THOMSON REUTERS ENDNOTE

[Bist* *et al.*, 6(2): February, 2017]
ICTM Value: 3.00

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

CLASSIFICATION OF METAMORPHIC VIRUS USING N GRAM ANALYSIS

Ankur Singh Bist*, Dr. Anuj Sharma
* PHD Scholar SVU Gajraula
SVU Gajraula

DOI: 10.5281/zenodo.291841

ABSTRACT

Recent research work shows that n-gram is widely used in metamorphic virus detection. Viruses generated from kits like NGVCK are detected effectively by n-gram approach. Our purpose is to examine various flavours of n-gram approach in virus detection.

KEYWORDS: n-gram, Metamorphism, Observation Sequence.

INTRODUCTION

Internet has become target of malicious codes due to its increasing use. Malicious codes are executable code and have the capability to replicate. It makes their survival strong. Viruses design and evolution attached with the area of programming. Similar to other computer programs viruses carry functions that are intelligent for providing protection in such a manner that detection remains not easy for virus scanner [1].

Viruses have to take various procedures of intellect for continued existence. That is why they may have complex encrypting and decrypting engines. These are the most frequent methods used by computer viruses in current scenario. They make use of these techniques to mask the antivirus and to adopt the certain environment for their expansion [2].

Figure 1: Assembly code of Virus File

Polymorphic viruses try to hide the decrypting module. More complex methods were developed enabling the virus designers to change the code of one virus file and make multiple morphed copies while maintaining its functionalities. These are the type of viruses which have the ability to mutate itself with the code changed but without changing its functionalities. Metamorphic virus can become a serious threat considering the fact that there can be thousands of variants of one virus file with their signature being totally different [4] [5] [6] [7].

Metamorphic viruses transform its code in a specific manner very frequently and require to be prohibited. Their analysis will lead to evolve a framework where the overall process of detection will be bounded in specific outcomes of continuing evolving results. It is essential to make a distinction between replicating programs and its similar forms. Reproducing programs will not necessarily damage your system [3] [8]. There is big fight between designers of virus and antivirus. The enhanced knowledge about the certain patterns, specifications can be

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

designed. Various malicious codes can be evolved and incremented in well precise and efficient manner. For perfect identification of a metamorphic virus, identification routines must be written that can generate the essential instruction set of the virus code from the actual occurrence of the infection.

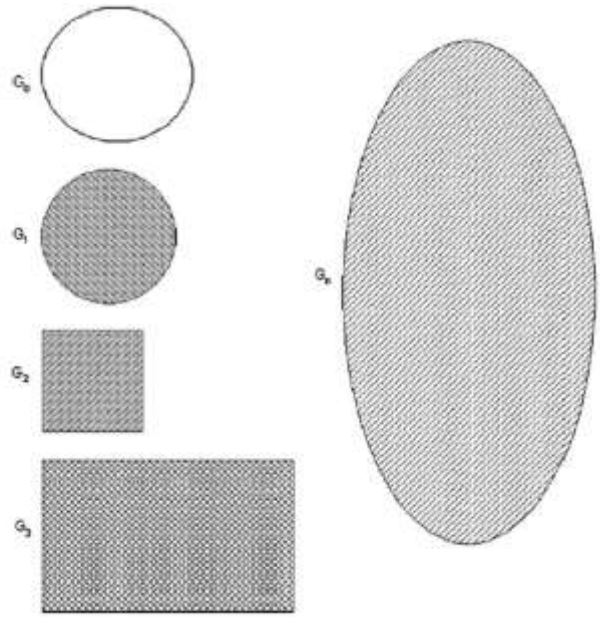

Figure 2: Analogy of Metamorphic Viruses

MALWARE CLASSIFICATION APPROACH USING N-GRAM

TABLE 1: WIN 32 COLLECTION: ACCURACY WITH A LIMIT OF 100,000

L	n														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
20	.45	.58	.52	.63	.69	.57	.55	.51	.50	.49	.44	.42	.38	.36	.37
50	.60	.63	.88	.89	.87	.84	.73	.68	.82	.63	.79	.79	.82	.85	.82
100	.77	.75	.90	.87	.87	.89	.87	.84	.84	.86	.86	.87	.87	.86	.85
200	.85	.76	.87	.89	.90	.90	.92	.89	.90	.90	.90	.89	.89	.89	.90
500	.85	.89	.89	.91	.90	.90	.91	.90	.88	.88	.87	.87	.84	.85	.85
1K	.85	.92	.93	.93	.91	.90	.89	.87	.85	.84	.84	.83	.85	.80	.80
2K	.85	.87	.93	.92	.90	.87	.86	.84	.83	.81	.83	.83	.86	.79	.75
3K	.85	.83	.92	.91	.90	.87	.86	.83	.82	.81	.81	.82	.84	.93	.93
4K	.85	.78	.91	.91	.89	.87	.85	.82	.81	.85	.92	.92	.92	.92	.92
5K	.85	.73	.91	.89	.89	.87	.84	.81	.92	.93	.93	.92	.92	.92	.92
6K	.85	.69	.91	.88	.88	.86	.84	.92	.93	.93	.93	.92	.92	.92	.92
7K	.85	.65	.91	.87	.87	.84	.92	.92	.93	.93	.93	.92	.92	.92	.92
8K	.85	.63	.91	.87	.87	.84	.92	.92	.93	.93	.93	.92	.92	.92	.92
9K	.85	.61	.90	.86	.87	.92	.92	.92	.93	.93	.93	.92	.92	.92	.92
10K	.85	.59	.89	.86	.86	.92	.92	.92	.93	.93	.93	.92	.92	.92	.92

Table 2: Worm Collection: Training accuracy for different values N-GRAM SIZE(n) and POFILE SIZE (L)

L	n									
	1	2	3	4	5	6	7	8	9	10
20	0.54	0.50	0.65	0.74	0.68	0.64	0.52	0.50	0.52	0.43
50	0.62	0.62	0.83	0.80	0.85	0.83	0.72	0.65	0.60	0.57
100	0.80	0.65	0.76	0.68	0.84	0.86	0.85	0.83	0.83	0.85
200	0.75	0.69	0.63	0.62	0.79	0.86	0.89	0.87	0.89	0.88
500	0.57	0.87	0.88	0.70	0.83	0.89	0.88	0.87	0.88	0.89
1000	0.57	0.85	0.89	0.90	0.90	0.89	0.88	0.88	0.89	0.87
1500	0.57	0.86	0.89	0.91	0.88	0.90	0.86	0.85	0.83	0.84
2000	0.57	0.83	0.90	0.90	0.88	0.87	0.84	0.79	0.73	0.74
3000	0.57	0.81	0.88	0.89	0.86	0.83	0.71	0.71	0.64	0.65
4000	0.57	0.78	0.88	0.87	0.84	0.82	0.68	0.64	0.61	0.62
5000	0.57	0.76	0.88	0.85	0.80	0.80	0.64	0.62	0.58	0.61

Tony Abou-Assaleh *et al.* explained signature based n gram malicious codes detection technique. In Table 1 and Table 2 Training accuracy is depicted for different value of n-gram size and profile size. CNG classification is

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

based on profiles for class representation. The similarity measure is used between instance profile and class profile. The following mathematical measure is used.

$$\sum_{s \in profiles}^{n} ((f1(s) - f2(s))/(.5f1(s) + .5f2(s)))^2$$

Where s is any n-gram from one of the two profiles, f1(s) frequency of the n-gram in one profile, f2(s) frequency of the n-gram in another profile. Table 3 and Table 4 show some important observation made by authors.

Table 3: WIN32 collection: Training Accuracy for different values of N-Gram Size (n) and Profile Size (L)

L	n									
	1	2	3	4	5	6	7	8	9	10
20	0.45	0.59	0.51	0.63	0.67	0.59	0.54	0.52	0.51	0.47
50	0.60	0.63	0.88	0.88	0.87	0.85	0.74	0.68	0.81	0.64
100	0.76	0.73	0.90	0.88	0.87	0.90	0.87	0.85	0.84	0.85
200	0.85	0.74	0.87	0.89	0.92	0.90	0.93	0.89	0.89	0.90
500	0.85	0.87	0.89	0.91	0.90	0.90	0.91	0.91	0.90	0.89
1000	0.85	0.90	0.93	0.93	0.91	0.90	0.89	0.88	0.87	0.87
1500	0.85	0.89	0.94	0.94	0.91	0.89	0.88	0.87	0.87	0.86
2000	0.85	0.87	0.94	0.92	0.91	0.89	0.87	0.86	0.85	0.82
3000	0.85	0.84	0.93	0.91	0.90	0.86	0.83	0.81	0.80	0.80
4000	0.85	0.79	0.93	0.92	0.87	0.86	0.81	0.80	0.80	0.79
5000	0.85	0.75	0.93	0.91	0.87	0.86	0.81	0.80	0.78	0.78

Table 4: Win32 collection: Average Accuracy in 5-fold cross validation for different values of N-GRAM SIZE(n) and PROFILE SIZE(L)

L	n									
	1	2	3	4	5	6	7	8	9	10
20	0.64	0.63	0.63	0.61	0.58	0.58	0.55	0.52	0.50	0.47
50	0.58	0.70	0.81	0.87	0.85	0.86	0.80	0.63	0.68	0.64
100	0.75	0.74	0.90	0.87	0.87	0.89	0.88	0.85	0.86	0.85
200	0.85	0.70	0.87	0.88	0.90	0.90	0.91	0.88	0.87	0.89
500	0.85	0.81	0.88	0.91	0.90	0.90	0.90	0.89	0.89	0.88
1000	0.85	0.88	0.90	0.91	0.89	0.89	0.86	0.86	0.87	0.86
1500	0.85	0.86	0.91	0.91	0.90	0.88	0.87	0.87	0.87	0.85
2000	0.85	0.86	0.91	0.91	0.89	0.88	0.87	0.85	0.84	0.84
3000	0.85	0.84	0.91	0.90	0.88	0.87	0.85	0.84	0.83	0.83
4000	0.85	0.84	0.91	0.91	0.89	0.86	0.86	0.84	0.82	0.82
5000	0.85	0.79	0.91	0.90	0.88	0.86	0.86	0.83	0.81	0.87

Adityaram Oza explained "HTTP attack detection is using N-Gram Analysis" In following figure the variation of Mahalanobis distance is depicted.

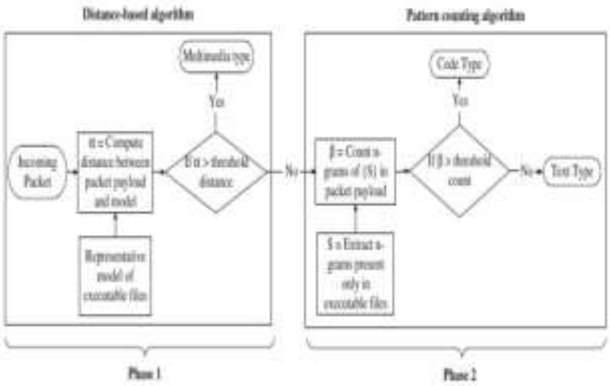

Figure 3: File Type Classification Scheme

The mean vector depicts the expected frequency distribution of bytes in a benign HTTP packet. Following figure shows the training and detection phase of χ^2 distance.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

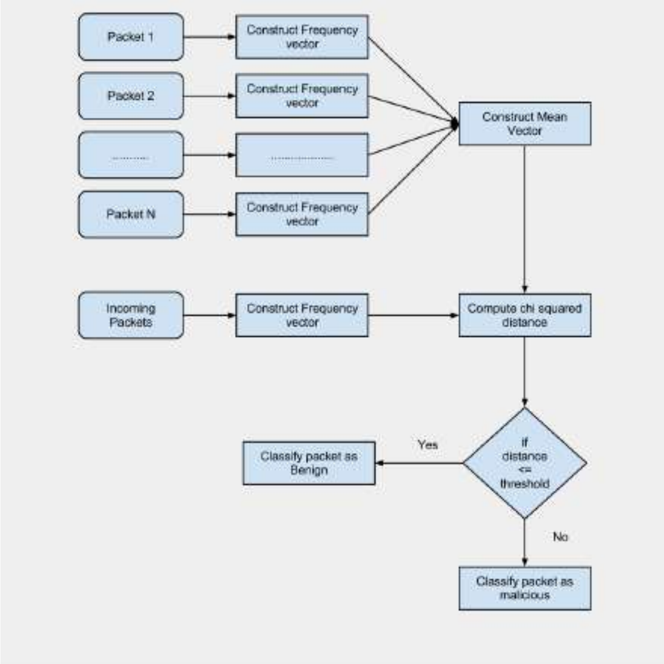

Figure 4: χ^2 ***distance***

Following figure shows training and detection phase of pattern computing technique. Packet count is an important term to classify the normal files and malicious files.

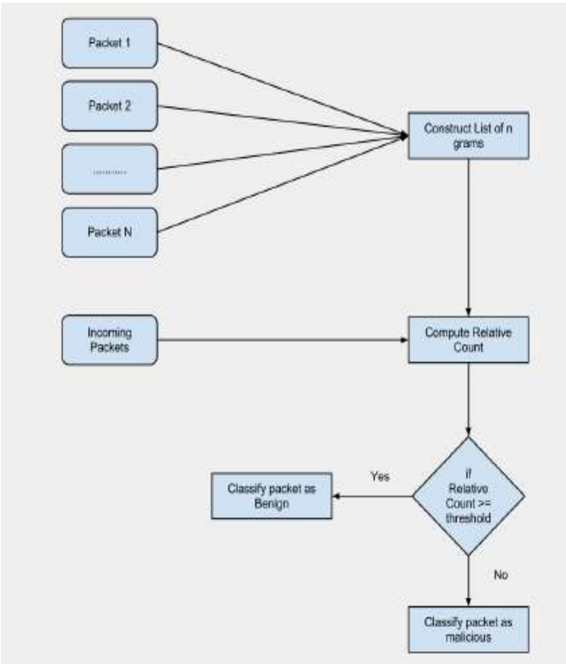

Figure 5: Pattern Counting

Following are some important results observed by researchers.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

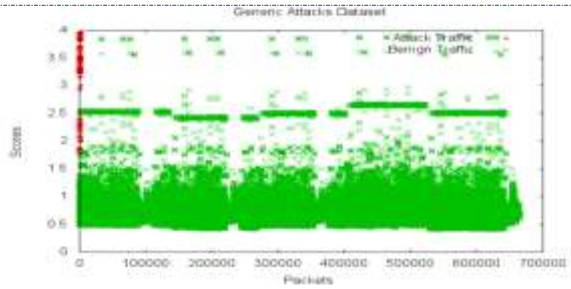

Figure 6: BOX plot-Generic attacks-Adhoc n-gram Distance 3-gram

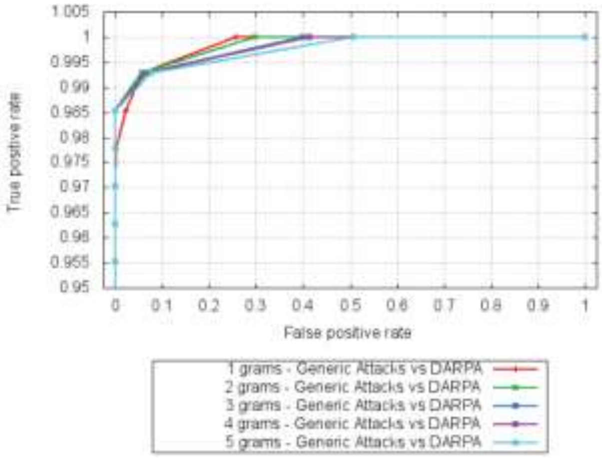

Figure 7: ROC- Generic Attacks- Adhoc n-gram distance

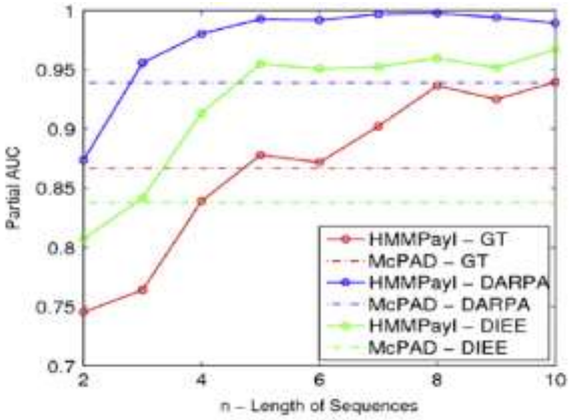

Figure 8: HMM Payl- Generic Attacks- AUCp values for n-gram sizes 2-10

Igor Santos *et al*. explained about N-Grams based file signatures for malware detection. Experiment is performed using 149882 malware files and 4934 benign files. Signatures are built on the set of n gram for n=2, n=4, n=6 and n=8. Some important observation is depicted in following figure and found that n gram detection can be used for malware detection.

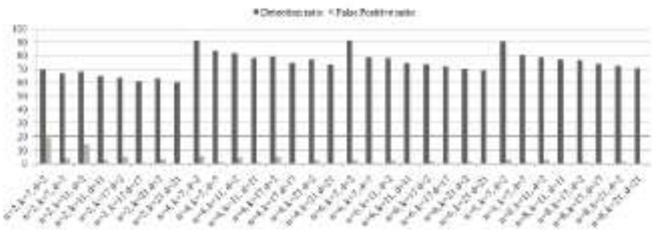

Figure 9: Detection and false positive ratio

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Ohm Sornil *et al.* explained about Malware classification using N-grams sequential pattern feature. N-grams are extracted from malicious program files, sequential n-gram patterns are determined, pattern statistics are calculated, and a classification technique is used to determine the family of malware. Classification models C4.5, multilayer perceptron, and support vector machines are used for classification and 96.64% accuracy are obtained.

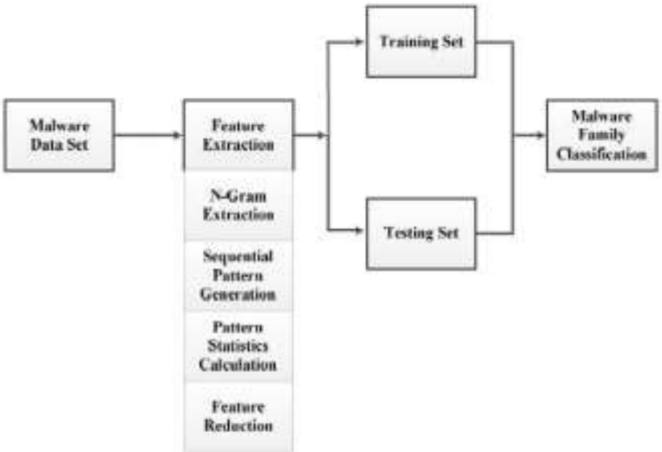

Figure 10: Malware classification method

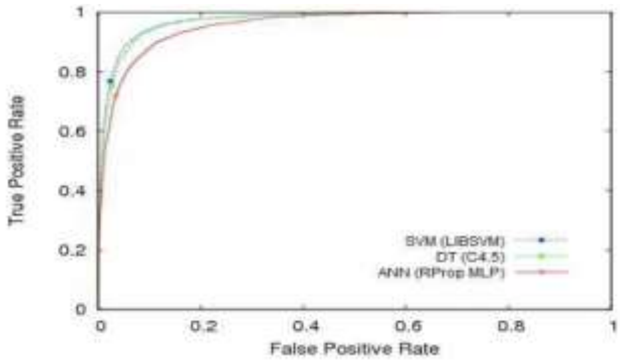

Figure 11: ROC plot for classification

Li et al. explained Fileprint (n-gram) analysis for the detection of malware. During the training part set of models are derived based on structural composition of file. Author applies 1 gram analysis technique to PDF files embedded with malware and achieved good detection rate. Another authors like Sekar *et al.* compared their developed approach with n-gram approach.

Earlier application part of n-gram is performed IBM research group; they used this method for the detection of boot sector viruses. They used n gram technique for different value of n in the range of 1 to 8 depending upon specific method and technique.

Kolter and maloof did the study to settle down the value of n in order to find out optimal solution. Abou-assaleh *et al.* try to find out the condition where n work best.

Kephart and Arnold used a range of n to build a recognition system and found that fix value of n is not sufficient to trace out the best results.

Karim *et al.* explained about n pern technique and how this technique can be used to analyze typical body changing viruses like metamorphic viruses.

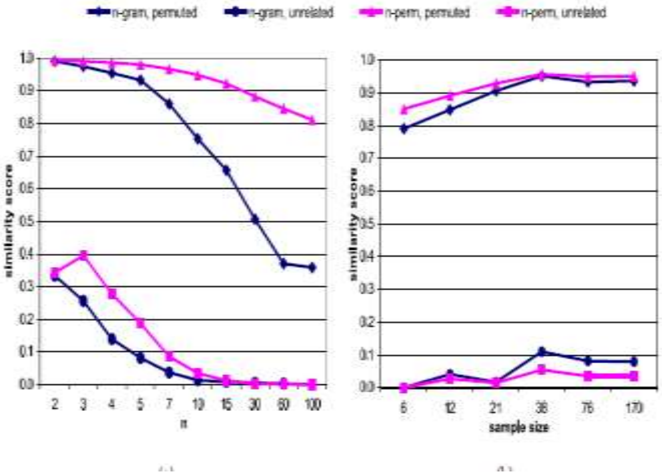

Figure 12: Similarity scores varying by n and sample size

CONCLUSIONS

N-gram is a powerful statistical tool for modeling generative sequences that can be characterized by an underlying process generating an observable sequence.

In this paper a detailed study is made to understand the impact of n-gram analysis in malware detection especially in metamorphic virus detection. Literature study depicts the various dimensions of n-gram that are being explored by researchers in order to enhance its utility in malware detection.

REFERENCES

[1] Bist, Ankur Singh, and Sunita Jalal. "Identification of metamorphic viruses."Advance Computing Conference (IACC), 2014 IEEE International. IEEE, 2014.
[2] Bist, Ankur Singh. "Detection of metamorphic viruses: A survey." Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on. IEEE, 2014.
[3] Bist, Ankur Singh. "Classification and identification of Malicious codes." IJCSE. 2012.
[4] M. Mangesh, Hunting for Metamorphic Java script Malware, Master's Project, pp 359, 2014.
[5] I. Santos, Y. Penya, J. Devesa and P. Bringas, N-Grams-Based files Signature for Malware Detection.
[6] O. Adityaram, HTTP Attack Detection using N-Gram Analysis, Master's Project, pp. 299, 2013.
[7] O. Sornil, C. Liangboonprakong, Malware Classification Using N-grams Sequential Pattern Features, Vol. 14 issue 5, IJIPM, 2013.
[8] Md. E. Karim, A. Walenstein, A. Lakhotia, Malware Phylogeny Generation using Permutation of Code. 2005, Journal of Computer Virology.