Expert Systems with Applications 39 (2012) 12487–12500

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

SDAI: An integral evaluation methodology for content-based spam filtering models

Noemí Pérez-Díaz, David Ruano-Ordás, Florentino Fdez-Riverola, José R. Méndez *

Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, University of Vigo, 32004 Ourense, Spain

ARTICLE INFO

Keywords:
Spam classification
Content-based techniques
Model comparison
Evaluation methodology

ABSTRACT

Tragedy of Commons Theory introduced by Hardin (1968) revealed how shared and limited resources get completely depleted as effect of human behaviour. By analogy, common spamming activities can be properly modelled by this solid theory and, consequently, a young Internet Security Industry has recently emerged to fight against spam. However, the massive intensification of spam deliveries during last years has led to the need of achieving a significant improvement in filter accuracy. In this context, current research efforts are mainly focussed on providing a wide variety of content-based techniques able to overcome common spam filtering inconveniencies. Although theoretical filtering evaluation is generally taken into consideration in scientific works, most of the evaluation protocols are not appropriate to correctly assess the performance of models during filter operation in real environments. In order to cover the gap between basic research and applied deployment of well-known spam filtering techniques, this work proposes a novel straightforward evaluation methodology able to rank available models using four different but complementary perspectives: *static*, *dynamic*, *adaptive* and *internationalisation*. In the present study, we applied our SDAI methodology to compare eight different well-known content-based spam filtering techniques using several established accuracy measures. Results showed the effect of the knowledge grain-size and evidenced several unexpected situations related with the behaviour of analysed models.

1. Introduction

The introduction of Internet mail service has provided a cost-effective way to communicate people worldwide. This particular way of contact is commonly used by billons of users generating hundreds of billions of messages every day (The Radicati Group, 2010). Unfortunately, low cost associated with Internet mail has boosted its utilisation as an advertisement platform. Nowadays, most messages delivered through Internet are sent to promote illegal and/or useless products bothering final users. These e-mails constitute one of the most important form of spamming that also includes (i) adds made by comments on blogs, (ii) advertisement bots on instant messaging communities, (iii) publicity included in wikis, (iv) promotion through newsgroups and (v) unwanted SMS communications. In this context, latest spam reports accounts that more than 70% of delivered messages are spam (Symantec Corp., 2011a).

Previous to the occurrence and proliferation of the spam phenomenon, Hardin (1968) introduced the *Tragedy of Commons* as one of the hard to overcome difficulties to sustainable development. This dilemma stands for the undesired destruction of common shared limited resources as effect of the human psychology. Since that, many economists have studied the impact of this event to explain different situations including the collapse of Soviet Union, pollution control and global warming or the regulation and use of natural resources. In the particular case of spam and other Internet missuses (e.g.: worms and viruses), a profitable business model based on providing security services to final users has evolved in line, guaranteeing Internet usage to everybody and allowing the exploitation of Internet for business. As an example of this situation, Symantec Corporation presented a GAAP revenue for a total amount of 6.19 billions of dollars in 2011 (Symantec Corp., 2011b).

During the last years, applied research carried out by the spam filtering industry has led to the commercialisation of relatively successful software products and services to fight spam. However, the accurate identification of illegitimate messages is a difficult task partly because (i) its subjective nature, (ii) the adverse effects of concept drift, (iii) different language issues, (iv) several delays introduced in the delivery process and (v) the asymmetric cost of filtering errors. By way of example of previously mentioned difficulties, a given message containing the word *Rolex* would be complicate to handle when the final user is a watchmaker. In addition, concept drift prevents the clear identification of new obfuscating trends and/or illegal advertisements (Tsymbal, 2004). As an example, the automatic detection of spam messages offering potassium

* Corresponding author. Tel.: +34 988 387015; fax: +34 988 387001.
E-mail addresses: npdiaz@uvigo.es (N. Pérez-Díaz), drordas@uvigo.es (D. Ruano-Ordás), riverola@uvigo.es (F. Fdez-Riverola), moncho.mendez@uvigo.es (J.R. Méndez).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.04.064

iodide pills after Fukushima nuclear disaster was almost impossible. Additionally, filters have to analyse e-mails written in a wide variety of languages which lead to complex models demanding high computational requirements. Last, but not least, are the different weights assigned by final users to the occurrence of false positive (legitimate e-mails classified as spam) and false negative (spam messages identified as legitimate) errors.

With the goal of both, assessing the performance of a given filter and compare its behaviour with existing models, previous research works have introduced several standardised methods and measures. In this context, Kohavi (1995) proved the importance of using cross-validation schemes to guarantee the quality of measurements carried out during empirical evaluations. Moreover, some preceding works introduced several measures specifically intended for correctly assessing filter accuracy such as (i) batting average (Graham-Cumming, 2004), (ii) recall, precision and *f*-score (Rijsbergen, 1979), (iii) balanced *f*-score (Shaw, Burgin, & Howell, 1997) and (iv) total cost ratio (Androutsopoulos, Koustias, Chandrinos, Paliouras, & Spyropoulos, 2000). Additionally, in order to consider evaluation from a more theoretical point of view, ROC (*Receiver Operating Characteristic*) curves have also been successfully imported from the biomedical domain (Egan, 1975).

Although classical spam filter evaluation seems quite consolidated from an academic perspective, current methods and measures do not correctly simulate the real environment in which a filter actually runs. In fact, common experimentation protocols do not reproduce scenarios characterised by the subjective nature of spam, concept drift effects, the presence of multiple languages in different e-mails or the delay added to the message delivery due to filter execution. This situation contributes to the generation of overoptimistic results for the spam filtering industry and makes actual filter deployment unsecure due to the existence of relevant untested conditions. In this context, the aim of the present work is twofold: to define a straightforward evaluation protocol taking into consideration ignored but fundamental issues concerning the real scenario in which spam filters actually work, and to evaluate a wide range of well-known content-based anti-spam techniques from different but complementary perspectives.

The rest of the paper is structured as follows: Section 2 presents a survey of previously successful works extracted from the spam filtering literature. Section 3 identifies main drawbacks found in current evaluation protocols and introduces our SDAI methodology together with a survey of available corpuses supporting experimentation. Section 4 details the experimental setup and discusses in detail the results obtained from the application of the proposed methodology. Finally, Section 5 summarises the main conclusions and outlines future research work.

2. Related work on spam filtering

Nowadays, professional spam filtering services are implemented by using an adequate combination of different classifiers and techniques supported by advanced filter platforms such as SpamAssassin (The Apache SpamAssassin Project, 2011) or Wirebrush4SPAM (Wirebrush4SPAM, 2011). These frameworks provide a clear separation between custom filters (usually modelled through a set of rules) and filtering techniques (implemented by the own framework). A common filter usually combines three different kinds of techniques: (i) domain authentication schemes, (ii) collaborative approaches and (iii) content-based classifiers.

The first group comprises a set of techniques implementing delivery authorisation and authentication procedures. Following this approach, SPF (*Sender Policy Framework*) (Wong & Schlitt, 2006) uses the TXT entries of a DNS (*Domain Name System*) zone to share lists of e-mail servers that are allowed to send e-mails from it. Another well-known alternative included in this group is DKIM (*Domain Keys Identified Mail*) (Allman et al., 2007), which implements a successful domain authentication method built on the top of PKI (*Public Key Infrastructure*) technology (Adams & Lloyd, 2002).

From another perspective, collaborative approaches are based on exploiting network-sharing capabilities to distribute information concerning spam messages and servers. Razor and Pyzor are Internet communities sharing nilsimsa sums of spam messages (Nilsimsa, 2011). In the same group, DNSBL (*DNS-based Black Lists*) and DNSWL (*DNS-based White Lists*) are widely used to share information about servers delivering spam and legitimate messages through DNS service[1] (Levine, 2010). Finally, URIBL (*Uniform Resource Identifier Black Lists*) also uses DNSBL/DNSWL technology to distribute information about Uniform Resource Identifiers found in spam messages.

In the third group, content-based techniques are intended to classify e-mails by carrying out a thoughtful analysis of their contents (i.e.: subject, body and attachments). Although there are available some alternatives able to provide useful information for classifying messages by processing attached images (Biggio, Fumera, Pillai, & Roli, 2011), the present work is mainly focussed on analysing the performance of classical text-based anti-spam classifiers. For this task, ML (*Machine Learning*) algorithms such as support vector machines (SVM), tree boosting schemes, random forests (RF), case-based reasoning (CBR) systems, rough sets (RS), neural networks (ANN), artificial immune systems (AIS) and different variants of the Naïve Bayes (NB) algorithm have been successfully applied to the spam filtering domain.

In the particular case of ML algorithms, perhaps the most common technique used for solving binary classification problems is SVM (Vapnik, 1999). Available SVM implementations are able to find an adequate transformation of an input *n*-dimensional space defined by *n* input features (those considered relevant to the classification problem) into a new *m*-dimensional space where positive and negative training examples can be separated linearly. Despite SVM is not commonly used in popular commercial spam filters, previous works have demonstrated its suitability for this domain (Druker, Wu, & Vapnik, 1999).

From another perspective, boosting techniques also emerged as meta-classifiers able to successfully take advantage of weak learning schemes (e.g.: decision stumps, C4.5, etc.) in order to improve classification results in spam filtering (Carreras & Márquez, 2001; Nicholas, 2003). In particular, there are available different well-known boosting alternatives like AdaBoost (Schapire & Singer, 1999), LogitBoost (Friedman, Hastie, & Tibshirani, 2000) and Boost by Majority (Freund, 1995). Related with boosting techniques, Random Forest (Breiman, 2001) also represents a successful ensemble learning scheme that has been previously applied to filter spam e-mails (DeBarr & Wechsler, 2009).

In addition, case-based reasoning systems are commonly used to solve problems by reusing knowledge from past situations (Aamodt & Plaza, 1994). This problem solving strategy has been widely used for years as a methodology to guide the construction of different successful spam filtering techniques (Cunningham, Nowlan, Delany, & Haahr, 2003; Delany & Cunningham, 2004; Delany, Cunningham, & Coyle, 2004; Delany, Cunningham, Tsymbal, & Coyle, 2005; Fdez-Riverola, Iglesias, Díaz, Méndez, & Corchado, 2007a).

Also, rough set theory (Pawlak, 1982, 1991, 1993) is achieving great popularity in the spam filtering community. RS are commonly used to generate a set of rules from previous collected data that can be further used to classify new incoming e-mails (Glymin & Ziarko,

[1] SpamHaus (http://www.spamhaus.org), SORBS (http://www.sorbs.net) and DNSWL (http://www.dnswl.org) currently provide this service.