TEST Engineering & Management

January - February 2020
ISSN: 0193-4120 Page No. 8479 – 8485

An N-Gram Analysis Approach for Sharing of Authentication of Data Using Model Based Techniques

[1]S.Praveen Kumar, [2]Dr.Y Srinivas, [3]K.Bhargav
[123]*Dept. Of CSE, GIT GITAM, Visakhapatnam,India*

Article Info
Volume 82
Page Number: 8479 – 8485
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 07 February 2020*

Abstract:
A data breach is the deliberate or involuntary disclosure of private information to illicit parties. In the present digital world, data has become one of the most vital components of an enterprise. Data leakage poses severe intimidation to organizations, including significant reputational harm and monetary losses. As the magnitude of data is mounting exponentially proportionately the frequency of data breaches have increased drastically. Therefore, it is necessary to propose methods which help in effectual detection of the breaches and thereby assisting the enterprises to overcome data losses. In spite a plethora of research efforts on protecting perceptive information from being leaked, it still remains as a dynamic research problem. This article address this issue by proposing a model based on N-Gram analysis together with statistical mixture model.

Keywords:*N-Gram Analysis, Secured, retrieved, Enormous Data, Model based Technique.*

1. Introduction

Textual data is omnipresent and plays a key role in information broadcasting. To assist distribution of Textual information, traditionally, analogous documentation discovery techniques are implemented in many application domains. Considerable work has been reported in the literature to uncover similar documents [1] [2][3][4,[5]. Most of the literature is driven in this direction by assuming that bulk of the information in the files is public. However, in certain applications like individual's medical data, it is necessary to cover most of the personnel information; therefore the traditional ways of similar document verification may not be applicable in these situations. Similarly, in cases of medical insurance policies, this approach is not suitable. There are many such instances, where it is deemed to consider privacy preserving models to uncover identical documents. Privacy Preserving Models (or) Secure Similar Document Detection (SSDD) models were mostly useful in cases of sharing the correlated criminal data across the investigation agencies. The data inside these investigative reports are formulated in the form of vectors and every term entered into the vector table contains the frequency of occurrence of each word. To identify the relative similarities among such documents, metrics like Cosine Similarity are used. However, these metrics fail to take into consideration the frequency of overlapping terms. To overcome this disadvantage, N-Gram analysis plays a vital role.

Another feature associated with N-Gram models are, they are language independent and possess the ability to model sensitive data. The objective of this article is to propose the usage of N-Gram Model to SSDD application. Here the data is chunked substrings, n-grams of similar size. If the data is clustered into a substring containing a character, it is termed as uni-gram approach, and subsequently, the other representations can be bi-gram, tri-gram or in general N-gram. Each of these blocks are taken into

January - February 2020
ISSN: 0193-4120 Page No. 8479 – 8485

a table and the similarities between these vectors are compared using metrics like cosine similarity, Jacquard similarity. Among these models most of the models are based on identifying the unstructured data and there by deleting (or) minimizing, the repeated data by formulating methodologies such as identifying the raw data in different formats, using methodologies based on query optimization and thereby helping out to convert the unstructured into structured data. Clustering techniques are also adopted to convert the unstructured into structured and thereby using classification techniques to retrieve required information.

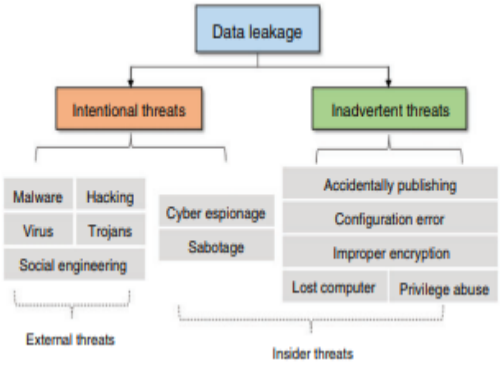
Therefore many models based on both generating and degenerating approaches have been evolved to cluster the data and thereby identifying the relative item set from the classified data. Oflate utility data sets, frequent item set pattern mining datasets, Apriori algorithms are therefore considered for this purpose. Among the literature, generative algorithm are assumed to be more significant compared to degenerative models (K.Naveen Kumar et al.),(G.V.S Rajkumar et al),(P.Chandrasekhar et al.)This is because of the very reason that the generative approaches try to estimate the inherit behavior of the data and thereby cluster the data into appropriate pools. Therefore these techniques yield maximized results in most of the cases when compared to degenerative approaches like apriori, Bayesian classification K-Means etc[6],[7],[8],[9].

Recently (S.Praveen Kumar et al.) has proposed a model based approach for identification of the frequent item sets based on latent semantic analysis. However most of the literature in this area is confined only towards the identification of a frequent item and no efforts have been made to utilize these methodologies for transfer of information globally. Hence with these intuitions, the present article makes an attempt to bridge the gap in this direction. Therefore in this article we proposed a methodology based on N-Gram analysis for text categorization and transferring of these categorization data across the globe by indulging model based approach ,using generalized Gaussian mixture Model. The rest of the article is presented as below. In section-2 of the paper, a brief review of N-gram analysis is presented. In Section-3 Gaussian Mixture model is presented. The dataset considered is highlighted in Section-4, the methodology is highlighted in Section-5. Section-6 of the paper portrays the experimentation together with the results derived and the final Section-7, summarizes the article.

1.1Classification of Data Leak Threats

Data Leakage threats can be classified basing on the type of threat and the necessity. Generally, the leakages of data may be done deliberately or accidentally. Most of the leakages within the enterprises are caused either by insider or outsider.

Deliberate leaks are mainly due to the participation of outside parties or malicious insiders. Outside data breaches are usually due to hacker break-ins, malware, virus, and social engineering. Phishing attacks become more and more complicated aligned with enterprises, by fooling workforce and passing the rich information source to the lawbreakers. In-house data leakage can be caused by either purposeful actions or involuntarily mistakes. The various types of data leakages are presented in the following Figure-1



2. N-Gram Analysis:

In the literature, to categorize the text several methodologies are presented. The foremost among them are frequent item set based approach. However there are certain limitations with these approaches such its inability to identify and mine the items having unstructedness and another disadvantage of

January - February 2020
ISSN: 0193-4120 Page No. 8479 – 8485

these methodologies is that it cannot minimize the unwanted/unused text which may be insignificant in same particular cases. Therefore to overrule these limitations utility based approaches have been considered. Here the utility of each item is considered and then the most frequent item sets is considered and then the most frequent item sets are generated. This approach generates reduction in documentation size. Nevertheless this methodology fails in case of text having minimal errors such as spelling mistakes, grammatical errors etc. to override these disadvantages, in this article we have considered N-Gram Analysis.

The primary Objective of any text categorization is to process the document and ensuring its ability to transform into electronic form. Most of the text contains grammatical errors. This methodology of N-Gram analysis[13] is considered because of its ability to override this disadvantage. One of the most fundamental issues in any document processing is its ease to handle massive data in electronic form. If the data consists of any typographic errors, special characters, white spaces then it is difficult to handle such data through other techniques like frequent pattern mining. Hence to overrule this disadvantage, the text data is processed such that the special characters, white spaces or removed. Therefore in this article we have considered this approach and as an experimental study we have implemented the concepts of N-Gram analysis, to ensure whether a particular email is sent by an authenticated user or not. The main advantage of N-Gram analysis is that it can identify the errors in the data most appropriately, with this approach the concepts of N-Gram are considered, the experimentation is conducted by considering the email corpus, ENRON.Also few other techniques can be considered like Deep learning for hate speech detection,New Malicious Code Detection Using Variable Length n-Grams Algorithms, Architectures and Information Systems Security, Hate speech detection, Detecting hate speech on the world wide web [10],[11],[12],[14],[15].

3. Gaussian Mixture Model (GMM)

The Probability density function of the Model is presented below.

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\left((z-\mu)/\sigma\right)^2} \text{ --- (1)}$$

The main advantage of the Gaussian Mixture Model is that it covers quite a few distributions as its specific cases and therefore it facilitates to identify different ranges of semantic values thereby helping out in identifying the appropriate links more aptly. It assists to classify the link analysis more robustly.

3.1.Dataset:

To present the proposed methodology, we have considered the dataset of emails from ENRON. This data set contains over 50,000 emails generated from the working employees, ranging across different tags, linguistic, featured, crime etc.

4.Methodology:

In order to present the methodology the architecture presented below in figure 1 is considered.

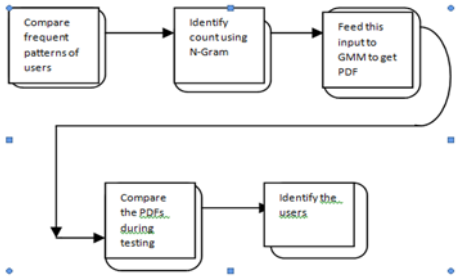


Figure 1- Architecture of the system

The general working of the N-Gram analysis, in particular bi-gram analysis, is explained by using the example below. Here are the 3 statements about a user generally used while writing a email, and want to ascertain whether the sentence of the email containing “I am Not” is a part of his emails or not [15],[16].

<S> I am Praveen <S>
<S> I am not Naveen <S>
<S> I Live at Simhachalam<S>

TEST Engineering & Management

January - February 2020
ISSN: 0193-4120 Page No. 8479 – 8485

This can be possible by counting the probability, using the formula

$P(W_i / W_{i-1}) = Count(W_{i-1}, W_i) / Count(W_{i-1})$

and by using the above formula, the probability using the bigram for the string "I am not" is calculated as follows

P(I/<S>) *P(I/I) * P(am/I) * P(not/am)

= 3/ 3 * ¼* 2/4 *1/2 = 0.0625

Since the probability is very low, the chance of the word "I am not" is not from the same user. In this approach, each of the sentences are scanned against a test sentence and the probabilities are identified.

As an extension to this approach, in our approach, we have identified the probabilities of each of the word , and calculate a frequency mapping table, by removing the special characters and . These probabilities are given to the Model based on GMM, to calculate the PDF. To test whether the email is from a particular user, the PDF's are compared using KL divergence, presented in section 5.1 of the article, and the relevancy is estimated. The predictive features andidentification of a frequent items can also be considered [17],[18].

4.1 KL-Divergence

This method is used to test the relevance between two PDF, the formula for calculating the same is given by

$$KL(p,q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

--- (3)

Where, μ_1, σ_1 and μ_2, σ_2 denote the mean andvariance of the trained and tested PDF's. If these probabilities are → 0, then there are very much identical.

5. Experimentation

The same methodology is tested by considering a small email from the dataset, and the results are presented in the following figure-2

Figure-2: Frequency count of an email

6.Performance Evaluation Metrics

In order to test the accuracy of the model, we have considered the metrics like precision and Recall, the formulas for calculation of the above are given below

a. Precision

It is the ratio of the number of relevant tokens to the total number of irrelevant and relevant tokens retrieved. It is typically articulated as a percentage.

Precision = (A / (A + C))* 100;

A: Number of relevant tokens retrieved.

C: Number of irrelevant tokens retrieved.

A + C: Total number of irrelevant tokens + relevant tokens identified

b. Recall

It is the ratio of the number of relevant tokens to the total number of relevant tokens in the database. It is usually expressed as a percentage.

Recall = (A / (A + B)) * 100

A: Number of relevant tokens retrieved

B: Number of relevant tokens not retrieved

A + B: The total number of relevant tokens

January - February 2020
ISSN: 0193-4120 Page No. 8479 – 8485

c. Error rate

Error rate = Number of non-relevant tokens / Total number of tokens

d. Retrieval efficiency

Retrieval efficiency=Precision=Number of relevant tokens retrieved/ Total number of tokens retrieved (If number of retrieved tokens > number of relevant tokens), otherwise

Retrieval efficiency= Number of relevant tokens retrieved/ Total number of relevant tokens.

The results derived using the proposed methodology on the considered data set is presented in the following Table-1.

No. of relevant Tokens in the database (R)	No. of non-relevant Tokens against the relevant tokens (NR)	Ratio of relevance to non-relevance (R/NR)	Precision	Recall	Error rate	Retrieval efficiency
80	15	0.18	83	6	08	84
77	13	0.16	79	29	16	94
60	29	0.33	62	58	18	72
50	22	0.088	53	54	27	73
25	14	0.56	42	63	18	80
35	26	0.74	35	72	15	76
28	20	0.71	27	77	14	75
15	14	0.93	16	89	14	70

Table 1: Variation of Recall value with Relative to Non-Relative Ratio

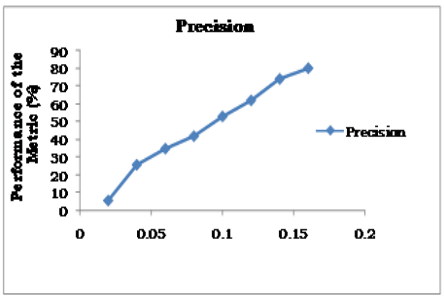


Figure-3 Variation of Precision value with Relative to Non-Relative Ratio

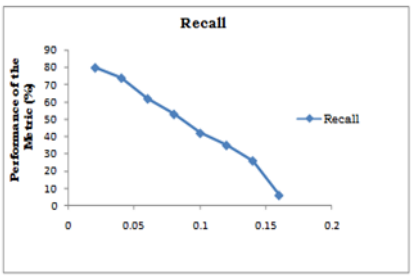


Figure-4 Variation of Recall value with Relative to Non-Relative Ratio

6.1 N-Gram Analysis

To investigate the n-gram output, we have considered the term frequency and TFIDF [5], which are defined as follows:

Term frequency, TF is the numerical measurement of how frequently a term take place with respect to other terms in the document. Term frequency of a specific term can be obtained by dividing its incidence by the frequency of maximum occurring term in the document, mathematically; it is evaluated by the formula

$$f = f(t,d)/ \max[f(t,d):w \in d]$$

Term Frequency–Inverse Document Frequency, TFIDF, is a numerical measurement that imitates the importance of a specific word in the collected pool of documents. TFIDF is calculated by multiplying TF wih IDF, and is given by

$$TFIDF = (t,d,D) = TF \times IDF$$

Where IDF denotes the inverse document frequency and the mathematical formula for the calculation is given by

$$IDF = Log2 (N/DF)$$

N denotes the number of documents and DF specifies the count number of files in which the term appears. We have collected 800 malicious documents for the mission and similar number of genuine documents. The experimentation is carried

January - February 2020
ISSN: 0193-4120 Page No. 8479 – 8485

out several times, however, the time consumed in case of larger programs is very large, therefore to handle large files, a threshold term frequency of 0.05 is set and all the terms having frequency of occurrence less than the threshold are not considered. We analyzed this data using machine learning using the formula

tuple = {tf,tfidf,mal0gen}

where mal0gen is term, which takes the binary values, 0 or 1. The value is set to 0 in case of genuine documents and is set to 1 for malicious document. The experimentation is carried out by using J48 algorithm using WEKA Tool [6] [7]. The results generated are presented in the following Table 2

	Classified correctly	Classified wrong
Genuine	41212	453
Malicious	53242	139

Table 2 Results with J48 Decision Tree

These results indicate that a model built with TF and TFIDF of PDF documents with the help of J48algorithm is very efficient with 0.01 percent false positives and 0.0039 percent false negative rate.

7. Conclusion

The experimentation is also carried out using the ratio of Relevant to Non relevant Tokens. The results derived are presented in Table -1 and figures 3 and 4. From the above table and figures, it can be seen that the accuracy of recognition of the model is minimum 70%.

The performance accuracy is carried out using various metrics like Precision, Recall, Error Rate and Efficiency and it can be clearly seen that the developed model is helping to identify the relevant tokens more accurately with a Retrieval efficiency of 89% and 94%. The Error rate is also very negligible which 8% to 14% is.

REFERENCES

1. T.V. Madhusudhan et al (2015) Model Based Approach for Content Based ImageRetrievals Based on Fusion and Relevancy Methodology, International Arab Journal of Information Technology, 12(6), pp 519-523
2. Qian F., Li M., Zhang L., Zhang H., and ZhangB., "Gaussian Mixture Model for Relevance Feedback in Image Retrieval," in Proceedings of IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, pp. 229-232, 2002.
3. K.Naveen Kumar et al (2016) Studies On Improving Texture Segmentation Performance Using Generalized Gaussian Mixture Model Integrating DCT and LBP,Journal of Therotical and Applied Information Technology, 92(2), pp 200-2007
4. Naveen Kumar. K, SrinivasaRao.K, Srinivas.Y and Satyanarayana. Ch(2015), "Texture Segmentation using multivariate generalized Gaussian mixture model under log DCT domain", International Journal of Applied Engineering Research, Vol.10(22), pp.43045-43051.
5. NageshVadaparthi, SrinivasYerramalle, and Suresh Varma.P: Segmentation of Brain MR Images based on Finite Skew Gaussian Mixture Model with Fuzzy C-Means Clustering and EM Algorithm", International Journal of Computer Applications, 28(10):18-26, August 2011.
6. G.V.S.Rajkumar et al(2011, Studies on Colour Image Segmentation Method Based on Finite Left Truncated Bivariate Gaussian Mixture Model with K-Means, Global Journal of Computer Science and Technology 11(18),pp 21-30
7. P.ChandraSekhar et al (2014), Image Segmentation for Animal Images using Finite Mixture of Pearson type VI Distribution, Global Journal of Computer Science and Technology: F Graphics & Vision
8. Asaf Shabtai1, Detecting unknown malicious code by applying classification techniques on OpCode patterns, Security Informatics, a Spinger Open Journal, 2012, 1(1), 1-22.
9. Georgios K. Pitsilis et al (2018), Detecting Offensive Language in Tweets Using Deep

January - February 2020
ISSN: 0193-4120 Page No. 8479 – 8485

Learning, arXiv:1801.04433v1 [cs.CL] 13 Jan 2018

10. PinkeshBadjatiya, Shashank Gupta, Manish Gupta, and VasudevaVarma. Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion, pages 759–760, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee. ISBN978-1-4503-4914-7.doi:10.1145/3041021.3054223 .
11. P. Barnaghi, P. Ghaffari, and J. G. Breslin. Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. In 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), pages 52–57, March 2016. doi: 12.1109/BigDataService.2016.36. BBC. Facebook, Google and Twitter agree german hate speech deal. Website, 2016. http://www.bbc.com/news/world-europe-35105003 Accessed: on 26/11/2016.
12. Ying Chen, Yilu Zhou, Sencun Zhu, and HengXu. Detecting offensive language in social media to protect adolescent online safety. In 2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012, pages 71– 80, 2012. doi: 10.1109/SocialCom-PASSAT.2012.55. URL https://doi.org/10.1109/ SocialCom-PASSAT.2012.55.
13. Subrath Kumar, New Malicious Code Detection Using Variable Length n-Grams Algorithms, Architectures and Information Systems Security, 2008, pp. 307-323
14. Fabio Del Vigna, Andrea Cimino, FeliceDell'Orletta, MarinellaPetrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017., pages 86–95, 2017. URL http://ceur-ws.org/Vol-1816/paper-09.pdf.
15. William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In Proceedings of the Second Workshop on Language in Social Media, LSM '12, pages 19–26, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL
16. ZeerakWaseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proceedings of the First Workshop on NLP and Computational Social Science, pages 138–142, Austin, Texas, November 2016. Association for Computational Linguistics. URL http://aclweb.org/anthology/W16-5618.
17. ZeerakWaseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, California, June 2016. Association for Computational Linguistics.
18. An Innovative Model Based Approach for Credit Card Fraud Detection Using Predictive Analysis and Logical Regression. International Journal of Innovative Technology and Exploring Engineering (IJITEE), Scopus, 2019, 8, 1683-1688.
19. Latent Semantic Indexing based Approach for Identification of Frequent Itemset.Jour of Adv Research in Dynamical & Control Systems, Scopus, 2018, Vol. 10, 686-690.
20. A mechanism for identifying the guilt agent in a network using vector quantization and skew Gaussian distribution.International Journal of Engineering & Technology, Scopus, 2018, 7, 149-151.
21. A Data Leakage Identification System Based on Truncated Skew Symmetric Gaussian Mixture Model. International Journal of Recent Technology and Engineering (IJRTE), Scopus, 2018, 7, 111-113.
22. An Enhanced Model for Preventing GuiltAgents and Providing Data Security inDistributed Environment. IEEE Conference International conference on Signal Processing, Communication, Power and Embedded System (SCOPES), 2016, 1, 337-339.