008 http://dx.doi.org/10.1017/CBO9780511809071

[9] Aljaber, B., Stokes, N., Bailey, J., & Pei, J. (2010). Document clustering of scientific texts using citation contexts. Information Retrieval, 13(2), 101-131. http://dx.doi.org/10.1007/s10791-009-9108-x

[10] Andrews, Nicholas O., and Edward A. Fox. "Recent developments in document clustering." (2007).

[11] Carpineto, Claudio, et al. "A survey of web clustering engines." ACM Computing Surveys (CSUR) 41.3 (2009): 17.

[12] Mubarak, Hamdy, and Kareem Darwish. "Using Twitter to collect a Multi-dialectal corpus of Arabic." ANLP 2014 (2014): 1.http://dx.doi.org/10.3115/v1/w14-3601

[13] Moh"d Mesleh, Abdelwadood. "Feature sub-set selection metrics for Arabic text classification." Pattern Recognition Letters 32.14 (2011):1922-1929. http://dx.doi.org/10.1016/j.patrec.2011.07.010