A Survey on Collaborative Reputation-Based Vector Space Model (CRVSM)

Masrath Parveen[1], Dr. Saurabh Pal[2], Dr. Venkateswara Rao CH[3]

[1]Research Scholar, Dept of CSE, V.B.S.Purvanchal University, Jaunpur
[2] Department of CSE, V.B.S.Purvanchal University, Jaunpur
[3]Department of CSE, Siddhartha Institute of Engineering and Technology, Hyderabad

Abstract: - The CRVSM has been built using the Java programming language and Map Reduce, and tests on a dataset of 1.4 million emails have been done. False Positive Rate (FPR), False Negative Rate (FNR), Detection Accuracy (DA), Spam Detection Time (SDT), Spam Detection Rate (SDR), Network Service Ratio (NSR), and Overall Throughput were usedto measure CRVSM performance (OT). With improved detection accuracy and in less time, CRVSM reduces the FPR and FNR values while detecting spam emails in vast spaces. To get the best collection of features and lower the dimensionality of the feature space, a unique improved Optimized Feature Selection Protocol (OFSP) has been developed. It works in four phases: Feature Selection Phase, Normalization Phase, Score Assignment Phase, and Optimal Feature Selection Phase. OFSP is a hybrid rule-based technique that combines two well-known feature selection approaches for email spam filtering. The Optimized Vector Search Algorithm (OVSA) is a tool used by OFSP to generate dynamic threshold values. The effectiveness of OFSP was examined through experiments to determine its false positive and false negative rates, detection accuracy, and spam detection time. The outcomes demonstrate that, by obtaining extremely strong performance and producing optimum results, OFSP surpasses CRVSM and PEP. In addition to the experimental investigation, the time complexity of the CRVSM, PEP, and OFSP procedures has been examined using complexity analysis. Results demonstrate that OFSP performs better than CRVSM and PEP protocols by generating much less overhead.

Keywords: - CRVSM, mail spam, SDT, SDR, OFSP and PEP.

1. Introduction: - According to reference [1], cyber-risk is the threat or crime related to a hostile event brought on by a malware assault in cyberspace that results in disruption and financial loss. Security risks including spamming, hacking, and phishing are common examples of it. Hackers deftly infiltrate online networks and abuse user data, including credit card information, leading to significant financial loss. According to studies, financial losses as a result of cyber attacks might total $20 trillion by 2020. Symantec software suffered a loss of $50,000 in 2012 as a result of a cyber exaction assault.

According to reference [2], cyber-security is a significant contributor to the financial risk of an enterprise. Reference [2] has created a Copula based Bayesian Belief Network (CBBN) model for assessing and cataloguing a company's cyber-risk. Reference [4] created an analytical approach to fend against cyber security risks. Using a model [5] are able to fight against cyberattacks that take use of space-based assets.

Reference [6] looked at whether cyberattacks constitute cyber warfare. Many additional researchers have successfully completed their work on guarding against cyber attacks [7-8].

Effects of Email Spam

Due to its clear lack of expense for message transport, email communication [9] is the most important and convenient method of communication [10]. Simple Mail Transfer Protocol (SMTP), which is described in RFC 821, is the protocol used by email users to send and receive messages. Email senders utilise the Multipurpose Internet Mail Extension (MIME) to deliver messages as plain text, which anybody may format. Email is therefore subject to spam assaults because of this. In Figure 1.1, the operation of an email is depicted.

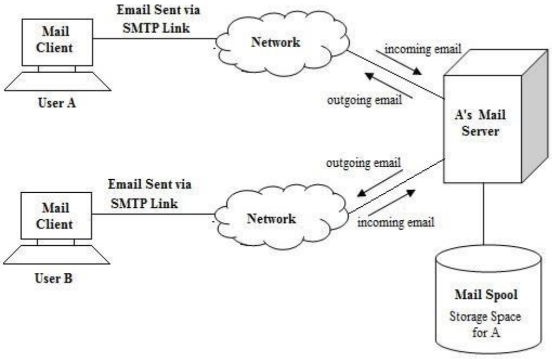

Figure 1.1 The Email System

Three parts make up the email system: the mail programme, mail server, and mailbox. The client-side application that enables reading, managing, and email composition is called the mail programme. Emails are received, stored, and delivered via the mail server. Emails are stored in folders called mailboxes. Let's say node A wishes to email node B. Node A uses a mail application to create and send the email, which is then sentto Node B's mail server over SMTP. Now that the email is stored on node B's mail server, node B may access it whenever it becomes accessible.

Spam emails, also known as suspicious, informal, and fraudulent emails, are unsolicited emails sent for a variety of reasons [11]. The definition of SPAM [12] is "Signal Processing to Analyze Malware." The most well-known type of cyber assault is email spamming [13], which heavily utilises cyber resources like memory, network bandwidth, computing power, etc. Web spam [14], public network spam, and email spam are diverse types of spam [15], Voice over Internet Protocol (VoIP) spam, instant messaging on mobile phone spam and Usenet newsgroup spam [16].Spam emails take different forms that are listed below:

- Unsolicited Advertisements [16]
- Nigerian 419 Scams [17]
- Phishing Scams [18]
- Trojan Horse Emails [19]
- Email Spoofing [20]
- Anti-Virus Spam [21]
- Commercial Advertisements [22]

- Political Or Terrorist Spam [23]
- Chain Letters [24]
- Porn Spam [25]

The email spam filtering techniques covered in this chapter each have advantages and drawbacks. In order to look for any similarities, content-based filtering compares the email content with a user profile. Filters that are content-based offer several benefits. The following is included in this: Effective use of online filtering results in fewer detectors being used, better efficiency, ease of deployment, accurate detection at a lower cost, and suitability for huge datasets, reduces the mistake Content-based filters have a few drawbacks: they produce few false positives and false negatives, need a lot of memory, necessitate frequent rule updates, produce a lot of false positives and negatives, are computationally complicated, execute slowly, demand correct datasets, and are time-consuming.

Non-content-based filters include email's non-content in their definition of spam. Filters that are not content-based offer certain advantages. This involves being quick and effective, and usable for big datasets. Filters that are not based on content have several drawbacks. This involves producing a lot of false positives and false negatives and being computationally difficult. Source-based filters create lists of email addresses and filter data by looking at the sender information in these emails. The benefits of source-based filtering are numerous. This covers the following: flawless authentication and minimal system resource consumption that lowers costs. Source-based filters come with certain drawbacks. This includes the fact that email address spoofing is simple to execute, produces many false positives, is more costly, and spreads spam backwards.

In order for collaborative filters to function, a group of users must agree onwhether an email is spam or not. The benefit of collaborative filters is that they effectively filter out spam. Collaborative filters have certain drawbacks, including the introduction of false positives and the inability to trust user trust levels.

3. APPROACHE:

Collaboration between several agents is the main emphasis of collaborative filters, which shorten the time it takes to filter out spam emails. Autonomous filters increase the workload on a single system and delay the process of filtering spam emails. Consequently, selecting collaborative filters is strongly advised for identifying and removing spam emails. Researchers are motivated to concentrate on effective collaborative filters in order to increase the security and effectiveness of email transmission by the growing demand for email usage and the absence of completely safe procedures for filtering out spam emails. These cooperative filters, however, have their own restrictions because they couldn't offer total security.

CONCLUSION:

The features of incoming emails are recovered during feature extraction after being represented as vectors using a vector space model. A soft cosine similarity metric is used to compare the similarity of two separate emails during the similarity detection process. A reputation function is created during the collaborative reputation evaluation to evaluate the accuracy of the results from various reporters.

CRVSM focuses on reputation-based detection of spam emails mainly at the receiver side in cyber space. This section describes the attacker model and the network model, section 3.2

describes the different phases of the proposed novel CRVSM protocol for detecting spam emails.

REFERENCES:

[1] P. Liu and T. -S. Moh, "Content Based Spam E-mail Filtering," 2016 International Conference on Collaboration Technologies and Systems (CTS), Orlando, FL, USA, 2016, pp. 218-224, doi: 10.1109/CTS.2016.0052. M. A. Shaik, M. Varshith, S. SriVyshnavi, N. Sanjana and R. Sujith, "Laptop Price Prediction using Machine Learning Algorithms", 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), Nagpur, India, 2022, pp. 226-231, doi: 10.1109/ICETEMS56252.2022.10093357.

[2] Mohammed Ali Shaik, Praveen Pappula, T Sampath Kumar, "Predicting Hypothyroid Disease using Ensemble Models through Machine Learning Approach", European Journal of Molecular & Clinical Medicine, 2022, Volume 9, Issue 7, Pages 6738-6745. https://ejmcm.com/article_21010.html

[3] M. A. Shaik, S. k. Koppula, M. Rafiuddin and B. S. Preethi, (2022), "COVID-19 Detector Using Deep Learning", International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022, pp. 443-449, doi: 10.1109/ICAAIC53929.2022.9792694.

[4] A. Subasi, S. Alzahrani, A. Aljuhani and M. Aljedani, "Comparison of Decision Tree Algorithms for Spam E-mail Filtering," 2018 1st International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 2018, pp. 1-5, doi: 10.1109/CAIS.2018.8442016.

[5] Mohammed Ali Shaik and Dhanraj Verma, (2022), "Prediction of Heart Disease using Swarm Intelligence based Machine Learning Algorithms", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020025-1–020025-9; https://doi.org/10.1063/5.0081719, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020025-1 to 020025-9.

[6] M. A. Shaik and Dhanraj Verma, (2022), "Predicting Present Day Mobile Phone Sales using Time Series based Hybrid Prediction Model", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020073-1–020073-9; https://doi.org/10.1063/5.0081722, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020073-1 to 020073-9

[7] Srinidhi, Jia Yan & Giri Kumar Tayi 2015, 'Allocation of resources to cyber-security: The effect of misalignment of interest between managers and investors', Decision Support Systems, July 2015, http://dx.doi.org/10.1016/j.dss.2015.04.011, vol. 75, pp. 49-62.

[8] Mohammed Ali Shaik, MD.Riyaz Ahmed, M. Sai Ram and G. Ranadheer Reddy, (2022), "Imposing Security in the Video Surveillance", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020012-1–020012-8; https://doi.org/10.1063/5.0081720, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020012-1 to 020012-8.

[9] M. A. Shaik, Geetha Manoharan, B Prashanth, NuneAkhil, Anumandla Akash and Thudi Raja Shekhar Reddy, (2022), "Prediction of Crop Yield using Machine Learning", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020072-1–020072-8;

https://doi.org/10.1063/5.0081726, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020072-1 to 020072-8.

[10] Mohammed Ali Shaik, Dhanraj Verma, (2021), Agent-MB-DivClues: Multi Agent Mean based Divisive Clustering, Ilkogretim Online - Elementary Education, Vol 20(5), pp. 5597-5603, doi:10.17051/ilkonline.2021.05.629

[11] Byrnea, DJ, David Morganb, Kymie Tana, Bryan Johnsona & Chris Dorrosa 2014, 'Cyber Defense of Space-Based Assets: Verifying and Validating Defensive Designs and Implementations', Conference on Systems Engineering Research (CSER 2014), Science Direct, vol. 28, pp. 522-530.

[12] Mohammed Ali Shaik and Dhanraj Verma, (2020), Enhanced ANN training model to smooth and time series forecast, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022038, doi.org/10.1088/1757-899X/981/2/022038

[13] M. A. Shaik, Dhanraj Verma, P Praveen, K Ranganath and Bonthala Prabhanjan Yadav, (2020), RNN based prediction of spatiotemporal data mining, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022027, doi.org/10.1088/1757-899X/981/2/022027

[14] Ashish Malviya, Glenn A Fink, Landon Sego & Barbara Endicott-Popovsky 2011, 'Situational Awareness as a Measure of Performance in Cyber Security Collaborative Work', Eighth International Conference on Information Technology: New Generations, pp. 937-942.

[15] Mohammed Ali Shaik and Dhanraj Verma, (2020), Deep learning time series to forecast COVID-19 active cases in INDIA: A comparative study, 2020 IOP Conf. Ser.:Mater.Sci.Eng. 981 022041, doi.org/10.1088/1757-899X/981/2/022041

[16] Mohammed Ali Shaik, "Time Series Forecasting using Vector quantization", International Journal of Advanced Science and Technology (IJAST), ISSN:2005-4238,Volume-29,Issue-4 (2020), Pp.169-175.

[17] Arunabha Mukhopadhyay, Samir Chatterjee, Debashis Saha, Ambuj Mahanti & Samir K Sadhukhan 2013, 'Cyber-risk decision models: To insure IT or not?', Decision Support Systems, http://dx.doi.org/ 10.1016/j.dss.2013.04.004, Volume 56, December 2013, pp. 11-26.

[18] Mohammed Ali Shaik, "A Survey on Text Classification methods through Machine Learning Methods", International Journalof Control and Automation (IJCA), ISSN:2005-4297,Volume-12,Issue-6 (2019), Pp.390-396.

[19] Andreas GK Janecek, Wilfried N Gansterer & Ashwin Kumar, K 2008, 'Multi-Level Reputation-Based Greylisting', in proc. of Third International Conference on Availability, Reliability and Security ARES 08, 4-7 March 2008, Barcelona, Spain.

[20] Mohammed Ali Shaik, P. Praveen, T. Sampath Kumar, "Integration and application of Fog, IoT and Edge Computing", Fog Computing: Concepts, Frameworks, and Applications (FCCFA) Aug-2022, CRC Press, ISBN: 9781003188230.

[21] Praveen, P, Mohammed Ali Shaik, T. Sampath Kumar, Choudhury T, "Smart Farming: Securing Farmers Using Block Chain Technology and IOT", Aug-2021, EAI/Springer Innovations in Communication and Computing, ISBN: 978-3-030-65690-4

[22] M. A. Shaik, "Protecting Agents from Malicious Hosts using Trusted Platform Modules (TPM)," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 559-564, doi: 10.1109/ICICCT.2018.8473278.

[23] Amani Mobarak & AlMadahkah 2016, 'Big Data In computer Cyber Security Systems', IJCSNS International Journal of Computer Science and Network Security, vol. 16, no. 4, pp. 56-65.

[24] lan Gray & Mads Haahr 2004, 'Personalised, Collaborative Spam Filtering', in proceedings of the First Conference on Email and Anti- Spam (CEAS), Mountain View, CA, USA, July-August, under grant no. CFTD/03/219.

[25] Aakash Atul Alurkar; Sourabh Bharat Ranade; Shreeya Vijay Joshi; Siddhesh Sanjay Ranade, Piyush A Sonewar, Parikshit N Mahalle & Arvind V Deshpande 2017, 'A proposed data science approach for email spam classification using machine learning techniques', Internet of Things Business Models, Users, and Networks, pp. 1-5.