Expert Systems with Applications 36 (2009) 4321–4330

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks

Chih-Hung Wu *

Department of Electrical Engineering, National University of Kaohsiung, Kaohsiung, Taiwan

ARTICLE INFO

Keywords:
Spam filtering
Classification
Back-propagation neural networks
Rule-based reasoning
Machine-learning

ABSTRACT

Earlier methods on spam filtering usually compare the contents of emails against specific keywords, which are not robust as the spammers frequently change the terms used in emails. This paper presents a hybrid method of rule-based processing and back-propagation neural networks for spam filtering. Instead of using keywords, this study utilize the spamming behaviors as features for describing emails. A rule-based process is first employed to identify and digitize the spamming behaviors observed from the headers and syslogs of emails. An enhanced BPNN with a weighted learning strategy is designed as the classification mechanism. Since spamming behaviors are infrequently changed, compared with that of keywords used in spams, the proposed method is more robust with respect to the change of time. The experimental results show that the proposed method is useful in identifying spam emails.

1. Introduction

With the popularity of the Internet, the inundation of unsolicited commercial emails, or spam, is an emerging problem. Anti-spam by determining whether or not an incoming email is spam has become an important problem. Various techniques for automatically detecting or filtering spam emails have been proposed. Many practical applications rely on building comprehensive databases for blocking emails whose addresses have been reported as black-lists or whose message bodies contain specific words or phrases defined as threatening terms. Among the others, machine-learning based techniques for context investigation receive a lot of attentions. Such machine-learning based methods construct rules or models with weighted scoring about the positions, frequencies and context associations of terms or phrases used in spam and estimate the likelihood that an incoming email is spam or ham accordingly. Methods based on such context filtering, or keyword-based filtering, are effective, if keywords are explicitly given. However, spammers usually attempt to make their messages as indistinguishable from legitimate email as possible and change the patterns of spam to foil the filters. Some spams are tailored by sophisticated programs to make them like normal messages which may not contain any specific keywords. From the point of view of machine-learning, the key to success of applying machine-learning based methods is the correctness of features which can precisely describe the training samples. However, with the limited spam corpus and the changes of email keywords, conventional approaches may not be able to precisely capture the characteristics of spam. Clearly, keyword-based filtering is a workable but limited approach for detecting spam. By observing the behaviors of spammers, we found that spam emails are generated and delivered according to some specific "*spamming behaviors*". For example, spam emails are sent with anonymous or forged user names, forwarded by illegal permission or accounts, delivered with a bunch of the same message repeatedly and unauthorizedly to many different recipients, and so on. Using specific keywords is only a class of these behaviors. Although, spam emails are changing their forms (Hall, 2000), human beings can easily recognize them no matter how they are generated (for example, image spam) and distributed. Spam filtering according to the concept of spamming behaviors is first presented in Tseng and Wu (2003) which claims that such behaviors can be used for identifying spams since they have better resistance with respect to the change of time. In this paper, a back-propagation neural network is designed and implemented for spam classification. Emails to be investigated by the neural network are described in terms of their spamming behaviors, not keywords them contains. The spamming behaviors of emails are first identified by a rule-based pre-processor. Next, the identified features are encoded as three-valued vectors and processed by the proposed neural network. Since spamming behaviors change inactively, in comparison with the changing frequency of keywords, so that classification of spams using behavior-based features may be more robust than keyword-based methods. Experimental results show that spam classification using behavior-based features is more robust.

* Tel.: +886 7 5919446; fax: +886 7 5919374.
E-mail address: johnw@nuk.edu.tw
URL: http://www.johnw.idv.tw.

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.03.002

The rest of the paper is organized as follows. Section 2 presents several typical methods of anti-spamming. Section 3 gives a brief description on the features for spam classification and the concept of spamming behaviors of emails. Section 4 presents a rule-based method for instantiating behavior-based features into discrete values. In Section 5 we present the design and implementation of back-propagation neural networks for spam classification using behavior-based features. The experimental results are presented in Section 6. Finally, we conclude this study in Section 7.

2. Related work

In most machine-learning based methods, filtering models using keyword-base features is intuitive since they are easier to implement. Commonly used learning techniques include decision trees (Crawford, Kay, & McCreath, 2001), case-based reasoning (Delany, Cunningham, Doyle, & Zamolotskikh, 2005; Mendez, Fdez-Riverola, Iglesias, Diaz, & Corchado, 2006), support vector machine (SVM) (Camastra, 2005; Wang, Yu, & Liu, 2005), artificial immune systems (Bezerra1et al., Wang, You, & Man, 2006; Wang et al., 2006), and boosting method Schapire and Singer (2000). Here are some examples. Bayesian-based spam filtering computes the probability that an email is spam based upon previous feature frequencies in spam and ham (Androutsopoulos, Koutsias, Chandrinos, & Spyropoulos, 2000; Graham, 2003; Wang, Hori, & Sakurai, 2006). Katirai (1999) employs genetic programming (GP) and Naïve Bayesian (NB) to filter out spam e-mails. The results demonstrate that NB slightly outperforms GP. Sakkis, Androutsopoulos, Paliouras, and Stamatopoulos (2003) develop a memory-based approach and claim that it outperforms NB methods. Brutlag and Meek (2000) compared SVM, TF-IDF, and unigram model in general e-mail classification. Diao, Lu, and Wu (2000) compare NB with decision trees in classification-based personal e-mail filtering. Applying neural networks for filtering spam e-mails is also studied in Clark, Koprinska, and Poon (2003), which produces a very good training accuracy. In Wang, Jones, and Pan (2006), two linear classifiers, Perceptron and Winnow, are integrated for spam filtering. A Hybrid method which combines neural networks and genetic algorithms for feature selection is presented in Gavrilis, Tsoulos, and Dermatas (2006) for robust detection of spam. Unfortunately, they do not give impressive results on practical tests. Also using keyword-based exploration, the concept drill problem is studied for filtering spam emails in Fdez-Riverola, Iglesias, Díaz, Méndez, and Corchado (2007), wherein lazy learning algorithms are applied to select relevant terms and the representative emails associated with the terms. The filtering method presented in Jiang (2006) applies the techniques of Latent semantic indexing and singular value decomposition to transform emails into statistically semantic vectors and classify them accordingly. A content-based mining method for author identification of e-mails is proposed in de Vel, Anderson, Corney, and Mohay (2001), where structural characteristics and linguistic patterns are analyzed. Yue, Abraham, Chi, Hao, and Mo (2007) employ "behavior-based" features in an artificial immune system for spam filtering. The features they used focus on the sender's IP addresses recorded in the MX record. Zhang, Liu, Zhang, and Wang (2006) present in a method for recognizing spam behaviors using decision trees learned from data maintained during transfer sessions. However, the so-called behaviors they defined are not the same as what we defined in this paper. More comparative studies on spam filtering techniques are available at Blanzieri and Bryl (2006), Hoanca (2006), Lai and Tsai (2004), Webb et al. (2005).

3. Behavior-based features

3.1. Spamming behaviors

A spam is generally recognized as an electronic message in which the recipient's personal identity and context are irrelevant because the message is equally applicable to many other potential recipients; and the recipient has not verifiably granted deliberate, explicit, and still-revocable permission for it to be sent; and the transmission and reception of the message appears to the recipient to give a disproportionate benefit to the sender (MAPS, 2003). The objective of sending spams is to sell products or services to the customers available on the Internet. For this purpose, spams are massively and repeatedly dispatched in order to broadly contact potential customers. However, in order not to be detected, spams are elaborately pretended as hams. The so-called *spamming behavior* is a pretending trick that spammers use for composing or delivering a spam for specific purposes. For example, in order to bypass spam filters, spambots fill the subject line with randomly generated characters and make the subject irrelevant to the message body. To hide the origin of spams, routing addresses or returning address are forged or invalid. Spams are massively sent to designated addresses which are collected manually by the spammer or automatically by spambots. Moreover, spams are usually delivered at non-office hours, such as 02:00 AM–06:00 AM, because of larger bandwidth available and lower risk to be coped. Unlike finding keyword-based features that can be extracted directly from the contents of emails, formatting behavior-based features for emails has more difficulty in direct extraction. Fortunately, they can be partially obtained from the transmission information associated with emails. Let us recall the process of delivering emails. Internet users usually compose or read emails using mail user agents (MUAs) which connect to mail servers. A mail transfer agent (MTA) is installed on the mail server for delivering emails and communicating with MUAs. Emails received by MTA are retained in a temporary file called mailbox until being downloaded by the recipients. Regarding to the structure of an email, the *header* is a piece of structured messages stating the organization and destination of the mail. Some of the information in the header can be given by the sender manually; some by MUA automatically. MTA delivers an email according to its header information. When an email is delivered by MTA, a record describing this delivery is added to an auditing file, referred to as *syslog*. Unlike the header parts, syslogs are generated by MTA and are not modifiable by users or MTA. Each record in an email's header or syslog consists of several "fields" each of which describes a piece of information related to the email. Emails transmitted with different communication protocols may have different formats in constructing headers and syslogs. Throughout this paper, we explore emails delivered by Simple Mail Transfer Protocol (SMTP) and sendmail in BSD 4.4 format. The readers may refer to Costales and Allman (2002) for more details. Fig. 1a and b present sample headers and syslogs of emails, where "`From:`", "`Reply-To:`", etc., are fields in the header and "`daemon`", "`nrcpts`", etc., are fields in the syslog. Spamming behaviors like forging messages, late or massively delivering, and the ones that cause inconsistency in the email messages are observable from headers and syslogs. The concept of spamming behaviors is first presented in Tseng and Wu (2003) which claims that such behaviors can be used for identifying spams since they have better resistance with respect to the change of time. In most cases, normal emails (or hams) are composed and delivered with valid and real transmission information. Conversely, the information associated with spams may contain inconsistent or abnormal information from which may reveal the existence of spamming behaviors.