ISSN (Online): 2455-7838
SJIF Impact Factor (2015): 3.476

EPRA International Journal of

Research & Development

(IJRD)

Volume:1, Issue:7, September 2016

Published By :
EPRA Journals

SJIF Impact Factor: 3.476 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 1 | Issue : 7 | September | 2016

PERSONALIZED AND EFFECTIVE SPAM FILTER BASED ON SOCIAL NETWORK

Bhagyashri A Shetage[1]
[1] ME(student)Computer Science and Engineering,
Ashokrao Mane Group of Institutions, Vathar,Kolhapur,Maharashtra,India

Amol B Rajmane[2]
[2] Computer Science and Engineering,
Ashokrao Mane Group of Institutions,Vathar,Kolhapur,Maharashtra,India

ABSTRACT

Spam is the abuse of electronic messaging systems to indiscriminately send unsolicited bulk messages. It forces users to search legitimate emails in the inbox which is filled with unwanted spam mails. In order to increase detection accuracy spam filter should be attack-resilient, personalized and user-friendly. To achieve the personalized feature proposed system are use which has Social Network Aided Personalized (SOAP) and effective spam filter. Personalized means an accurate spam filtering should be done by considering social parameters of a particular individual. Proposed spam filter will first consider the relationship between correspondents and secondly it will consider different area of interest/disinterest of individual. The intention of this paper is to prove the ability of proposed system. Social network based spam filter (SOAP) are much better than existing spam filter because of its personalized feature.

KEYWORDS: *SOAP, Bayesian filter, Spam Emails, Social network*

I. INTRODUCTION

The Bayesian filter is a step ahead of static keyword filter due to its unique ability to continuously tackle new spam by learning keywords in new spam emails. The reason Bayesian filter is so intriguing is its high accuracy rate and low false-positive rate, making it one of the most useful anti-spam tool.

The proposed system has Social Network Aided Personalized (SOAP) and effective spam filter which is accurate and user friendly. By using SOAP, instead of building blacklists and focusing on parsing keywords, it exploits the social relationship among email correspondents to detect spam adaptively and automatically. SOAP integrates four components into basic Bayesian filter: social closeness-based spam filtering, social interest-based spam filtering, adaptive trust management and friend notification. SOAP can greatly improve the performance of the accuracy, attack resilience and efficiency of spam detection.

Personalized means an accurate spam filtering should be done by considering social parameters of a particular individual. Proposed spam filter will first consider the relationship between correspondents and secondly it will consider different area of interest/disinterest of individual, for example an email about “cricket” is not spam to cricket fans, but is spam

to those who are not interested in cricket. Thus to express a legitimate mail is different from person to person and making it user-friendly according to user's requirements.

II. RELEVANCE

Especially spam is assuming as alarming proportions due to the fact that genuine emails get buried in the horde of Spam. Bayesian spam filters calculate the probability of a message being spam based on its contents. Unlike simple content-based filters, Bayesian spam filtering learns from spam mails and good mail resulting in a very robust, adapting and efficient anti-spam approach that returns hardly any false positives. In order to prevent inboxes from getting flooded with spam emails, these techniques are used which is Bayesian Filtering and Gzip Compression Technique. The gzip algorithm was compared to the Bayesian algorithm using the Spam Assassin Corpus so based on research for accuracy and speed, with fixed corpus sizes on the same machine Bayesian Filtering proved to be better in terms of accuracy and speed.

Previously spam filter approaches can be branched into two major categories: (1) Content based spam filters in which emails are separated on the basis of keywords and patterns which are typical in spam. (2) Identity based spam filters target on the identities of email senders. User manually creates a blacklist and whitelist of addresses of the emails and accordingly the emails are distinguished. Thus, the previous spam filters are not user friendly, they required much user efforts to manually differentiate spam from legitimate emails for training. Therefore, a spam filter should be attack-resilient, personalized and user-friendly to provide a secure and a reliable system

III. SYSTEM OVERVIEW

A. PROPOSED WORK

The functionalities of proposed system are divided into following parts:

1. Keyword Parsing – When a client machine receives an emails, proposed system parses out the keywords from the mail and after parsing the keywords, Bayesian filter assign each keyword weight accordingly.
2. Social Data Extraction – To identify the social relationship between the two clients/users, their social data needs to be extracted from online social network.
3. Spam Filtering – The spam filtering is further divided into four more functionalities, social-closeness based spam filtering, social-interest based spam filtering and adaptive trust management, friend notification.
4. Training – The accurate results from SOAP become training data to automatically train the Bayesian filter, thus making the filter user-friendly and personalized.

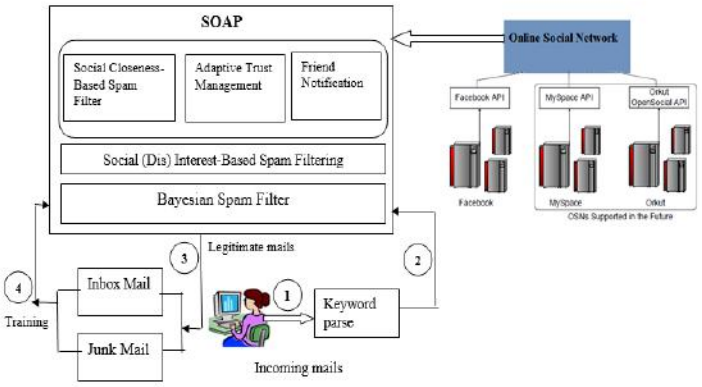

Fig 1: System Architecture

B. METHODOLOGY

SOAP integrates with following four components:

(1) Social closeness-based spam filtering-

It calculates the node closeness based on social relationships. Since nodes with higher closeness have a lower probability of sending spam emails to each other, emails from nodes with lower closeness are checked more strictly and vice versa. This component makes SOAP resilient to poison attacks.

Algorithm 1: Fetch Social Friend List to calculate social closeness

Input: User Name (UN) and Password (PW) for login social account.
Output: List of Friend_Names [LFN]
Step 1- call social media API with UN & PW.
Step 2-Get the list of friends(LF)
Step 3- For each friend(F) in LF
Step 4- Get name from F
Step 5- Add F into LFN
Step 6-Next
Step 7-Return LFN.
Step 8- End.

(2) Social interest-based spam filtering-

It calculates the user area of interest based on likes or post which is fetch from social profiles. This information helps the filter to enhance the accuracy of spam detection by considering individual preferences. This component contributes to the personalized feature of SOAP.

Algorithm 2: Fetch Social Post or Comments to calculate social Interest.

Input: User Name (UN) and Password (PW) for login social account.
Output: List of Post [LPS]
Step 1- call social media API with UN & PW.
Step 2-Get the list of Post(AP)
Step 3- For each Post(P) in AP
Step 4- Remove all stop words
Step 5- Add P into LPS

Step 6-Next
Step 7-Return LPS
Step 8- End.

(3) Adaptive trust management-

In order to tackle impersonation attacks, SOAP relies on the additive-increase/ multiplicative-decrease algorithm (AIMD) to adjust the trust values of nodes. The trust value is used to tune closeness values in order to block emails from low-trust nodes or normal nodes impersonated by spammers.

(4) Friend notification-

In order to strengthen SOAP's capability to combat impersonation attacks, a node quickly notifies its friends and friend of friend about a detected suspicious compromised node.

C. SYSTEM IMPLEMENTATION

I) Steps to Implement Bayesian Spam Filter

1) Split e-mail in tokens.
   a) Need number of e-mails for detection spam and legitimate mails.
   b) Need frequency of each word for each type based of Social Closeness and likes on social profile. This system use social network API to collect the likes and social relationship with the email sender and receiver.

2) Calculate probabilities
   i. P (legitimate) = Social friends/ Likes/post
   ii. P (spam) = Social friends/ Likes/post

3) Calculate likelihood of being spam (spamicity) using a special form of Bayes' Rule
   Where likelihood = a / (a + b), where a is the probability of a legitimate word / Social Closeness/ Likes and b is the probability of spam word./ Social Parameters

4) calculate spam keywords probability by using Baye's Rule.
   Choose tokens whose combine probability is farthest from 0.5 either way. This is because the farther it is from 0.5 (neutral), with more certainty we can say it belongs to either strategy.
   i. Do this for n numbers for instance choose to have 15 limits?
   ii. Combine their probability to get a figure for message using Bayes' Rule. In basic terms, Baye's Rule determines the probability of an event occurring based on the probabilities of two or more independent evidentiary events. For three evidentiary events a, b, and c, the probability is equal to:

$$\frac{a\ b\ c}{a\ b\ c + (1-a)*(1-b)*(1-c)}$$

   In this fashion, the rule can be expanded to accommodate any number of evidentiary events.
   iii. If the end result is closer to 1.0, then the message is classified as legitimate mail and if it is closer to 0.5 then the message is classified as spam mail. The cut-off range that has been specified for spam is that it should be less than 0.5, but spam results are above 0.4.

5) Generate final Result of percentage for spam or regular mails.

II) To design SOAP system with Bayesian spam filter combine all four steps using following algorithm. Based on probability of spam keywords, if the mail is specified as spam mail then you can check with another factors getting from social site which is closeness-ship and area of interest. If probability value of these factors is more, then system can decrease the probability value of spam mail and that mail declares as regular mail which is came into inbox.

Algorithm 3: Implement SOAP System with Bayesian Spam Filter

Input: Email List (EL).
Output: List of Spam mails [LS] and List of Regular mails [LR].
Step 1-Call social closeness and interest based function.
Step 2-For each Emails (E) in (EL).
Step 3-Get message (M) from E
Step 4-Check spam keyword probability using basic Bayesian algorithm
Step 5-IF Not a spam mail then
Add E into LR
Step 6-Else IF E is from LFN then
Add E in LR
Step 7-Else IF E subject are match with LPS then Add E into LR
Step 8- Else Add E into LS declare as spam mail.
Step 9- End IF.
Step 10: Return LR and LS.

IV. EXPERIMENTAL RESULTS

In our system we are checking SPAM mail by using different techniques.

1. Checking by social account like Facebook Friend List
2. Checking by social account like Facebook Post
3. Checking by contents wise
4. Checking all together (contents, Post, Friend List)

When we checked above techniques, total mails were in inbox are 132.The result we get is as follows

Out of 132 emails	SPAM Email	REGULAR Email
Friend List	117	15
Posts	97	35
Contents	95	67
All together	76	56

SO the graph is as shown below.

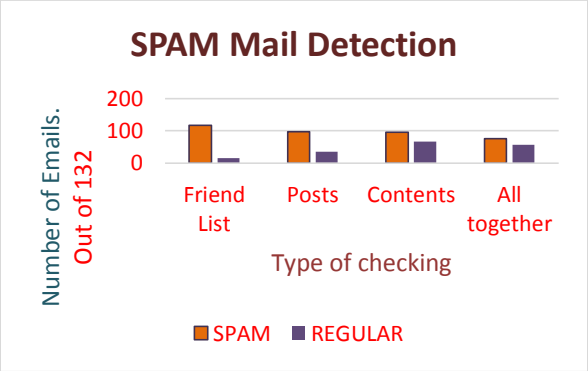

Fig 2- Graph of spam mail detection

From the graph it clearly shows that the last technique is best technique to detect the SPAM mail. In this technique we are checking first content's then is mail is spam we check for Subject, if it is spam then we check for Friend list else the mail is spam.

V. CONCLUSION

In order to prevent inbox from getting flooded with spam emails spam filters are used. The SOAP improves the performance of Bayesian networks in term of spam detection accuracy and training time. SOAP is a personalized and user-friendly social network based Bayesian spam filter. SOAP can build result of spam or regular mail detection based on like comments posted on walls of user social networking profile. SOAP can be used as a plug-in in current on-line social networking sites such as Facebook, MySpace, Twitter. In this paper the complete system review are given to help out for finding unwanted spam mails from inbox.

REFERENCES

[1] *Leveraging Social Networks for Effective Spam Filtering HaiyingShen, Senior Member, IEEE, and Ze Li, Student Member, IEEE IEEE TRANSACTIONS ONCOMPUTERS, VOL. 63, NO. 11, NOVEMBER 2014*

[2] *Ze Li and HaiyingShen ,"SOAP: A Social Network Aided Personalized and Effective Spam Filter to Clean Your E-mail Box" ,IEEE infocom 2011.*

[3] *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 3, March 2015 ACCURATE KEYWORD BASED SPAMS FILTERING IN SOCIAL NETWORKS.*

[4] *Al-Khwarizmi Engineering Journal, Vol. 6, No. 2, PP 83-92 (2010) Software Engineering-Based Design for a Bayesian Spam Filter Mumtaz Mohammed Ali AL-Mukhtar.*

[5] *BAYESIAN FILTERING INTRODUCTION TO Using Bayes' Formula to keep spam out your Inbox.*

[6] *Statistical Spam Filtering David Anderson EECS595, Fall 2006.*

[7] *A review of machine learning approaches to Spam filtering Thiago S. Guzella *, Walmir M. Caminhas Expert Systems with Applications 36 (2009)*

[8] *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 3, March 2015. ACCURATE KEYWORD BASED SPAMS FILTERING IN SOCIAL NETWORKS.*

[9] *Naïve Bayes classifier[Available:http://en.wikipeda.org/wiki/Naive_Bayes_classifier.*

[10] *A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm,[Available:http://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest algorithm.*