ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

An Analysis and Review of Spam Detection Techniques in Various Forms

Ms. R.Priya[1], Dr. P. Sumathi[2],

Research Scholar, PG & Research, Department of Computer Science, Government Arts College (Autonomous), Coimbatore, India[1]

Assistant Professor, PG & Research, Department of Computer Science, Government Arts College (Autonomous), Coimbatore, India[2]

ABSTRACT: Spam are the most occurred threat in the online usage system which would be send periodically to the multiple users continuously which are invalid message for the corresponding users. These spam messages might cause the performance of users in various forms like system corruption, performance degradation and so on. This spam messages needs to be prevented for the system performance improvement. There is various research works has been conducted previously for better prevention of the spam messages from the hackers. This analysis work attempts to review of various research works that has been conducted with the goal of prevention of spam messages. This research works also provides the discussion of merits and demerits of the different research methodologies to predict the better approach that can provide the way of better detection of spam messages.

KEYWORDS: Spam detection, hackers, filtering, bot nets

I. INTRODUCTION

Spam is defined as unwanted or unnecessary electronic junks which come frequently to emails. Some unwanted ads or posts that are displaying in the web pages also referred as spam which should be analysed efficiently for the better performance improvement. Spam is considered as the most concerned threat in the online usage system which might cause the performance of system by consuming larger storage consumption and leading to overload of the system. The overload of the system would degrade the system performance which might lead to corruption and loss of the system files. This is the biggest research concern in the biggest industries and the organization in the real world who is about to maintain various files and information.

Some automated mechanisms that are implemented for spam detection would prevent the unsolicited messages from displaying. This automated mechanism might prevent the email from known persons too due to usage of new email ids. This would be greater trouble in the online usage systems, thus differentiation of unsolicited messages from the known mail ids should be done. The automated mechanism must be programmed to find the differentiation between the unwanted ads and the wanted emails.

These spam mails needs to be prevented for the better utilization of storage spaces and preventing system from failures. Various methodologies has been introduced earlier which attempts to prevent the spam storage emails. However, those methods are mostly similar to each other which can be easily detected by the various anti spam techniques. Thus, there is a chance of race course between the spam detection techniques and the anti spam techniques. This analysis work attempts to discuss the various features, merits and demerits of spam detection and anti spam techniques with the concern of the different terminologies and factors. The following section would provide the detail of the different features that can be used for detection of spam messages.

There are three types of spam are present in the real world environment. Those are link spam, content spam, and cloaking. Link spam considers the link based applications to detect the spread the spam message too many persons through linkage analysis. Some of the examples of link based applications are page ranking; web application ranking and trust based ranking and so on. Content spam is a kind of spam techniques to modify the various terminologies and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

the contents that are present over the web pages and the real world applications. Cloaking is a technique to modify the contents that are present in the web pages and preventing the real world pages to display the original contents. Among these most popular spam detection technique is the take usage of web page application features to modify or create the contents that are present in the system.

The main contribution of the analysis work is to research the various previous research methodologies that have been conducted previously and discussing the merits and demerits of those methods to predict the better approach from which research can be focussed. The overall organization of the research methodology is given as like as follows: This section provides the detailed introduction about the spam detection techniques. In section 2, different research methodologies that has been conducted previously for detection the spam messages has been discussed. In section 3, merits and demerits of the previous research methodologies that have been conducted in the section 2 has been given. In section 4, overall analysis of the research methodology is concluded.

II. RELATED WORK

This section provides the detailed discussion of the different previous research methodologies that has been conducted by various researchers has been given. The methodology used and the working procedure of the different research methodologies has been discussed in depth and the merits and demerits of those methodologies has been given in the following section.

Brian Whitworth et al (1) discussed and analysed about the spam details and the introduction about the detection methodologies. The author provides the role of spam detection in the real worlds and the effects of spam in the application in terms of their causes. The author has given detailed introduction about the impacts of the spam detection techniques with their effects of reduction of gap in the real world application. To do so, author has developed a software system that concern on finding the spam occurrence. This software system is capable of finding the spam occurrence in the automated manner with improved performance.

Linda Dailey Paulson (2) analysed the resolution for the spam prevention in the automated manner permanently. The author has analysed the many ways that can be used for complete prevention of the spam mails from the different sources. This work provides the various issues that might be arise due to arrival of more junk messages. These causes and effects of these issues are given in the detailed manner with the concern of the different terminologies. The side effects of spam in the long running application are also discussed in the detailed manner. The author has concluded that there is no way for prevention of the spam mails in the complete manner due to their various threat occurrence and their side effects.

VedatCoskun et al (3) proposed a quarantine region scheme for effective detection of the spam attacks in the wireless sensor networks. The main goal of this proposed scheme is to detect the issues like delay tolerance and the energy problems that might arise in the wireless sensor network due to the occurrence and arrival of spam messages in to the environment. This detection mechanism is applied in the quarantine region for detection of the spam mails to improve the performance of the system. The author has proved that the QRS can achieve the trade off between the resilience against spam attacks and the number of authentications. The experimental results of this work concludes that the propose scheme can achieves the better prevention of the spam message into the wireless sensor network system.

Adam J.O'donnell (4) designed a framework called the microcosm which is called as stock spam. The stock spam is nothing but the group of the spam emails and the stock messages that are incorporate with each other in terms of analysis of various merits. The stock spam allows users to define the users to group the various emails along with their advertisement in the secured manner. This framework allows users to define the different terms and concepts that are reasonable for the provisioning of the emails such as junk mails. This framework attracts various industry people to implement in their organization to provision and handling of spam messages in the considerable manner.

Kang Li et al (5) introduced ALPACAS—a privacy-aware framework for collaborative spam filtering which focus on spam filtering process in the web pages with the concern of privacy of users without leaking their sensitive information. This is achieved by following two steps. Those are preserving the features of the environment with the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

concern of the various contents and collusion items present in the environment. Improving the privacy of the users by securely transmitting messages between ports by using the privacy preserved protocols. This process is done in the collaborative manner to improve the spam detection process by filtering out the contents that are present in the real world applications. The overall research of this work proves that the proposed research can contribute towards better detection of the spam mails with improved privacy and less false detection rate.

Lourdes Araujo et al (6) proposed a link based features with language model to predict the spam arrival in the effective manner. This method assures the detection of the spam mails and junk mails can be done in the effective manner by learning the feature that are main reasonable for those implementation. This is done by introducing the classification mechanism into the system that can assure the learning of knowledge. This proposed methodology would be applied on the contents that are retrieved from the various web pages in order to allow them to learn the different methodologies. This mechanism leads to a complete knowledge learning about the proposed mechanism, thus the experimental results of this work provides better result by detecting the spam mails in the effective manner.

Chi-Yao Tseng et al (7) proposed a novel Email abstraction scheme for better prediction of the spam junks that are entering into the system. This mechanism is based on near duplicate detection scheme which attempts to preserve the Email junk detection capability in the effective manner by implementing the methodology that can identify the near duplicate items. This methodology is based on Email layout system which can identify the knowledge of the system in the effective manner by updating the changes occurred in the dynamic manner. The overall research of this work concludes that the propose mechanism can provide the accurate result. The layout of the Emails would be updated periodically by learning the knowledge of information system using the learning mechanism. This methodology is evaluated on the real live data set to evaluate the performance improvement.

ZhenhaiDuan et al (8) proposed a SPOT framework which is a spam detection mechanism. The SPOT works based on monitoring the messages that flowing in the environment. The divergence behaviour of different messages that are transferred between the different mediums would be calculated for identifying the pattern of the spam messages. This is done by collecting the two-month e-mail trace from various sources. This mechanism can automatically detect the spam messages that are arriving into the system in terms of evaluating the performance and divergence of the different spam messages. The performance improvement of the proposed mechanism is evaluated by comparing it with the other zombie Spam detection mechanism in terms of performance measures called the accuracy and precision rate.

HaiyingShen et al (9) proposed SOAP methodologies which make utilize the social networking environment for the better prediction of the spam messages that are entering into the system in terms of improved utilization of the system. This methodology is based on the distributed approach which would collect the spam node details from the nodes which would be shared with other nodes present in the environment to improve the security. The proposed methodology of this work provides the performance improvement over the real time environment by adapting the social network environment knowledge information. SOAP integrates four components into the basic Bayesian filter: social closeness-based spam filtering, social interest-based spam filtering, adaptive trust management, and friend notification.

JiHua et al (10) introduced the novel mechanism that learns the non spam content features that are present in the environment. This is done to analyse the behaviour of non spam contents, thus the spam mails can be prevented in the effective manner. This is done to differentiate the spam contents from the non spam contents in the effective to prove the better knowledge of the system. To do so, calculation system probability equation is proposed which can differentiate the spam contents from the non spam contents in the distributed and efficient manner. The overall research of this work proves that the proposed research mechanism provides better result in terms of optimal detection of spam behaviours.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

III. COMPARISON ANALYSIS

This section provides the merits and demerits of different research methodologies which have been conducted previously with the concern of the effective detection of spam messages and mails.

Table 1. Comparison Analysis of different research methodologies

S. NO	TITLE	AUTHORS	METHOD	MERITS	DEMERITS
1	Spam and the Social-Technical Gap	Brian Whitworth, Elizabeth Whitworth	Spam detection	Provides detailed overview of role of spam detection techniques in the real world applications. Improved performance in case of arrival of more load of memory too	It concludes that the various features are present in the system can affect the spam detection process. Developed software system is not that much efficient in detection of the spam detection
2	No Quick Fix for Spam	Linda Dailey Paulson	Analysed the threat of spam occurrence	There are numerous methods has been proposed for detection of spam mails Better analysis of the spam and junk mail causes and effects	This work concludes that there is no methodology is present for the complete detection of spam mails arrivals
3	Quarantine Region Scheme to Mitigate Spam Attacks in Wireless Sensor Networks	VedatCoskun, ErdalCayirci, Albert Levi, and SerdarSancak	Quarantine Region Scheme	Better detection of spam messages Improved accuracy of	It can authenticate about only 50 percent of authentication messages

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

				detection of spam emails Achieves better trade off between detection of spam emails	Can lead to more computation overhead
4	The EvolutionaryMicrocosm of Stock Spam	Adam J.O'donnell	Microcosm	Providing user flexible environment to allow them to publish their ad contents in the environment Can lead to better handling of the spam message with the concern of system performance improvement	More computation overhead Increased detection capability of researchers to provide a flexible environment.
5	Privacy-Aware Collaborative Spam Filtering	Kang Li, ZhenyuZhong, andLakshmishRamaswamy	ALPACAS—a privacy-aware framework for collaborative spam filtering	More privacy which provides the secured environment for the users to prevent from the performance degradation which occurs due to spam mails Improved false positive rate by detecting the spam in the effective and accurate manner	More computation time required to resolve the system configuration details Classification result might get affected in case of more arrival of incoming messages dynamically

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

6	Web Spam Detection: New ClassificationFeatures Based on Qualified Link Analysisand Language Models	Lourdes Araujo and Juan Martinez-Romo	link-based features with language-model (LM)	Improved detection of spam mails that are arriving into the system by learning the features of the different web pages More secured environment can be assured by detecting the spam in dynamic manner	Accuracy might get reduced in case of presence of more computation overhead
7	Cosdes: A Collaborative Spam DetectionSystem with a Novel E-Mail AbstractionScheme	Chi-Yao Tseng, Pin-Chieh Sung, and Ming-Syan Chen	novel e-mail abstraction scheme	Better prediction of the knowledge of the system by learning the system spam detail Periodical update of Email structure leads to the better prediction of spam junk mail arrivals	Upto date information maintenance might violate the performance of the near duplication checking mechanism
8	Detecting Spam Zombies byMonitoring Outgoing Messages	ZhenhaiDuan, Peng Chen, Fernando Sanchez,Yingfei Dong, Mary Stephenson, and James Michael Barker	spam zombiedetection system	More accuracy in prediction of spam zombie messages Improved performance of real world system in better prediction of	Less system performance due to handling of large volume of data's that are entering into the system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

				the spam messages More precision and accuracy rate	
9	Leveraging Social Networks forEffective Spam Filtering	HaiyingShen, and Ze Li	SOcial network Aided Personalized and effective spam filter (SOAP)	Better detection of the spam messages Less computation overhead More privacy in terms of analysing the various parameters that are considered	Less accuracy in detection of spam messages
10	Analysis on the Content Features and Their Correlationof Web Pages for Spam Detection	JiHua, Zhang Huaxiang	calculating probabilityformulae of the entropy and independentn-grams	Mo re accuracy in prediction of spam messages Improved system performance	More content features might degrade the system performance.

IV. NUMERICAL ANALYSIS

This section provides the numerical analysis of the various methodologies that has been introduce with the concern of detection of spam contents that are present in the system. This performance analysis is done based on the performance measure called the accuracy.

Accuracy is defined as the detection of spam contents that are present in the system in the accurate manner. This is the ratio between the number of web pages that are analysed and the number of web pages that are correctly analysed without fault. Accuracy of the proposed methodology should be high than the existing methodologies for the better system performance. Numerical evaluations of the various methodologies are depicted in the following graph.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

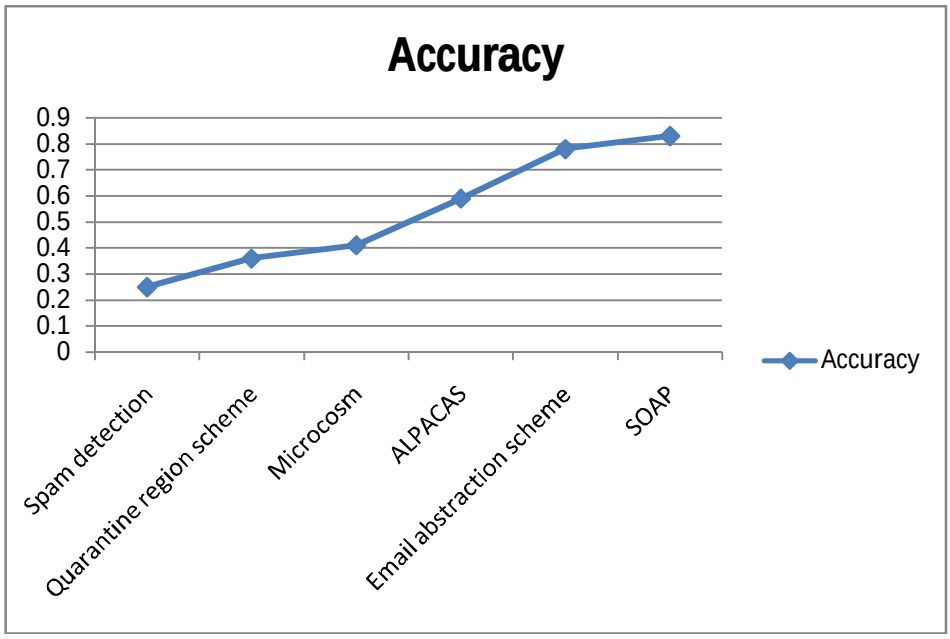

Figure 1. Accuracy Comparison

The above graph depicts the accuracy comparison for the various methodologies. This graph proves that the SOAP approach provides better result than the other methodologies with improved accuracy rate.

V. CONCLUSION AND FUTURE WORK

Spam is a most concerned threat in the online usage application where the adversaries attempt to violate the system performance by continuously sending junk mails or spam mails. This unwanted and repeated receiving message behaviour would violate the system performance by corrupting operating system. This analysis paper discusses the various methodologies in terms of different spam detection techniques to find the properties. The numerical evaluation of this work proves the SOAP methods can provide better result than the other mechanism in terms of improved accuracy rate.

REFERENCES

1. Brian Whitworth, Elizabeth Whitworth, “Spam and the Social-Technical Gap”, Published by the IEEE Computer Society, 2004
2. Linda Dailey Paulson, “No Quick Fix for Spam”, IT programmer, June 2005
3. VedatCoskun, ErdalCayirci, Albert Levi, and SerdarSancak, “Quarantine Region Scheme to Mitigate Spam Attacks in Wireless Sensor Networks”, IEEE transactions on mobile computing, vol. 5, no. 8, august 2006
4. Adam J.O’donnell, “The Evolutionary Microcosm of Stock Spam”, published by the IEEE computer society, 2007
5. Kang Li, ZhenyuZhong, and LakshmishRamaswamy, “Privacy-Aware Collaborative Spam Filtering”, IEEE transactions on parallel and distributed systems, vol. 20, no. 5, may 2009
6. Lourdes Araujo and Juan Martinez-Romo, “Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models”, IEEE transactions on information forensics and security, vol. 5, no. 3, september 2010
7. Chi-Yao Tseng, Pin-Chieh Sung, and Ming-Syan Chen, “Cosdes: A Collaborative Spam Detection System with a Novel E-Mail Abstraction Scheme”, IEEE transactions on knowledge and data engineering, vol. 23, no. 5, may 2011
8. ZhenhaiDuan, Peng Chen, Fernando Sanchez, Yingfei Dong, Mary Stephenson, and James Michael Barker, “Detecting Spam Zombies by Monitoring Outgoing Messages”, IEEE transactions on dependable and secure computing, vol. 9, no. 2, march/april 2012
9. HaiyingShen, and Ze Li, “Leveraging Social Networks for Effective Spam Filtering”, IEEE transactions on computers, vol. 63, no. 11, november 2014
10. JiHua, Zhang Huaxiang, “Analysis on the Content Features and Their Correlation of Web Pages for Spam Detection”, Communications System Design, March 2015

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

BIOGRAPHY

Dr.P.Sumathiis working as an Assistant Professor in the Department of Computer Science, Government Arts College, Coimbatore. She completed PhD in the area of Grid Computing at Bharathiar University. She completed M.Phil in the area of Software Engineering at Mother Teresa Women's University. She completed MCA at Kongu Engineering College at Perundurai. She has published many national and International journals. She has about seventeen years of teaching and research experience. Her research interests include Data Mining, Distributed Computing and Software Engineering.

Ms.R. Priyais an M.Phil Research Scholar in the Department of Computer Science, GovernmentArts College, Coimbatore. She completed M.Sc Computer Science at Government ArtsCollege, Coimbatore. She completed B.Sc Computer Science at NKR Government Arts College for Women, Namakkal.