Efficient Detection of Spam by Monitoring Outgoing Messages using Reverse Dictionary

S.Gavaskar[1] , LGXAL.Agnel Livingston[2]
[1]Assistant Professor, [2]Assistant Professor
Dept. Computer Science and Engineering
St.Xavier's Catholic College of Engg
Nagercoil,India.

Abstract—**Many attacks are present in the internet; one of them is spam attack. Unwanted or intrusive advertising message send on on the Internet is called spam. There are two types of Spam detection namely outbound and inbound.The outbound detection will detect the spam by watching the system sending phishing,spam,virus messages and emails.The Inbound detection prevent the system by securing from malware and virus and prevent from compromising the system.Many effective method are used to detect bots in local networks but there may be a chance to detect normal message as spam message so we focus on the detection of the Spam messages in the email through Reverse Dictionary approach.**

Keywords-Semantic, spam zombies, Reverse Dictionary.

I. INTRODUCTION

The term internet is known as interconnection of computers and computer network.The network and computer are subjected to wide range attach such as altering and modification of data/information,destroying and exposing of sensitive data/information,unauthorized use of resource and assets.The message received by the user is known as spam if it consist of irrelevant message for the context of the user.The spam messages are sent by the attacker for the purpose of promoting advertisement,sending virus and malware message.The spam is also known as junk message in which identical unwanted message is send as link to many number of host/user simultaneously.when user click this they were taken to phishing website and malware web sites.In spam email the malware is included as executable file or script as attachment

In internet terms Spam is flooding of host or computer with many replica of the same un useful message.It pave way to force the mail,message and attachment on user who is not willing to receive that actually. Many of the spam are business advertising message by the third part advertiser and it is frequently made for unworthy products, getting positive feedback on the product etc.The cost of destruction due to spam is mostly spend by carrier of the message and recipient than the sender of the message

The computers/device used by user is taken control by attacker/spammers without the knowledge of the device/computer user. This computer/device are termed as spam zombies which send the spam message to entire user connected to network without the knowledge of the user. The term Botnet is used to refer the collection of spam zombies present in a particular network

Zombie detection is used for both inbound and outbound protection.The outbound detection will detect the spam by watching the system sending phishing,spam,virus messages and emails.The Inbound detection prevent the system by securing from malware and virus and prevent from compromising the system

Botnet, a collection of spam zombies is a major threat for security in internet and intranet.The done based on botnet is difficult and hard to to detect and defend.Many methods were used but in this methods there may be possibilities of detecting the normal message as a spam message,so we are using the reverse dictionary for accurately detecting the spam message present in the message.

In Reverse dictionary it takes a phrase as user input and it compares that phrase with the phrases of traditional dictionary if both the phrase matches it return a set of candidate word, from the set of candidate words we can find the exact word that we want.From reverse dictionary we can get word from the phrase or sentence ,that we are intending to think or we have in our message. The matching technology in reverse dictionary depends both on the type of the query and on the type of the forward dictionary. To achieve reverse dictionary two steps are to be maintained (1)the given user input is matched with the definition of a word in the forward dictionary (2)the responsiveness of the reverse dictionary should be similar to the responsiveness of the forward dictionary that is it does not take too much time. Our goal is to construct a reverse dictionary from the given two or more forward dictionaries. Here, WordNet is used to get the semantics of the spam word.

II. RELATED WORKS

Z. Duan, K. Gopalan, and X. Yuan(2007) differentiate the attacker/spammer character at computer network and server present in the remote.It also relate the incoming of spam/malicious message with the update made in BGP router to find the network usage pattern of attacker/spammers. the usage characteristics of attacker/spammers such as the information of spam messages/email from many number of spammers/attackers, the spam received patterns across the IP addressing, the number of remote email servers in different (spam) networks, and the time at which spammers are active can significantly affect the effectiveness of many anti-spam mechanisms.

Fernando Sanchez, Zhenhai Duan & Yingfei Dong(2010) provide the study on the header fields received from spam emails. They found, at what extend spammers/attacker can trace information of spam emails. It is well known that spammers/attacker can get the header of an email and trace the information/data for hiding the true origin of the email. The ability of spammers/attacker to change email headers often make spam control ver difficult and makes it hard to detect the true spammers/attackers.

J.P. John et al(2009) present Botlab, a system that frequently watch and find the characteristics of botnets which are spam oriented found in the network. Botlab collect various real-time information/data about spam botnets present in the network. By Analyzing and watching these streams, Botlab will generate accurate, timely, and collective data/information about spam botnet character and properties..

Botlab get control of all incoming spam message received and it allow to find new botnet present in the network. It then run many prior , sandboxed compute/device for finding various botnets, allowing it to observe the accurate sent spam message from the computer/device. It identify the spam message for URLs, by gathering information/data on scams, and identifies the malicious links. Finally, it relate the received and sending spam message to identify the most active running botnets and all of the compromised computer/device present in each botnet.

M. Xie, H. Yin, and H. Wang (2006) propose a technique DBSpam, to find and avoid spamming functionality done inside a computer network. Watching the incoming/outgoing traffic send through a network gateway, DBSpam uses a simple method on statistical,known as Sequential Probability Ratio Test; to find the activity of spamming in a frequency manner. It is done to be placed at a network intermediate point such as the edge router or gateway that connects a computer network/device to the Internet. Due to the protocol characteristics of SMTP and timing the character of proxy-based spamming/attacks illustrate the unique characteristics of packet symmetry and connection correlation.By using this type of spamming behavior, the suspicious TCP connections that are involved in spamming are identified.

G. Gu et al(2007) present a new kind of computer network monitoring technique called BotHunter, which focuses on finding the infection pattern that occurs during a malware infection process. The BotHunter uses Snort ,a customized malware-focused ruleset, which is further used with two additional bot specific anomaly-detection plugins for malware analysis:SLADE and SCADE.

SLADE uses a lossy n-gram analysis of receiving traffic flows, using byte-distribution divergences in selected protocols that are indicative of frequent malware intrusions. The BotHunter correlator is driven by Snort with a customized malware-focused ruleset, which is further augmented with two additional bot specific anomaly-detection plugins for malware analysis: SLADE and SCADE. SLADE implements a lossy n-gram payload analysis of incoming traffic flows, targeting byte-distribution divergences in selected protocols that are indicative of common malware intrusions.

III. PROPOSED SYSTEM

The proposed system uses the reverse dictionary technique to identify, prevent, and report the spam. The reverse dictionary checks ‘Best health insurance’ phrase with the phrases or definitions of forward dictionary and retrieve a set of words like health checkup, health care, health insurance etc. Like that it returns many words that match this given phrase. Then it extracts each word and performs semantic matching and the appropriate word ‘health insurance’ is retrieved.

A. Synonym Matching

Synonym matching represents a fundamental technique in many applications in areas such as discovering of resource, integration of data, migration of data, translation of query, peer to peer networks.. It is also being used in other areas such as event preprocessing. Synonym matching is a technique used to identify information which is meaningfully related to the given text.

B. Retrieval of set of words

The second step is the retrieval of set of words. The given input definition or phrase is matched with all the dictionary definitions, all dictionaries contain more than hundred thousand words in it .It checks the given phrase with all these ten thousand words dictionaries and if matches it return some list of words. This retrieval of set of words can be done based on synonym matching.

C. Semantic Extraction

The third step is semantic extraction. All the retrieved words are analyzed for finding the nature of words. From this nature of words structuring of words can be done for extracting the semantic features. Then these features are checked with list of forward dictionaries and finally the word matching can be done.

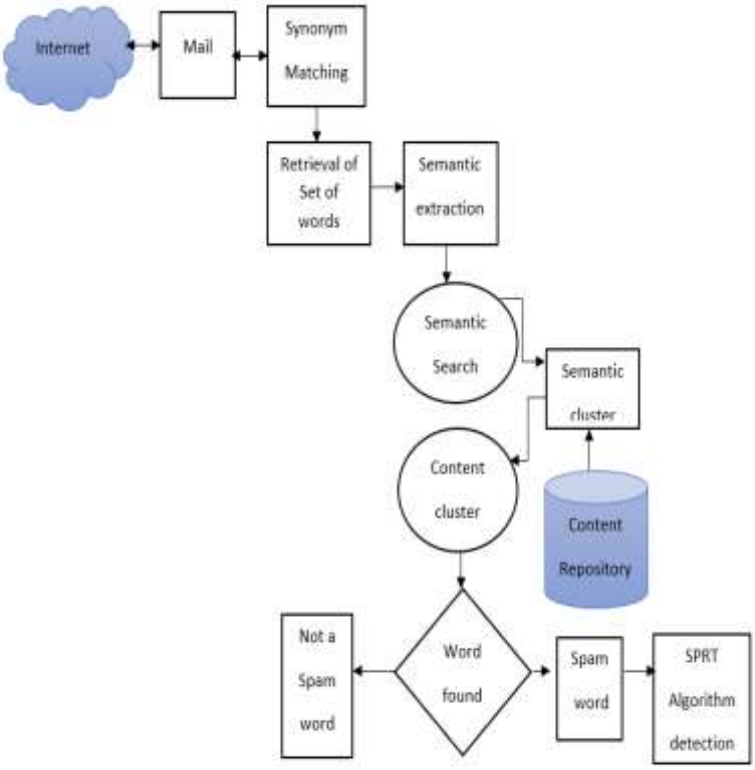

Fig1:Spam Detection Using Reverse Dictionary

D. Content Repository

A database is used to store the retrieved words. This contains the Metadata, the definition of the retrieved words and related nouns and verbs of the retrieved words. It can also hold all the information that is needed for dictionaries.

E. Semantic Text Matching

Next step is the Semantic text matching. From the semantic clusters each text in the given input phrase is matched with the contents of semantic groups. If it is exactly matched the matched contents are retrieved. Semantic text matching means searching the web on the basis of meaning of user's query. For an effective and more relevant search results, it is necessary to understand and interpret the query in the way user wants. The meaning of the query is hidden in the query itself. Words like what, when, why etc. can change the total meaning of the query.

F. Word found

Finally it checks whether the appropriate word is spam message or not. If the word present is a spam message it checks for the SPRT detection algorithm. If the given word is not a spam word it ignores the message. For this Wordnet Dictionary is used it checks them by checking spelling, links etc.

G. SPRT algorithm

It is an machine learning algorithm and it is more efficient algorithm for detecting the spam messages. First the SPRT algorithm gets a threshold value from the user and marks his IP address and initializes a count value as one. If the next time a message is came from the same IP address it increase the value count as two and so on. Likewise the count increases if the count is equal to the threshold value then the receiving message is a spam message. It is an effective and efficient spam detection system.

IV. CONCLUSION

In this study a reverse dictionary based on SPRT spam detection is proposed. This method is comparatively efficient than the previous methods. It is an effective and efficient method for spam detection. False positive and false negative can be bounded by user specified threshold.

REFERENCES

[1] Zhenhai Duan, Yingfei Dong(2012)" Detecting Spam Zombies by Monitoring Outgoing Messages", IEEE Transactions On Dependable And Secure Computing, Vol. 9, No. 2, March/April 2012

[2] Ming-Yen Chen, Hui-Chuan Chu, Yuh-Min Chen (2009), "Developing a semantic enable information retrieval mechanism", Elsevier Journal on Expert Systems with Applications, May 2009.

[3] Zongli Jiang and Changdong Lu, "A latent semantic analysis based method of getting the category attribute of words" 2009 International Conference on Electronic Computer Technology.

[4] Hongwei Yang, "A document clustering algorithm for web search engine retrieval system",2010 International Conference on e-Education, e-Business, e-Management and e-Learning.

[5] Jianpei Zhang, Zhongwei Li, Jing Yang, " A divisional incremental training algorithm of support vector machine" Mechatronics and Automation, 2005 IEEE Conference.

[6] Gang Lv, Cheng Zheng, Li Zhang, "Text information retrieval based on concept semantic similarity" 2009 Fifth International Conference on Semantics, Knowledge and Grid.

[7] Trong Hai Duong, Geun Sik Jo, Ngoc Thanh Nguyen, "A method for integration across Text Corpus and WordNet based ontologies" 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[8] Zhongcheng Zhao, "Measuring semantic similarity based on WordNet" 2009 Sixth Web Information Systems and Applications Conference.

[9] Trong Hai Duong, Geun Sik Jo, "Semantic similarity methods in WordNet and their application to information retrieval on the web" 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[10] Wei-Dong Fang, Ling Zhang, Yan-Xuan Wang, Shou-Bin Bong, "Toward a semantic search engine based on ontologies" Network Engineering and Research center, South China University of Technology, Guanghou 510640, China.

[11] Qinglin Guo, Ming Zhang (2007), "Multi-documents automation abstracting based on text clustering and semantic analysis", Elsevier Journal on Knowledge based systems, 22, 482-485.

[12] Berry, M.W. (1992), "Large scale singular value computations", International Journal of Supercomputer Applications.

[13] Jiuling Zhang, Beixing Deng, Xing Li, "Concept based query expansion using WordNet" 2009 International e-Conference on Advanced Science and Technology.