ARTICLE IN PRESS

Information Sciences xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Study on the effectiveness of anomaly detection for spam filtering

Carlos Laorden *, Xabier Ugarte-Pedrero, Igor Santos, Borja Sanz, Javier Nieves, Pablo G. Bringas

Laboratory for Smartness, Semantics and Security (S³Lab), University of Deusto, Avenida de las Universidades 24, 48007 Bilbao, Spain

ARTICLE INFO

Article history:
Received 5 November 2011
Received in revised form 24 May 2013
Accepted 15 February 2014
Available online xxxx

Keywords:
Spam filtering
Anomaly detection
Secure e-commerce
Computer security

ABSTRACT

Spam has become an important problem for computer security because it is a channel for spreading threats, including computer viruses, worms and phishing. Currently, more than 85% of received emails are spam. Historical approaches to combating these messages, including simple techniques such as sender blacklisting or using email signatures, are no longer completely reliable on their own. Many solutions utilise machine-learning approaches trained with statistical representations of the terms that usually appear in the emails. Nevertheless, these methods require a time-consuming training step with labelled data. Dealing with the limited availability of labelled training instances slows down the progress of filtering systems and offers advantages to spammers. In this paper, we present a study of the effectiveness of anomaly detection applied to spam filtering, which reduces the necessity of labelling spam messages and only employs the representation of one class of emails (i.e., legitimate or spam). This study includes a presentation of the first anomaly based spam filtering system, an enhancement of this system that applies a data reduction algorithm to the labelled dataset to reduce processing time while maintaining detection rates and an analysis of the suitability of choosing legitimate emails or spam as a representation of normality.

1. Introduction

Electronic mail is a powerful communication channel. However, as with all useful media, it is prone to misuse. Flooding inboxes with annoying and time-consuming messages, more than 85% of received emails are spam.[1] Bulk email is not only annoying to everyday email users but also constitutes a major computer security problem that costs billions of dollars in productivity losses [6]. It is also commonly used as a medium for *phishing* (i.e., attacks that seek to acquire sensitive information from end-users) [18] and the spread of malicious software (e.g., computer viruses, Trojan horses, spyware and Internet worms) [6].

Different studies have shown that spam has a notorious and prejudicial effect on the worldwide economy. Leung and Liang [25] presented an analysis of the impact of phishing on the market value of global firms, which showed that phishing alerts lead to a significant negative return on stock. In a similar vein, Mostafa Raad et al. [30] offered another study to assess

* Corresponding author. Tel.: +34 944139003; fax: +34 944139166.
E-mail addresses: claorden@deusto.es (C. Laorden), xabier.ugarte@deusto.es (X. Ugarte-Pedrero), isantos@deusto.es (I. Santos), borja.sanz@deusto.es (B. Sanz), jnieves@deusto.es (J. Nieves), pablo.garcia.bringas@deusto.es (P.G. Bringas).

[1] http://www.spam-o-meter.com/ (October 17, 2011).

http://dx.doi.org/10.1016/j.ins.2014.02.114
0020-0255/

the influence and impact of spam in several companies whose email advertisements were considered spam. Both examples clearly illustrate the necessity to detect undesired messages and, perhaps more importantly, the need to restore users confidence in their email filtering systems.

The academic community has proposed several approaches to solve the spam problem [34,9,39,10]. Among them, the *statistical approaches* [42] use machine-learning techniques to classify emails. These approaches have proven their efficiency in detecting spam and are the most utilised techniques to fight it. In particular, Bayes' theorem is widely used by anti-spam filters (e.g., SpamAssassin [27], Bogofilter [33] and Spamprobe [7]).

Statistical approaches are usually supervised, as they require a training set of previously labelled samples. These techniques perform better as more training instances are available, which requires a significant amount of previous labelling work to increase the models' accuracy. This work includes a gathering phase, in which as many emails as possible are collected. Nonetheless, the availability of labelled training instances is limited, which slows the progress of anti-spam systems.

In light of these difficulties, we propose the application of anomaly detection to spam filtering. Our approach can determine whether or not an email is spam by comparing word frequency features with a dataset composed only of what is considered normal (i.e., usually legitimate emails). If the email under inspection presents a considerable deviation from what is considered typical, it is considered an anomaly, or spam. This method does not need updated data about spam messages and thus reduces the efforts of labelling messages, working, for instance, only with a user's valid inbox folder.

By studying our method, we noticed that the number of comparisons needed to analyse each sample was considerably high (i.e., comparison against every legitimate email), resulting in a high processing overhead. We therefore present an enhancement to our approach by applying partitional clustering to reduce the number of vectors in the dataset used as normality. This improvement boosts scalability due to the reduction in processing time.

Finally, because the amount of spam within all email messages greatly exceeds the number of legitimate emails, a question regarding the suitability of choosing legitimate emails, instead of spam, as a representation of normality, may arise. Therefore, we performed a thorough study on the issue, providing comparisons between the two approaches, using both legitimate emails and spam as representations of normality.

In summary, our main findings presented in this paper include the following:

- We present an anomaly-based approach for spam filtering by proposing different deviation measures to determine whether an email is spam.
- We adapt a method for email dataset reduction based on the partitional clustering algorithm Quality Threshold (QT) and generate reduced datasets of different sizes.
- We empirically validate the reduction algorithm by testing its accuracy results and comparing them to the approach using the unreduced datasets.
- We prove that a unique, synthetically generated sample of legitimate emails is representative enough to implement an anomaly detection system without compromising accuracy results.
- We show that labelling efforts can be reduced in the industry, while still maintaining a high rate of accuracy.

The remainder of this paper is organised as follows. Section 2 provides background regarding the representation of emails based on the Vector Space Model (VSM). Section 3 details our anomaly based method. Section 4 describes the experiments and presents the results of the approach without the dataset reduction. Section 5 details the application of the dataset reduction step to our anomaly based method. Section 6 describes the experiments and presents the results of our anomaly based approach enhanced with the dataset reduction, offering a comparison with the approach that does not perform the reducing step. Section 7 compares the use of legitimate emails against spam as representations of normality. Section 9 discusses the implications of the obtained results and shows the limitations of the proposed approach. Finally, Section 10 concludes the paper and outlines avenues for future work.

2. Vector space model for spam filtering

Spam filtering software attempts to accurately classify email messages into 2 main categories: spam or legitimate messages. We thus use information found within the body and subject of an email message and discard every other piece of information (including the sender or time-stamp of the email). To represent messages, we remove the stop-words [40], which are words devoid of content (e.g., 'a', 'the', 'is'). These words do not provide any semantic information and add noise to the model [36].

We then represent the emails using an Information Retrieval (IR) model. Formally, let an IR model be defined as a 4-tuple $[\mathcal{E}, \mathcal{Q}, \mathcal{F}, R, (q_i, e_j)]$ [3] where.

- $\mathcal{E}$ is a set of representations of email.
- $\mathcal{Q}$ is a set of representations of user queries.
- $\mathcal{F}$ is a framework for modelling emails, queries and their relationships.
- $R(q_i, e_j)$ is a ranking function that associates a real number with a query $q_i,\ (q_i \in \mathcal{Q})$ and an email representation $e_j,\ (e_j \in \mathcal{E})$. This function is also called a similarity function.

Please cite this article in press as: C. Laorden et al., Study on the effectiveness of anomaly detection for spam filtering, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.02.114