Volume 8, No. 5, May-June 2017

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Review on Different types of Spam Filtering Techniques

Pooja Revar, Arpita Shah, Jitali Patel & Pimal Khanpara
Institute of technology,
Nirma University,
Ahmedabad, India

***Abstract*:** Nowadays the use of internet is increasing; there are also different type sites like social networking sites, blogs, and different email services, different portals are there. Using that user can share the different types of information. Here information does not means that it can be always good information, but some time s it may be harmful that we can call as spam. This spam can be redirects the user to the different other pages and also it can make user fool, it can also be helpful for spreading harmful contents. In this paper we have make review on the different spam filtering techniques and different works which has been done in this area.

***Keywords*:** Spam filter, Spam URL, UBSF, analysis of spam

I. INTRODUCTION

Spam email has become a serious problem today with the popularity of Internet and network services [1]. Spam emails are not desirable because i) they waste a huge amount of network resources which are very important for the users ii) they highly affect the daily work of many users. People have to waste a significant amount of time to deal with spams every day. Also they give rise to problems like private information leaking, malware infection and one click fraud. In worst cases, malicious attachments contained by spam mails crack the system of the user. People have been struggling with spam for about 10 years. Though many techniques have been suggested by researchers, it seems like the problem has not got any effective solution yet. According to a study, people receive more than 50% of spam mails on an average. Moreover, the percentage of spam messages received is increasing exponentially. Thus, it is very impoant to deal with spam [1].

There is some unsolved problems in the spam filtering can be given as below:

(1) There is still problem names “false negatives” and “false Positives” is unsolved and in that false positives problem is bigissue[1].

(2) Spam can be relative problem for the people, some link or mail can be spam for some people but some people does not consider that as spam so the spam can be relative to the people and the link which it redirectsto[1].

(3) In the different techniques consider content based techniques for filtering spam contents we requirehighcomputational cost to perform their function until whole email is received. It consumes more time and it cannot be much useful in online processes in which there is heavy internettraffic[1].

II. DESIGN GOALS AND DIFFERENTAPPROACHES

For the spam filtering techniques now days the design goals can be give as below.

(1) Results in Real-time: some services like social networking and many others are working I real time. So it is needed that spam filtering can be done with smalldelay.

(2) Accuracy of decision: the system or technique should give accurate result within the time in order to mistake minimization of non-spamURLs.

(3) Classification should be context independent: Classifier should allow services for different webservices.

(4) Fine-grained classification: The system should be easily recognize a difference between spams, which is hosted on public-services with ‘non-spamcontent’.

There are various spam filtering techniques can beconsidered now days, major of them are described asbelow.

(1) Based on list using email-addresses: DNS address and ‘ip address’ two types of list names white list and black lists are created first. Then, this list will useful for filtering spam content.one of the list names real-time black lists is used based on thismethod[1].

(2) Techniques based on the content: In this technique different type of machine learning algorithms like ‘Support Vector Machine (SVM)’ ‘Naïve Bayesian Classifier’ can be applied over the content to filter spam Contents.in this type of technique large amount of research work has been done it gives successful result but still some problems there e.g. it is more time consuming and sometimes there may be less accuracy inresults[1].

(3) Techniques based on set of rules: In this technique some weighted rules sets are defined for the emails parts. If the emails triggers it then the score of that rules are calculated and added. Here one threshold value is also defined, if calculated value is more than threshold then the mail is identified as spam.one method names ‘spam assassin’ is classic filter method which we can put in thiscategory[1].

(4) Technique based on URL: This is most recently used technique in spam filtering.in this technology different researchers have done different works for differentiate wanted and unwanted email links. .In the different type of this approach 24/7 spam filtering is done and by that and reviews are made for the systems. In this paper we have given idea of how different system is works in the area of spam filtering techniques. There is also some software like Spatmo. There is also spam filters names: Early Grey Filter ,Domainator and Razor filters working for this area [1]. Domainator is now days working on this approach with Google’sdatabase.

(5) Other techniques: ‘client puzzle approach’ is one of the technique the spammers works on “fire and go” approach

so this technique can determine the spammers. In this approach sender requires effort to deliver a message to one particular recipient. Moreover, sender have to solve puzzles after that confirm messages. So, human involvement needed to complete this work prevents the mechanism from being widelyaccepted.

In this paper we are going to discuss different techniques as below.

III. DIFFERENTSYSTEMS

A. Stasticalanalyzingapproach:

Below the system design of this approach is given inn Fig 1.

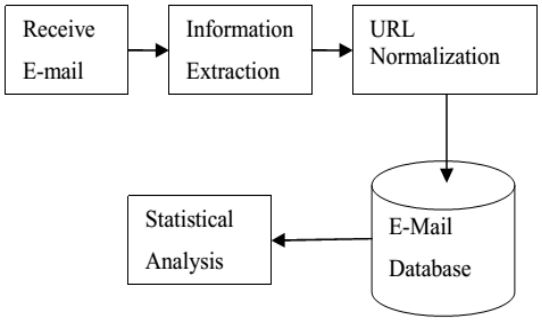

Fig 1: Analysis of spam technique

This filtering techniques is useful for the e-mail services. Filtering of spam techniques steps are as below:

STEP 1:Information extraction phase

There is mainly two parts in the e-mail (1)HTML tag Part (2) plain texts.URL is mainly located in the HTML part only using tags like SRC tag, 'ACTION' tag, 'BACKGROUND' tag.

STEP 2: URL Normalization

Main work of this task is convert the URL in the canonical forms, by that it can become easy to detect equivalent but lexically different URLs. It can be effective for reducing the

URL database libraries made for the Spam filter processing and also helpful in reducing the space storage of db.

STEP 3: Statistical Analysis

In this task main work is done at statistic levels, this task can be as below.

(1) Analysis of HTML tagging: mainly URL are inserted in the email with help of argument of different html tags. This tags can be 'SRC','HREF','ACTION' etc. so mainly tag related analysis isdone.

(2) Dynamic or Static: the URL found in the email can be dynamic or static. By this dynamic URL is directed to the different database, and also the page can be also generated dynamically using different types of scripts. Mostly commercial sites and spam related sites are found dynamic during earlier research works. Example: http://www.wooha.com/?T.mc_id=INAV100120-1h.[2]

(3) Origin of URLs: According to the paper [2] it is found that, if host component can be checked of any sites URL than it found that spam-advertised sites are using short- lived 'ip address' so that it cannot be found easily that it is running for spam relatedworks.

(4) Path length of URLs: as we know URL path is made in such hierarchal manner. On particular site we can got to particular page of the site using this path. If the path length is 1 or 2 than it is easy to remember.as in paper [2] described, there is 63% spam-advertised URL only of 1 to 2 pathlength.

Summary and Outcomes of the paper

By tis paper the statistical analysis of the spam related work done successfully and it gives analysis as below

(1) Majorities from the spam messages are encoded in the html bodies, more than 95% are encoded with help of HREF and SRCtags

(2) Most of the URLs are corresponds to the static data, and minority being up to 22.69% aredynamic.

(3) About 63% URL which are related to spam are found have length of URLPath=3.

B. UBSFTECHNIQUE

Here the working diagram of UBSF is as n Fig 2 found in paper [1]. The working of different blocks are described as below.

(1) **Decoder:** emails are encoded in Base64, Unicode and different other format. The decoder are use to decode thatinto the normalform.

(2) **E-mail processors:** There is tags any Meta tags areused in the emails so that should be processed first for filter process. So using this meta-information can be extracted, which destined to the fake or spam relatedaddress.

This is the half part of the system now how mainly system works in system identifier is described.

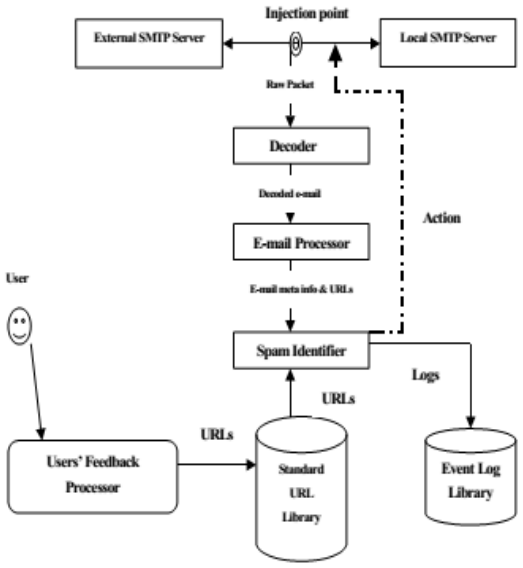

Fig 2: UBSFSystem Flow

The Fig2 shows different part of the UBSF system.

(3) **Spam-identifier:** this is the main block of the systemuses the LCS algorithm. Useful for make SUL that is 'Standard URL LIBRARY'. It also creates event log libraries using XML formats.it also send command to the SMTP servers.it can also make log for false negative and false positives evaluation for filteringspams.

(4) **User feedback Processor:** As we know some of the

links cannot be spam for some users, but it can be spam for some user so if the spam is found by the user. The feedback is given to SUL and saves according to thecalculation

Result analysis of the technique

sFirstly as seen this paper [1] false positives of UBSF and precision rate of 'naïve bayes' classifier based system and SVM on mailing system is compared[4]. For comparing two main precision and recall terms were used.

Here *Nl→l ,Ns→s,* denote the no of legitimate messages and spam mess rectly. *Nl→s*classifi *ls→l*is defined as leg

$$Precision(p) = \frac{Ns \to s}{Ns \to s + Nl \to s}$$

Here table 1 be veen the different techni

$$Recall(r) = \frac{Ns \to s}{Ns \to s + Ns \to l}$$

Table 1: comparison between different techniques of spam filtering

Approach	Precision	Recall	Time (Seconds)
Naïve Bayes	97.18%	94.83%	179.32
SVM	98.79%	96.33%	199.48
UBSF	99.30%	98.42%	24.92

C. *MONARCH :Real Time Spam FilteringSystem.*

In this section the description of the one real time system is given

The simple flow diagram of any spam filter system can be given as below fig:

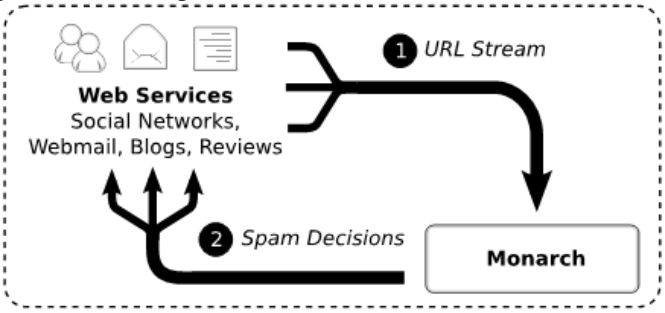

Fig 3: simple diagram of monarch system work

The system flow of the monarch system is given as below in Fig 4.

The systems' different parts are described as below.

(1) URL Aggregation: Fig shows flow of monarch system.In the system firstly 'Dispatch-queue' mechanism is used in the 'URL aggregation' block. In this firstly different email streams and different URLs are collected from different sources and firstly it is put in thequeue[5].

(2) Feature collection: Then it comes to the second block of the system called 'feature collection'. In this block different types of data are collected. The data here means features related to site sand URLs, like if site uses java script, which site the URL redirects, and many others[6]. Here in thisstructure or system cloud environment can be alsoused.

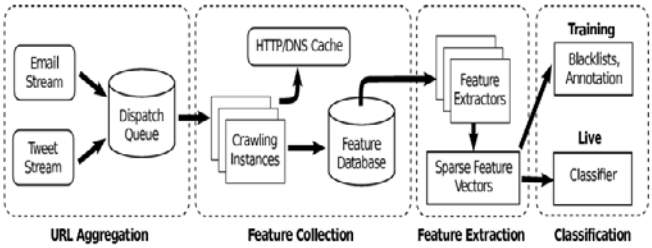

FIG 4: The System Architecture of Real Time Spam filter MONARCH

(3) Feature Extraction: The output of the 'feature collection' block used as input for this block. Raw data generated during data collection block converted into a feature vector block which can be understand by the classifier block. In this URLs are converted into binary feature block, also Bag of words are made using the html contents. Here the data which is used as input stored temporary. In this URLs are also use canocalize to present 'ip address' into hex code and check thee path traversal operation in the URL path. This hex code easily can be converted into binary form here. Html contents are also used here in same manner and here different text will be converted or here we can say that it will be tokenize. By this way sparse hash map or vector will be generated, which will be used by the next classificationblock[3].

(4) Classification: It is the final phase of system flow to make classification decision. Classifier training occurs offline, which is not depends on the main system, make it-self free from live decision as simple summation of weights of classifiers. In one phase names training, labeled data set by taking URL found during the phases above can be put in the black-list or traps as spam. For make good decision we train systemdaily[3].

In the phase different algorithm are used are as below

Algorithm 1 Distributed LR with L1-regularization

Input: Data $\mathbf{D}$ with m shards
Parameters: λ (regularization factor), I (no. of iterations)
Initialize: $\vec{w} = \vec{0}$
for $i = 1$ to I **do**
  (gradient) $\vec{g}^{(j)} = \text{LRsgd}(\vec{w}, \mathbf{D}_j)$ for $j = 1..m$
  (average) $\vec{w} = \vec{w} - \frac{1}{m}\sum_{j=1}^{m} \vec{g}^{(j)}$
  (shrink) $w_\alpha = \text{sign}(w_\alpha) \cdot \max(0, |w_\alpha| - \lambda)$ for $\alpha = 1..d$
end for

Algorithm 2 Stochastic gradient descent for LR (LRsgd)

Input: $\vec{w}$ (weight vector), $\mathbf{D}_j$ (data shard)
Parameters: η (learning rate)
Initialize: $\vec{g}_0 = \vec{w}$
for $t = 1$ to $|\mathbf{D}_j|$ **do**
  Get data point $(\mathbf{x}_t, y_t)$ from $\mathbf{D}_j$
  Compute margin $z = y_t(\vec{x}_t \cdot (\vec{w} - \vec{g}_{t-1}))$
  Compute partial gradient $\vec{h} = y_t([1 + e^{-z}]^{-1} - 1)\vec{x}_t$
  $\vec{g}_t = \vec{g}_{t-1} + \eta\vec{h}$
end for
Return: $\vec{g}_{|\mathbf{D}_j|}$

Herefirstlyusingalgorithmondifferentweightvector $\vec{w}$ willbe generated. This will be used in the training algorithm given as algorithm 2.

Training algorithm:
For this firstly training data generated by algorithm 1 is divided into the m shards (this may be done on the different filesystemlikehadoop).Afterthatinitialmodelvector $\vec{w}$ is distributed over mshards.
This vector is updated using algorithmDuring this whole process. Here work of shrink step is done in algorithm 1 and then it used in algorithm 2's step.

Run time Performance of system:
This monarch frame work is firstly run on the amazon server for implementation purpose for twitters contents spam filtering. This work is gives result and performance of this as given in below table in accordance to the time.

The table 2 and table 3 are made as below[3]:
Table 2: Time taken by system

Component	Median Run Time (seconds)
URL aggregation	0.005
Feature collection	5.46
Feature extraction	0.074
Classification	0.002
Total	**5.54**

Table 3: Cost for the system for spam filtering process

Component	AWS Infrastructure	Monthly Cost
URL aggregation	1 Extra Large	$178
Feature collection	20 High-CPU Medium	$882
Feature extraction	—	$0
Classification	50 Double Extra Large	$527
Storage	700GB on EBS	$70
Total		**$1,587**

IV. CONCLUSION

In this Review Paper, we have discussed different technique related to spam filtering. There are many techniques now days available and different tools are there working in this area, but still it is hard to find the spam content because one type of spam content may not be sometimes spam content for one user. So we have to make improvement on it, in this paper we have discussed one technique 'UBSF' can be very useful but still not only spam content can be detected using this because the content may be harmful. So we have to do more research work in content based and fast spam filter technique.

V. REFERENCES

[1] Li, Y.,Fang, B. X., Guo, L., Tian, Z. H., Zhang, Y. Z., & Wu, Z. G. UBSF : A Novel Online URL-Based Filter,IEEE Symposium on (pp. 332-339).ISCC 2008

[2] Zhongtao, W.,Xin,P.,Yuling,W.,Yaohua,L.Li,H., & Biao,C. (2012, March). Analysis on the Characteristics of URL Spam. In Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on (Vol. 1, pp. 246- 249).IEEE.

[3] L. Pelletetier, J. Almhana, and V. Choulakian, "Adaptive Filtering of SPAM," Proceedings of the Second Annual Conference on Communication Networks and Service Research(CNSR'04),2004. W.K.Chen, LinearNetworksandSystems. Belmont, CA:Wadsworth, 1993, pp. 123–135.

[4] I. Androutsopoulos, J. Koutsias, K. Chandrinos, and C. Spyropoulos, "An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages," Proceedings of the International ACM SIGIR Conference, 2000.E. P. Wigner, "Theory of traveling-wave opticallaser,", vol. 134, pp. A635–A646, Dec.1965.

[5] Kurt Thomas, Chris Grier,JustinMa,VernPaxson ,"Design and Evaluation of a Real-Time URL Spam Filtering Service Dawn Song", University of California, Berkeley, International Computer ScienceInstitute.

[6] Blum, A., Wardman, B., Solorio, T., and Warner, G., "Lexical Feature Based Phishing URL Detection Using Online Learning". inAISec '10 Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security, Illinois, USA, 2010, ACM New York, NY, USA, pp. 54-60. DOI= http://dl.acm.org/citation.cfm?id=1866423.1866434&coll= DL& dl= ACM&CFID= 237444071&CFTOKEN=87140042.