ISSN: 0975-766X
CODEN: IJPTFI
Available Online through
www.ijptonline.com
Research Article

A CONTENT-BASED SPAM FILTERING APPROACH VICTIMISATION ARTIFICIAL NEURAL NETWORKS

B.Sundarraj[1], Dr.K.P.Kaliyamurthie[2]
Assistant Professor, Department of CSE, Bharath University, Chennai[1]
Professor, Department of Computer Science and Engineering, Bharath University, Chennai[2]
Email: sundarrajboobalan@gmail.com

Received on: 15.10.2016 *Accepted on: 22.11.2016*

Abstract

As the quality of the web exaggerated, electronic mails (emails) became a really common and convenient medium for daily communications. The spam outlined as uninvited business bulk emails, or uninteresting emails has vulnerable the web security and email services. Since the spammers perpetually improve their techniques to compromise the spam filters, building a spam filter which will be incrementally learned and custom-made became a lively analysis field.

1. Introduction

Today, laptop has replaced all suggests that of ancient communication considerably. Several distant communication tools claim to be interactive, however few offers two-way communication. (Email) “Electronic Mail” is that the most widespread suggests that of communication medium today. Email and net square measure effective tools for interaction yet on create a bridge of communication between individuals [8].

Email security may be a priority concern for several organizations. There square measure numerous threats to email security. Email security is vulnerable by a spread of problems. One amongst the foremost publicized and high risk of all problems is spamming [4].

Internet users dissent wide in what they understand to be spam. Some users contemplate all advertisements, jokes and chain letters or even all unwanted messages to be spam, whereas others try and outline it in terms of existing acceptable used policies or network prescript rules. Leung defines spam as uninvited email messages or news articles sent in bulk to recipients while not their permission. The Center for Democracy and Technology uses a broader definition and refers to spam simply as junk. Junk mail will consist of jokes and chain letters from business colleagues, friends and family. Solkin identifies the 2 most common definitions of spam as being (UCE) “unsolicited business email” and (UBE) “unsolicited bulk email” [5].

Spam emails square measure flooding networks merely as a transmitter will create a profit or accomplish its narcissistic functions by causation spam emails. but spam emails do hurt to others, each networks and human society. Spam emails price individuals time and cash, cause the consumptions of computing and network resources, degrade the network performance, and result in lots of security issues from the networks [11].

1.2 The Evolution Of Spam

The era of spamming began in 1978 and lasted till the center of the 90’s. At the primary years, spam was sent manually. Spamming required immense quantity of human resource then it didn't reach many users. Spammers wont to send the messages one by one and used inner address lists of little communities [6].

Spams were distributed over newsgroups. though the terribly 1st incident is controversial, 2 events square measure aforementioned to mark the start of uninvited bulk mailings. the primary came about in Gregorian calendar month 1994 once a young worker at a Michigan Christian school sent a message to almost 5000 newsgroups asserting the approaching of Christ. The second followed on its heels once 2 lawyers, Laurence Canter and Martha Seigal, sent a message to some 7000 teams giving facilitate getting a positive identification. The provide concerned the completion and mailing of forms obtainable to anyone freed from charge from the United authorities [4].

2. Spam-Based Dataset Preprocessing

The purpose of preprocessing is to remodel the e-mail messages into an even format which will be understood by NN. The preprocessing includes 2 halfs: the primary elective half is options extraction method and therefore the second default part is standardisation method. the subsequent subsections illustrate every half in details.

3. Options Extraction

Extract specific options from the spam-based dataset is another essential operation as a result of the dataset contains lots of redundancy, and extracting specific options from lots of fifty seven options eliminates this redundancy and solely the distinguishedoptions square measure unbroken. PCA is employed during this thesis to extract the distinguished options from the spam-based dataset. rule (3.1) outlines however PCA is applied to extract distinguished options.

4. Error Signal Analysis

The performance of the BP and OBP structures might be evaluated reckoning on the error signal against range of epochs. depicts the performance in terms of error signal of 4 BP Structure. BP-1 takes ten epochs to get the suitable error signal (1.94886), BP-2 takes nineteen epochs to urge acceptable error signal (2.9632), BP-3 takes twenty five epochs .

5. Conclusions

The work represented during this thesis considerations the appliance of 2 totally different techniques of neural network (BP and OBP) to filter incoming email messages. The filter distinguishes the spams from legitimate emails. The neural network ought to be trained to be able to distinguish between spams and legit emails. 9 teams coaching of coaching} dataset square measure employed in training section and 3 testing dataset teams square measure wont to appraise the performance of the planned spam filtering. Also, four totally different feature set sizes (57, 42, and 14) square measure employed in BP and OBP to point out the impact of feature set size on the performance of planned spam filtering. In general, the results of the planned spam filtering approach supported OBP square measure higher than the results of BP. The coaching dataset size and testing dataset size play a major role on performance on the planned spam filtering. Increasing the coaching dataset size will increase the flexibility of neural network.

6. Future Works

Several points haven't been enforced that can be self-addressed in future researches and experimental works.

1. The present work can be derived on a true client-server system. this could contemplate real emails and the way to extract correct feature set for coaching and testing. Moreover, it ought to contemplate computation time overhead required to separate spams from legitimate emails.

2. One amongst the long run works is a way to contract the setting of parameters so as to create out a correct NN structure rather than path error setting.

References

1. D. Anderson and G. McNeill, "Artificial Neural Networks Technology", a DACS progressive Report (1992).
2. W.A. Awad, And S.M. Elseuofi, "Machine Learning ways for Spam E-Mail Classification", International Journal of computing &amp; data Technology (IJCSIT), Vol. 3, No. 1, (2011). [Ben05] B. Bencsáth, I. Vajda, "Efficient Directory Harvest Attacks",Proceedings of the 2005 International conference on cooperative Technologies and Systems, pp. 62-68., IEEE laptop Society, (2005).
3. J. Chen, C. Guo, "Online Detection and interference of Phishing Attacks", Communications and Networking in China (2006).
4. W. H. Cheung, "Neural Network power-assisted Aviation Fuel Consumption Modeling", M.Sc. thesis, college of the Virginia technical school and State University (1997).
5. W. Chigona, A. Bheekun, M. Späth, S. Derakhashani and J. Van Belle, "Perceptions on SPAM in a South African context", fifth International Business data Management Conference, pp. 283 - 291, (2005).
6. P. Cocca, "Email Security Threats", version 1.4b, SANS Institute (2004).
7. M. Cofield, "Symantec Norton Anti-Virus nine.0.1for Macintosh transfer, Installation, and Basic Use Tutorial", based mostly on nformation found at:
8. M. Davy, "Feature Extraction for Spam Classification", M.Sc. Thesis, Department of computing, University of Dublin, Trinity school (2004).
9. A. Dave, M. George, "Artificial Neural Networks Technology", Kaman Sciences Corporation (1992).
10. T. Grance, J. Hash, S. Peck, J. Smith, and K. Korow-Diks, "Security Guide for Interconnecting data Technology Systems" , laptop Security Division data Technology Laboratory National Institute of Standards and Technology Gaithersburg, (2002).