International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

An Efficient Framework for Spam Mail Detection in Attachments using NLP

Amol Malge[1], Dr. S. M. Chaware[2]

[1]ME student, Department of CE, TSSM's BSCOER college of Engineering & Research, Narhe, Pune, India.

[2]Head of Dept., Department of CE, TSSM's BSCOER college of Engineering & Research, Narhe, Pune, India.

Abstract: ***Now a day's mail becomes a vital medium for communication as it offers a lot of advantage over SMS service. But spams also known as junks are immerged as hurdle for the email service. Spam attracting more attention as internet fields are widely affected due to such unwanted spams. Recently conducted study shows that 70% of today's business mails are spams. Therefore many serious problems such as wasting storage space, growing volume of spams etc. are immerged. The main motto behind such spams is to damage financially to the company and individual. Hence number of approaches had been proposed to protect systems from such unwanted spams, and filtering is one of them. Proposed approach is working on the Gmail host id's where it identifies the spam Emails by detecting signature of the Email with the support of the strong NLP protocols.***

Keywords : Correlation, Ontology, NLP, inverted index, hardness

1. Introduction

In today's digital era World Wide Web becomes an integral part of users. As the numbers of users are relying on WWW, spammers become more active. They got internet as a powerful way to harm the users. And among all the WWW entities spams are the best suits for the spammers. To harm the user's activity spam mails are sent by the spammers. Spam mails are the subsets of electronic mail which are junk mail or bulk mail. In spamming identical mails are sent to the numerous users at a same time without their agreement. Spam normally contains the suspicious links which are intended to steal the user data or harm the user system. There are numerous ways used by the spammers to collect the email address of the user such as chat rooms, news groups, websites etc. numbers of techniques are used for spamming as below.

- Appending.
- Image spam.
- Blank spam.
- Backscatter spam.

Numbers of anti-spam techniques are proposed still the numbers of spam messages are keeping increasing rapidly. So to get rid of the spam emails, many organizations makes use of filtering gateways, anti-spam services, end user training etc. for restriction of spam mails, spam filters can be established at any layer such as firewall, mail transfer agent, anti-virus etc.Spam filtering architecture can be shown in the below figure 1.

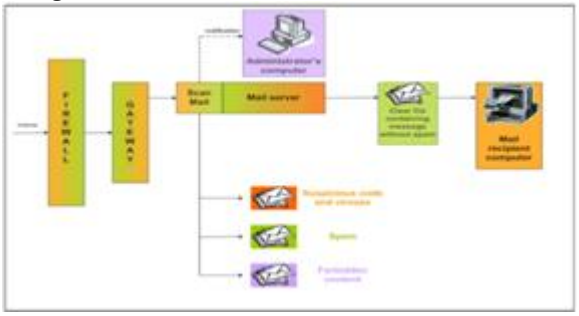

Figure 1: Spam filtering architecture.

Given below the few spam identification techniques.

1) Whitelists/Blacklists
Name itself indicates the use of two list. One is whitelist and other is blacklist. Whitelist contains all the email addresses from which healthy mails are received. While blacklist contains the mails from unintended users.

2) Mail header checking
In mail header checking, the header of mail is observed. Header such as blank subject, to many receivers, large numbers of digits in body etc. is checked to come on conclusion.

3) Bayesian analysis
Here probabilistic functions are used to scrutinize the mails on either category.

4) Keyword checking
Here both i.e. subject and body of the mail is considered for spam detection. Combinations of keywords are used for making conclusion.
Like spam identification, spam filtering approaches are there as.

A. Distributed adaptive blacklists
B. Rule based filtering:
C. Bayesian classifier
D. K nearest neighbors
E. Support vector machine (SVM):
F. Content based Spam Filtering Techniques - Neural Networks:
G. The multi-layer networks

Artificial neural network is an area to find the intelligence models in the field of computation. The main inspiration of ANN is BIS i.e. biological immune system. If the neural network is compared with the human body then we can correlate the brain of our body. Human brain takes the decisions based on the internal and external situations of human body.

Licensed Under Creative Commons Attribution CC BY

In spam filtering emails are need to read in order to classify them as a spam or not. Once the emails are read, it contains huge amount of unwanted things. So it is necessary to remove all these things before sending the data for filtration process. Once unwanted things are removed, lot of time and space is saved required for their handling. Preprocessing comprises of following sub algorithms as stop word removal, stemming and special symbol removal.

MessageLabs intelligence a tracking system, records 77 targeted spams each day by March 2010. Figure 2 is given below which clearly shows the day by day increment in the spamming rates.

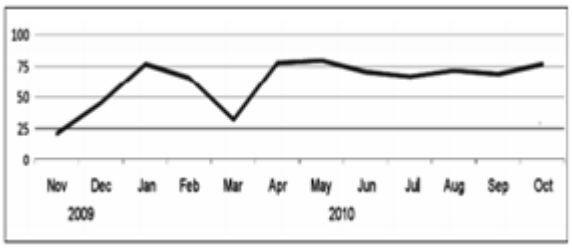

Figure 2: Targeted spam attacks 2009-2010 (Symantec. cloud Message Lab)

1978 is the year during a first spam is recorder and Peter J. Denning's is the first person who reviews it. Bays algorithm is the first technique used for the spam filtering. Table 1 compares the few techniques used for the spam filtering.

Table 1: Comparison of spam filtering techniques

Anti-spam software	FPP	FNP	FP	Detected		FN
				TP	TN	
Matador 1.0.0	4.9%	4.7%	35	410	256	256
SpamBrave	1.1%	7.0%	8	400	283	283
SpamArrow	2.5%	5.8%	18	405	273	273
Espresso 1.06.94	2.9%	20.9%	21	340	270	270
Spam Bully	2.1%	16.3%	15	360	276	276
Spam Fighter	1.9%	19.1%	14	348	277	277
Qurb 2.0	5.8%	2.3%	42	420	249	249

TH-True Negatives, TP-True Positives, FN-False Negatives, FP-False Positives.

In case of spam filtering techniques relying on single algorithm for spam detection is never a good choice because single algorithm is merely incompatible to deal with all the associated issues. So for improving efficacy and accuracy of the system two or more than two algorithms are incorporated into a single technique. The plus point of using more algorithms is we can easily overcome the advantages of one algorithm in other. The systems which combine few algorithms are known as composite intelligent algorithms. For showing the experimental evaluation composite algorithms are compared with the existing algorithms such as blacklist, rule based, bayes algorithm etc.

The rest of the paper is organized as follows. Section 2 discusses some related work and section 3 presents the design of our approach. The details of the results and some discussions we have conducted on this approach are presented in section 4 as Results and Discussions. Sections 5 provide hints of some extension of our approach as future work and conclusion.

2. Literature Survey

To bring the output pre-processing is required on the input data. Normalization is one of such method used for pre-processing. It is the process of bringing the word to its original form i.e. root form. This can be done by the stemming process. Normally normalization is done to find and remove the suffixes attached to the words in order to find the occurrences of the same word repeatedly in specific context (e.g. going and go gives the same meaning, computing and computed also have same meaning). Again the suffixes to be searched should be known in advance. Sometimes it may possible that normalization carried by the stemming process will change the meaning of the word (computing and computed will give compute), a solution on this is to use lemmatization. But a condition that doesn't have the linguistic knowledge in prior will support the stemming as a best method.

To do so first list should be created that contains the stems i.e. suffixes. [1] Gives the six stemming and lemmatization approaches where they conclude that these approaches are best over the word based baseline. Table differentiate the six normalization approaches i.e. OpenOffice based lemmatizer, HPS Stemmer, HMM tagger, PDT 2.0 Lemmatizer, PDT 2.0 Lemmatizer and MorphoDiTa.

[2] Represents a stemming algorithm that based on the context aware. The given approach is intended to reduce the morphological variation of the input query caused because of stemming process. The stated algorithm takes well known port stemmer algorithm as a base for development. The rule based approach is used to do so. Affix removal approach [3, 4, 5, 6, & 7] is one of the good approaches used for the purpose of stemming. This algorithm is a comes under the classical approach of stemming technique. There are number of algorithms like Dawson stemmer, Lovins stemmer, Paice-Husk stemmer came under the same category. Lovins stemmer works on the principle of longest match. Dawson stemmer also makes use of principle of longest match and it replaces the recoding rule used in the Lovins stemmer to make it reliable. Paick huck stemmer is another algorithm for stemming word removal which finds out the answer in indefinite steps. Among all the above stated methods porter stemmer gain popularity because its performance over another algorithms.

[8] Illustrates the n- gram stemmer, an interesting and language independent method which makes use of string similarity approach. The basic idea behind the approach is that the similar word will have high proportions of n-grams in common. E.g. for n=2, n=3 words will be diagrams and trigrams respectively. N-gram technique is one of the common techniques used in the approaches stated in [5]. But one of the biggest disadvantage of the method is it requires the significant amount of memory. [9] Presents a new approach based on the Hidden Markov Model (HMMs)

Licensed Under Creative Commons Attribution CC BY

which are finite state automata where probability functions are used for the rules between the transitions.

$$\text{miss rate} = \frac{\text{nonspam samples misclassified}}{\text{total nonspam examples}}$$
$$\text{false alarm rate} = \frac{\text{spam samples misclassified}}{\text{total spam examples}}.$$

Yass stemmer [10] stands for Yet Another Suffix Striper and another stemming method based on the statistical and corpus method. Also it will not require the prior language knowledge and it is language independent. [7, 11, 12] tries to elaborate are some of the hybrid approaches of stemming word removal. It includes Linguistic Lexical Validation Stemmer, Corpus Based Stemmer, and Context Sensitive Stemmer. A study [5] gives a methodology of Linguistic Lexical Validation Stemmer. The main motto behind the method is to reduce the stemming errors and increase the accuracy of the overall system.

Corpus Based Stemmer was being proposed by the [9] where author tries to overcome some of the disadvantages of the well-known stemmer Port Stemmer. The biggest problem with the port stemmer is that sometime it generates the stems which are not real words. This problem is easily overcome by the stated stemmer. Context Sensitive Stemmer fins an interesting method of stemming because morphological variants necessary for the search are predicted before the query is submitted to the search engines. This experiment dramatically reduces the unwanted expansions. Also precision can be increase too much by the method.

For efficient categorization of spams i.e. healthy or infected. [13] Gives spam mail detection technique by using the text clustering methodology. For the data representation vector space model is used by the system. But the problem of vector space model is that the size of the data is increases to much so to reduce the size of the data clustering of vectored data is done. The main reason behind the use of clustering is to categorize the data based on their patterns.

[14] Narrates the multi neural based spam detection system using the spam words as a base of their operation. Here in this method neurons are trained by observing the weights obtained from the ASCII value of the characters. One care is needed to be taken before supplying the words to the classifier i.e. preprocessing should be done. The experimental evaluation of preprocessing shows that the system gives positive results. Drawback of the system is, it is implemented on the small size database. So the implementations of the system on large scale database are kept as a future work of the system. Apart from this the user feedback is need to consider as it can improve the quality of the spam word detection can be improved to the great extent.

[15] Presents support vector machine based technique for spam detection. The reason behind the use of SVM is it makes use of three different algorithms for spam detection i.e. Ripper, Rocchio, and boosting decision trees. For experimental evolution purpose the system is tested on two datasets. In the first dataset the numbers of features are restricted to 1000 where in another dataset the features are extended to 7000. The experiment shows that the SVM gives the best result when the data is binary. Also the observations show that the SVM significantly requires less training time. Finding of False Alarm and Miss Rate is accomplished by using two formulas as mentioned below.

[16] Elaborates the one more SVM based spam filtering system. Here all the problems associated with the spam scrutinization are well discussed by the author. To classify the spams online and active algorithms are used. For showing the experimental evaluations TREC2006 spam filtering benchmarked dataset is used by the author who gives a promising result. Still authors states that the numbers of drawbacks are associated with the system. And it can be overcome as the feature work of the work.

3. Proposed Methodology

In this section, we describe our framework for spam mail detection system using NLP rules with the below mentioned steps as shown in figure 3.

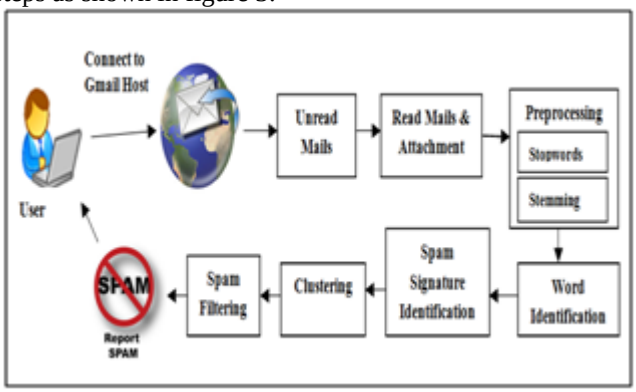

Figure 3: Overview of the proposed work

Step 1: Here in this step system accepts the user's Gmail id and password and then connects to the respective ID using Gmail host. And then reads all the unread mails. Once it read the unread mail all the content are fetched and stored in a vector and then this vector is passed to further preprocessing process to identify spam.

Step 2: This is the step where preprocessing is conducted, where string is processed to its basic meaning words by the following four main activities: Sentence Segmentation, Tokenization, Removing Stop Word, and Word Stemming.

- *Sentence segmentation* is boundary detection and separating source text into sentence.
- *Tokenization* is separating the input query into individual words.
- *Stop word removal* :In any document narration the conjunction words does not play much role in the meaning of the document, so by discarding these words (like: is, the, for, an) from the documents which greatly reduces the overhead of processing
- *Stemming:* Many of the elongated words in the English language generally fail to provide proper meaning in the given scenario and also they increases the computational time. So it is necessary to bring the words to their base form by replacing its extended

Licensed Under Creative Commons Attribution CC BY

Step 3:-Feature extraction- Term weight.-The most repetitive words in text are obviously the important words. So system identifies the list of most repeated words and considers some top n elements (where n is user defined) as the important word for text to store in vector. And this can be extract as in below shown algorithm

ALGORITTM 2: TOP WORD IDENTIFICATION

Step 0: Start
Step 1: Read string
Step 2: divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: **for** i=0 to N (Where N is length of V)
Step 5: for ith word of N check for its frequency
Step 6: Add frequency in List Called L
Step 7: end of **for**
Step 8: return L
Step 9: stop

Step 4: Spam Signature Identification: In this step all extracted top words are checking for their past occurrences and their duration of stay in the inbox of the Email. Based on these parameter a protocol is designed where Emails are been constantly checking for immediate deletion from the user and their respective top words. Based on this signature spam Email is identified and cluster them to show in a proper report.

The whole proposed system of spam mail detection can be depicted in the below algorithm 2 .

ALGORITHM 2: SPAM MAIL DETECTION

Input: // Set E = E0, E1, E2,. . . , En As set of Emails of a Id U
Output: // Spam Emails sets as S = S0, S1, S2, . . . , Sn
Step 0: Start
Step 1: Read Email data as String Es
Step 2: create a list L = Es, d, t, f
Where d=date
t=time
f=from
Step 3: Get spam words in a list SL.
Step 4: Set Interval Time as T
Step 5: if L not presents Time t Then Check L SL
Step 6: if Step 5 is true
Step 7: then Add in S
Step 8: return S
Step 9: Stop

4. Results And Discussions

To show the effectiveness of proposed system some experiments are conducted on java based windows machine. To measure the performance of the system we set the bench mark by selecting Gmail ID mails which containing attachments along with the Email body.

To determine the performance of the system, we examined how many relevant Spam Email Clusters are formed based on the NLP rule and signature identification process.

To measure this precision and recall are the best measuring techniques. So precision can be defined as the ratio of the number of relevant Spam Emails identified to the total number of irrelevant and relevant Spam Email Emails identified. It is usually expressed as a percentage. This gives the information about the relative effectiveness of the system.

Whereas Recall is the ratio of the number of relevant spam Emails are identified to the total number of relevant Spam Email identified. It is usually expressed as a percentage. This gives the information about the absolute accuracy of the system.

The advantage of having the two for measures like precision and recall is that one is more important than the other in many circumstances. In contrast, various professional searchers and intelligence analysts are very concerned with trying to get as high recall as possible, and will tolerate fairly low precision results in order to get it. Individuals searching their hard disks are also often interested in high recall searches. Nevertheless, the two quantities clearly trade off against one another.

For more clarity let we assign
• A = The number of relevant Spam Emails identified,
• B = The number of relevant Spam Emails not identified, and
• C = The number of irrelevant Spam Emails identified.
So, Precision = (A/ (A+ C))*100
And Recall = (A/ (A+ B))*100

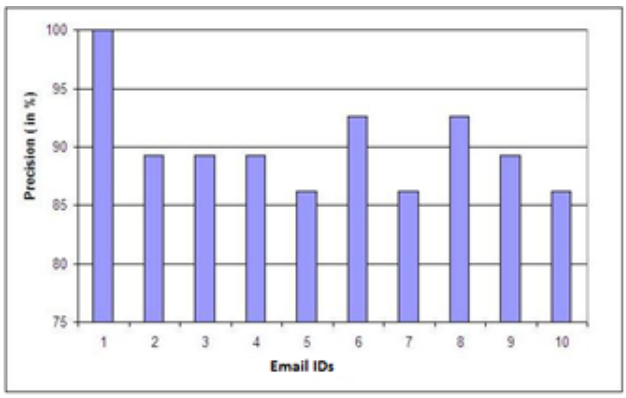

Figure 4: Average precision of the proposed approach

In Fig. 4, we observe that the tendency of average precision for the identified spam Emails are high compared to other systems.

Licensed Under Creative Commons Attribution CC BY

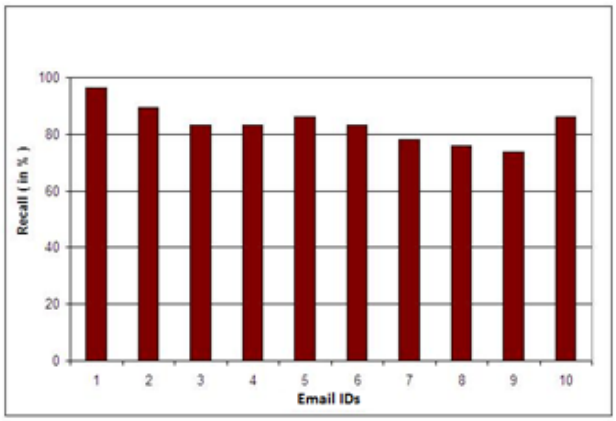

Figure 5: Average Recall of the proposed approach

In Fig. 5, we observe that the tendency of average Recall for the identified spam Emails are high compared to other system. So this shows that our proposed system is achieving high accuracy than any other method.

5. Conclusion and Feature Scope

Proposed system successfully fetches the unread Email subject, body and attachment from the Gmail host. After this process system properly performs pre-processing on the data to get rid of the over burden while detecting the spam Emails. Then Top words are been identified which actually plays a vital role in identification of the important keywords for the spam Emails. Then by using these important words a signature is created based on the previous history and the presence important word factor to identify spam Emails.

This framework can be enhance to detect the Spam Emails from many other host Email services like yahoo, Hotmail and many more. System can also be enhancing as a readymade plug-in for multi host Email service provider with minimal settings.

References

[1] Michal Konkol and Miloslav Konopík , "Named Entity Recognition for Highly Inflectional Languages: Effects of Various Lemmatization and Stemming Approaches" Named Entity Recognition for Highly Inflectional Languages
[2] K.K. Agbele, A.O. Adesina, N.A. Azeez , A.P. Abidoye "Context-Aware Stemming Algorithm for Semantically Related Root Words" © 2012 Afr J Comp & ICT.
[3] J. B. Lovins. (1968). Development of a Stemming Algorithm, Mechanical Translation and Computational Linguistics, vol.11, no. 12, pp: 22-31.
[4] J. Dawson. (1974). Suffix removal and word conflation. ALLC Bulletin, vol. 2, no. 3, pp: 33-46.
[5] M. Porter (1980). An Algorithm for Suffix Stripping. Program, vol. 14, no. 3, pp: 130 – 137.
[6] D. Paice Chris. (1990). Another Stemmer. ACM SIGIR Forum, Volume 24, No. 3, pp: 56-61.
[7] R. Krovetz. (1993). Viewing morphology as an inference process. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA, USA – June 27 th –July 01, 1993, pp: 191-202.
[8] G.E. Freund and P. Willet. (1982), 'Online identification of word variants and arbitrary truncation searching using a string similarity measure'. Information Technology: Research and Development, vol. 1, pp: 177-187.
[9] M. Melucci and N. Orio. (2003), A novel method for stemmer generation based on hidden Markov models. Proceedings of the 12 th international conference on Information and knowledge management, New Orleans, LA, USA, Nov 03 – 08, pp:131-138.
[10] M. Prasenjit, M. Mandar, K. Swapan K. Parui, K. Gobinda, M. Pabitra and D. Kalyankumar. (2007). YASS: Yet another suffix stripper. ACM Transactions on Information Systems. vol. 25, no. 4, article 18.
[11] J. Xu, W.B. Croft, (1998). Corpus-based stemming using co- occurrence of word variants, ACM Transactions on Information Systems, vol. 16, no. 1, pp: 61-81.
[12] P. Funchun, A. Nawaaz, L. Xin and L. Yumao (2007), Context sensitive stemming for web search. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval Amsterdam, July 23 – 27, pp: 639-646.
[13] M. Basavaraju, "A Novel Method of Spam Mail Detection using Text Based Clustering Approach", International Journal of Computer Applications (0975 – 8887) Volume 5– No.4, August 2010.
[14] Ann Nosseir , Khaled Nagati and Islam Taj-Eddin, "Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784.
[15] N. Cristianini, B. Schoelkopf, "Support vector machines and kernel methods, the new generation of learning machines". Artificial Intelligence Magazine, 23(3):31–41, 2002
[16] Wang, Qiang, Yi Guan, and Xiaolong Wang. "SVM-Based Spam Filter with Active and Online Learning." *TREC*. 2006.

Licensed Under Creative Commons Attribution CC BY