Journal of Engineering and Applied Sciences 13 (13): 4875-4879, 2018
ISSN: 1816-949X

Generic Approach for Classifying Spam Mails by Machine Learning Techniques

[1]Banumathy Rajesh and [2]Shanmugasundaram Hariharan
[1]Research and Development Centre, Bharathiar University, 641046 Coimbatore, India
[2]Department of Computer Science and Engineering, Saveetha Engineering College, India

Abstract: Email communication is one of the fastest means of information sharing and has become successful among online users. This predominant success has made web users to generate anonymous contents which are called as spam. Research on identifying these fraudulent information has been a major research issue in recent years and continues to be a major threat. Spam occurs in the information in textual form, short messages and images. Variety of methods exists to ensure security like Naive Bayes, machine learning, Genetic algorithm and others. Machine learning techniques now days used to automatically filter the spam e-Mail in a very successful rate. Classifying the emails as genuine or vice versa is a major research concern. This study attempts to provide a study on this context and there by provide a framework for improving the security. Descriptions of the algorithms are presented and the comparison of their performance on the SpamAssassin spam corpus is presented.

Key words: Email, spam, machine learning, blacklist, Naive Bayes, classification

INTRODUCTION

Information users depend heavily on email's system as one of the major sources of communication. Due to tremendous growth in e-Content and communication there has been a great deal of interest of late in the problem of automatically detecting and filtering out unsolicited commercial e-Mail messages which is commonly referred to as spam. The internet has brought about fundamental changes in the way peoples generate and exchange media information. A variety of approaches have been discussed in this context identify fraudulent information (Lai, 2007). To reduce the spamicity, several algorithms are in existence. Machine learning approach is more efficient than knowledge engineering approach, it does not require specifying any rules (Guzella and Caminhas, 2009). A set of training samples these samples is a set of pre classified e-Mail messages. These unsolicited bulk electronic mails (spam e-Mail) were expressed in different formats and have become one of the most serious problems in internet era. Internet Service Providers (ISP), business firms and general end users the global rate of spam in email traffic in 2014 was 60% internet security threat (Corporation, 2015). Effective image spam detection is of importance in many different domain applications (Soranamageswari and Meena, 2010). Different kinds of features have been used in existing image spam filters including e-Mail header features, image metadata, text-based features and visual-based features. The success of email spam identity have been possible with multimodal textual spam filtering for mobile devices using dendritic cell algorithm (El-Alfy and AlHasan, 2016), Naive Bayes approach (Almeida *et al*., 2011). Its importance and usage are continuously growing despite the evolution of mobile applications, social networks, etc. A large set of personal emails is used for the purpose of folder and subject classifications. Algorithms were developed to perform clustering and classification for this large text collection. Classification based on NGram is shown to be the best for such large text collection especially as text is bi-language (i.e., with English and Arabic content) (Alsmadi and Alhami, 2015).

Academia and industry have shown their concern to accurately detect and effectively control web spam, resulting in a good number of anti-spam techniques currently available. Though there exists many solutions available for the spam mails, each method has its own significance. Moreover, many spam mail incidents are also observed in today's electronic mail system by the Messaging Anti-Abuse Working Group (MAAWG) found that (IDGCI., 2010).

Emails are used on both the personal and professional levels. They can be considered as official documents in communication among users. Email's data mining and analysis can be conducted for several purposes such as: spam detection and classification, subject classification, etc. Email is a convenient means of communication throughout the entire world today. The increased popularity of email spam in both text and images requires a real-time protection mechanism for the media flow. The effective communication is used by large number of users generating hundred of billion of messages every day. It has several advantages viz., less expensive, high speed, reliable and large data transfer

Corresponding Author: Banumathy Rajesh, Research and Development Centre, Bharathiar University, 641046 Coimbatore, India

when compared with other mode of transfers. Tools exists for filtering spam contents, SpamAssassin (ASF., 2016) mail filtering tool is one such tool. SpamAssassin uses a large manually-generated feature set and a simple perceptron classifier with hand-tuned weights to select ham (non-spam) messages and discard spam. Several set of spam words can be filtered easily which could improve the accuracy of the filtering process. The methods available to control spam include anti-spam laws, email protocol exchanges, challenge based systems and response filters. The email originates from various sources. Figure 1 presents the origin of spam mails.

The spam is mixed information which is unwanted to the user. This could be the data originating from multiple sources, information which is irrelevant to the context, representation of data in unusual formats, etc. Figure 2 shows a sample image with spam content. Spam datasets are shown in Table 1 which falls into two categories as email or text. Openly accessible datasets are available for investigation and we extract data from SpamAssassin for experimental.

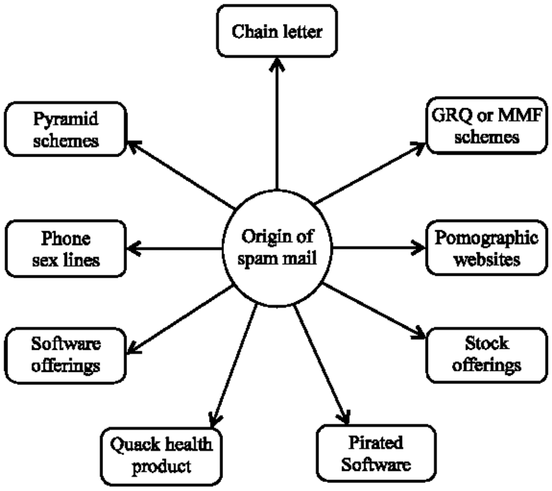

Fig. 1: Origin of spam mails

Literature review: Researches on detecting unsolicited message images (image spams) have become one of the most serious problems for Internet Service Providers (ISPs), business firms and general end users. The researchers have presented a novel system called RoBoTs (Robust BoosTrap based spam detector) to achieve a robust image spam filter. The system is developed based on multiple visual properties extracted from different levels of granularity, capturing discriminative contents to achieve effectiveness (Shen *et al*., 2015). Also the research focuses on to develop a learning frame research to effectively integrate random forest and Linear Discriminative Analysis (LDA). Supervised machine learning algorithms is adopted for extracting subjective information from text documents online. This has facilitated in accurate spam classification process using public datasets used for training and test.

Over the last years, internet spam content has spread enormously inside web sites mainly due to the

Table 1: Spam datasets and its characteristics

Dataset	Focus	Nature	Spam	Legitimate
SpamAssassin	Email	Preprocessed	1813	2788
TREC 2005	Email	Raw	1323	1390
SpamBase	Email	Raw	1323	276
Spam collection	SMS	Raw	747	4827

Are you having a hard time paying your mortgage? Do you need help restructuring the mortgage to get a lower payment? If so, reply HELP. To cancel, reply STOP.

STOP. I was in the middle of a wonderful masturbation session and now it's ruined all thanks to you assholes. Next time I'll reach through the phone and rip your dick off.

Fig. 2: Sample content with spam

emergence of new web technologies, online sharing of resources and information. A vast approach uses ensemble learning paradigms for sentiment analysis to enhance the predictive performance (Onan *et al*., 2016). A paradigm integrating multi objective, optimization based weighted voting scheme is used. An appropriate weight value is assigned to existing classifiers to each class based on the predictive performance of classification algorithms. The ensemble method based on static classifier selection involves majority voting error and forward search as well as a multi objective differential evolution algorithm.

Some illustrative schemes incorporates Bayesian logistic regression, Naive Bayes, linear discriminant analysis, logistic regression and support vector machines as base learners whose performance in terms of precision and recall values determines weight adjustment. Our experimental analysis of classification tasks including sentiment analysis, software defect prediction, credit risk modeling, spam filtering, and semantic mapping, suggests that the proposed classification scheme can predict better than conventional ensemble learning methods such as AdaBoost, bagging, random subspace and majority voting.

The successful integration of different algorithms for web spam classification is still a challenge. In this context, a present study introduces WSF2, a novel web spam filtering framework specifically designed to take advantage of multiple classification schemes and algorithms. The approach encodes the life cycle of a case-based reasoning system with appropriate knowledge and different parameters to ensure continuous improvement in filtering with passage of time leading to good precision. The evaluation of the dynamic model leads to improved efficiency with set of experiments involving a publicly available corpus. The approach using well known classifiers and ensemble approaches were used which revealed that WSF2 performed well, being able to take advantage of each classifier and to achieve a better performance when compared to other alternatives (Heydari *et al*., 2016).

Earlier approaches were limited by the adaptive nature of unsolicited email spam. To alleviate this research, email detection system was introduced by some researchers. Particle Swarm Optimization (PSO) was implemented to improve the random detector generation in the Negative Selection Algorithm (NSA). The algorithm generates detectors in the random detector generation phase of the negative selection algorithm. The combined NSA-PSO uses a Local Outlier Factor (LOF) as the fitness function for the detector generation. The detector generation process is terminated when the expected spam coverage is reached. A distance measure and a threshold value are employed to enhance the distinctiveness between the non-spam and spam detectors after the detector generation. The implementation and evaluation of the models are analyzed. The results show that the accuracy of the proposed NSA-PSO Model is better than the accuracy of the standard NSA Model. The proposed model with the best accuracy is further used to differentiate between spam and non-spam in a network that is developed based on a client-server network for spam detection (Idris *et al*., 2015).

Existing reports clearly indicate that volume of spam over instant messaging and SMS is dramatically increasing year by year. It represents a challenging problem for traditional filtering methods nowadays, since, such messages are usually fairly short and normally rife with slangs, idioms, symbols and acronyms that make even tokenization a difficult task. In this scenario, the approach proposes a method to normalize and expand original short and messy text messages in order to acquire better attributes and enhance the classification performance. Preprocessing is done based on lexicographic and semantic dictionaries along with state-of-the-art techniques for semantic analysis and context detection. This technique is used to normalize terms and create new attributes in order to change and expand original text samples aiming to alleviate factors that can degrade the algorithms performance, such as redundancies and inconsistencies (Almeida *et al*., 2016).

MATERIALS AND METHODS

Framework for email filtering systems: The framework for email filtering system is presented in this study. Figure 3 presents the framework of our email intrusion detection system. The sequence of steps involved in the proposed system is shown in Fig. 4. The proposed e-Mail Personalization system will be able to prioritize the e-Mail messages into different categories. The performance is evaluated using precision and recall using Eq. 1 and 2. Table 2 presents the spam datasets used for experiments and the characteristics exhibited by the spam datasets.

Table 2: Spam datasets and its characteristics

Message collection	Training set	Testing set
Ham message	842	432
Spam message	736	521

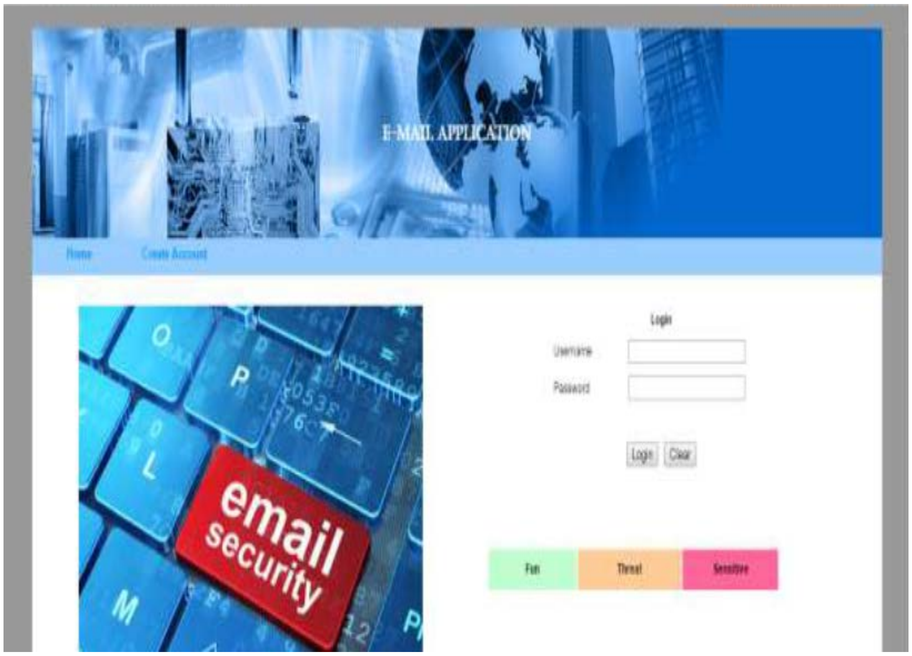

Fig. 3: Proposed framework for spam identification

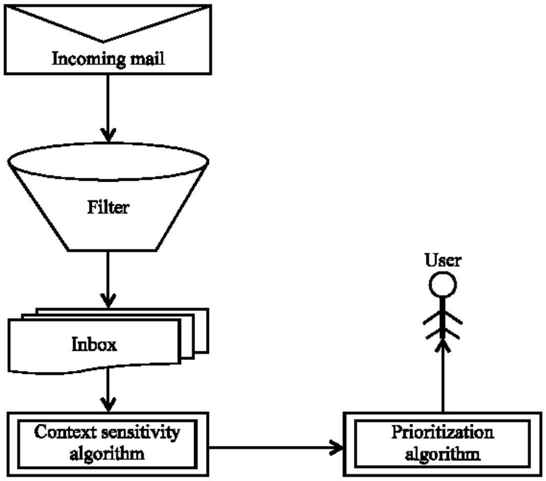

Fig. 4: Architecture for spam detection

RESULT AND DISCUSSION

The results from Table 3 shows that, the SVM approach is dominant and yields better results in the context spam identification. The result shows significant improvements in terms of precision and recall measures:

$$\text{Recall} = \frac{\text{No. of mails correctly classified}}{\text{Total No. of messages}} \quad (1)$$

$$\text{Accuracy} = \frac{\text{No. of mails correctly categorized}}{\text{Total No. of emails}} \quad (2)$$

Table 3: Performance using different classifiers

Methods	Recall	Precision	Accuracy
SVM	0.98	0.99	99.46
KNN	0.95	0.93	96.90

CONCLUSION

The study has presented a study on email filtering using SVM approach. The research shows some improvements as compared to KNN approach improvements in the existing schemes. Experimental illustration have significance using SpamAssassin and TREC 2005 datasets for emails which are raw contents. The proposed approach investigates on the significance of the preprocessing role to reduce the spam mails. Currently research is in progress on to identify image spam.

ACKNOWLEDGEMENTS

The researcher would like to express our sincere thanks to anonymous reviewers for their insightful suggestions. The researcher indeed would extend their acknowledgement for the support by all well wishers, faculty members and others who helped out in achieving this research task to get completed.

REFERENCES

ASF., 2016. Apache spam assassin. Apache Software Foundation, Forest Hill, Maryland, USA. http://www.spamassassin.org.

Almeida, T.A., J. Almeida and A. Yamakami, 2011. Spam filtering: How the dimensionality reduction affects the accuracy of Naive Bayes classifiers. J. Internet Serv. Appl., 1: 183-200.

Almeida, T.A., T.P. Silva, I. Santos and J.M.G. Hidalgo, 2016. Text normalization and semantic indexing to enhance instant messaging and SMS spam filtering. Knowledge Based Syst., 108: 25-32.

Alsmadi, I. and I. Alhami, 2015. Clustering and classification of email contents. J. King Saud Univ. Comput. Inf. Sci., 27: 46-57.

Corporation, S., 2015. Internet security threat report. Symantec, Mountain View, California, USA. https://www4.symantec.com/mktginfo/whitepaper/ISTR/21347932_GAinternet-security-threat-report-volume-20-2015-social_v2.pdf

El-Alfy, E.S.M. and A.A. AlHasan, 2016. Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm. Future Generation Comput. Syst., 64: 98-107.

Guzella, T.S. and W.M. Caminhas, 2009. A review of machine learning approaches to spam filtering. Expert Syst. Appl., 36: 10206-10222.

Heydari, A., M. Tavakoli and N. Salim, 2016. Detection of fake opinions using time series. Expert Syst. Appl., 58: 83-92.

IDGCI., 2010. Sign up for network world newsletters. IDG Communications Inc. Framingham, Massachusetts. http://www.networkworld.com/newsletters/

Idris, I., A. Selamat, N.T. Nguyen, S. Omatu and O. Krejcar *et al*., 2015. A combined negative selection algorithm-particle swarm optimization for an email spam detection system. Eng. Appl. Artif. Intell., 39: 33-44.

Lai, C.C., 2007. An empirical study of three machine learning methods for spam filtering. Knowledge-Based Syst., 20: 249-254.

Onan, A., S. Korukoglu and H. Bulut, 2016. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Syst. Appl., 62: 1-16.

Shen, J., R.H. Deng, Z. Cheng, L. Nie and S. Yan, 2015. On robust image spam filtering via comprehensive visual modeling. Pattern Recognit., 48: 3227-3238.

Soranamageswari, M. and C. Meena, 2010. Statistical feature extraction for classification of image spam using artificial neural networks. Proceedings of the 2010 2nd International Conference on Machine Learning and Computing (ICMLC), February 9-11, 2010, IEEE, Bangalore, India, ISBN:978-1-4244-6007-6, pp: 101-105.