Electronic Commerce Research and Applications 11 (2012) 290–298

Contents lists available at SciVerse ScienceDirect

Electronic Commerce Research and Applications

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

Word sense disambiguation for spam filtering

Carlos Laorden [a,*], Igor Santos [a,1], Borja Sanz [a,1], Gonzalo Alvarez [b,2], Pablo G. Bringas [a,1]

[a] Laboratory for Smartness, Semantics and Security (S³Lab), University of Deusto, Avenida de las Universidades 24, 48007 Bilbao, Spain
[b] Instituto de Física Aplicada, Consejo Superior de Investigaciones Científicas (CSIC), C/Serrano 144, 28006 Madrid, Spain

ARTICLE INFO

Article history:
Received 19 May 2011
Received in revised form 27 September 2011
Accepted 22 November 2011
Available online 28 December 2011

Keywords:
Spam filtering
Word sense disambiguation
Secure e-commerce
Computer security

ABSTRACT

Spam has become a major issue in computer security because it is a channel for threats such as computer viruses, worms, and phishing. More than 86% of received e-mails are spam. Historical approaches to combating these messages, including simple techniques such as sender blacklisting or the use of e-mail signatures, are no longer completely reliable. Many current solutions feature machine-learning algorithms trained using statistical representations of the terms that most commonly appear in such e-mails. However, these methods are merely syntactic and are unable to account for the underlying semantics of terms within messages. In this paper, we explore the use of semantics in spam filtering by introducing a preprocessing step of Word Sense Disambiguation (WSD). Based upon this disambiguated representation, we apply several well-known machine-learning models and show that the proposed method can detect the internal semantics of spam messages.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Spam has become a significant problem for e-mail users over the past decade; an enormous amount of spam arrives in peoples' mailboxes every day. At the time of writing, 86.6% of all e-mail messages are spam, according to the Spam-o-meter website.[3] Spam is also a major computer security problem: it is a medium for phishing (i.e., attacks that seek to acquire sensitive information from end-users) (Jagatic et al. 2007) and for spreading malicious software (e.g., computer viruses, Trojan horses, spyware, and Internet worms) (Bratko et al. 2006).

Nevertheless, different studies show that the effect of spam in worldwide economy is notorious and prejudicial. Leung and Liang (2009) presented an analysis of the impact of phishing on the market value of global firms, which showed that phishing alerts pose significantly negative return on stock. In a similar vein, Mostafa Raad et al. (2010) offer another study to assess the influence and impact of spam in several companies whose email advertisement was considered as spam. Both examples clearly show the necessity to detect undesired messages and maybe more important, the need to restore the confidence of users in their e-mail filtering systems.

The simplest methods for filtering junk e-mail are usually blacklisting or signature-based (Carpinter and Hunt 2006). Blacklisting is a simple technique that is broadly used in most filtering products; such systems filter out e-mails from certain senders. In contrast, whitelisting systems (Heron 2009) deliver messages only from designated senders to reduce the number of misclassified legitimate e-mails (also known as 'ham' by the spam community). Another popular variant of these so-called banishing methods entails DNS blacklisting, in which the host address is checked against a list of networks or servers known to distribute spam (Jung and Sit 2004, Ramachandran et al. 2006).

In contrast, signature-based systems create a unique hash value (i.e., a message digest) for each known spam message (Kołcz et al. 2004). The main advantage of these methods is that they rarely produce false positives. Examples of signature-based spam filtering systems are Cloudmark,[4] a commercial implementation of a signature-based filter that is integrated with the e-mail server, and Razor,[5] a filtering system that uses a distributed and collaborative technique to spread signatures (Carpinter and Hunt 2006).

However, these simplistic methods have several shortcomings. First, blacklisting methods produce a high rate of false positives, making them unreliable as a standalone solution (Mishne et al. 2005). Second, signature-based systems are unable to detect spam messages until they have been identified, properly registered and documented (Carpinter and Hunt 2006).

A large amount of research has been dedicated to finding better spam filtering solutions. Machine-learning approaches have been effectively applied to text categorisation problems (Sebastiani 2002), and they have been adopted for use in spam filtering

* Corresponding author. Tel.: +34 944139003; fax: +34 944139166.
E-mail addresses: claorden@deusto.es (C. Laorden), isantos@deusto.es (I. Santos), borja.sanz@deusto.es (B. Sanz), gonzalo@iec.csic.es (G. Alvarez), pablo.garcia.bringas@deusto.es (P.G. Bringas).
[1] Tel.: +34 944139003; fax: +34 944139166.
[2] Tel.: +34 915618806; fax: +34 914117651.
[3] http://www.junk-o-meter.com/stats/index.php.
[4] http://www.cloudmark.com.
[5] http://razor.sourceforge.net.

1567-4223/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.elerap.2011.11.004

systems. Consequently, substantial work has been dedicated to naïve Bayes filtering (Lewis 1998); several studies on its effectiveness have been published (Androutsopoulos et al. 2000a,b,c; Schneider 2003, Seewald 2007). Another broadly embraced machine-learning technique is the Support Vector Machine (SVM) method (Vapnik 2000). The advantage of SVM is that its accuracy is not diminished when a problem involves a large number of features (Drucker et al. 1999). Several SVM approaches have been applied to spam filtering (Blanzieri and Bryl 2007, Sculley and Wachman 2007). Likewise, decision trees, which classify samples using automatically learned rule-sets (i.e., tests) (Quinlan 1986), have also been used for spam filtering (Carreras and Márquez 2001). All of these machine-learning-based spam filtering approaches are known as statistical content-based approaches (Zhang et al. 2004).

Machine-learning approaches model e-mail messages using the Vector Space Model (VSM) (Salton et al. 1975). The VSM is an algebraic approach for Information Filtering (IF), Information Retrieval (IR), indexing and ranking. This model represents natural language documents mathematically as vectors in a multidimensional space where the axes are terms within messages. As in any other IR system, the VSM is affected by the characteristics of the text, with one of those characteristics being *word sense ambiguity* (Sanderson 1994). The use of ambiguous words can confuse the model, permitting spammers to bypass spam filters.

We propose here the application of WSD for spam filtering to recover the filtering capabilities of content-based methods. Our approach pre-processes e-mails disambiguating the terms before constructing the VSM. Thereafter, based on this representation, we train several supervised machine-learning algorithms to detect and filter junk e-mails. In summary, we advance the state of the art through the following contributions:

- We present a method to disambiguate terms in e-mail messages.
- We provide an empirical validation of our method with an extensive study of several machine-learning classifiers.
- We show that the proposed method improves filtering rates; we discuss the weakness of the model and explain possible enhancements.

The remainder of this paper is organised as follows. Section 2 addresses the impact of electronic undesired mail on e-commerce. Section 3 describes the problem of WSD and the effects that ambiguity has on spam filtering systems. Section 4 introduces our method to improve detection rates by using WSD. Section 5 provides an empirical evaluation of the experiments performed and presents the results. Section 6 presents the main limitations of the proposed method and proposes possible enhancements. Finally, Section 7 presents the conclusions and outlines the avenues for future work.

2. Impact of undesired e-mail on e-commerce

Spam is a serious issue in the e-commerce arena, affecting many actors from the end users, to business offering commerce opportunities, to intermediaries. Correctly identifying spam, can have an impact on e-commerce, since false positives result the recipient not receiving legitimate e-mails (e.g., those used to conduct an advertising campaign chosen by the user itself), while false negatives can leave the recipient susceptible to spam attacks such as phishing.

On a thorough report back in 2004, Cashell et al. (2004) brought together different statistics on the economic impact of cyber-attacks. This report includes the analysis of a British firm, called Mi2g, which publishes analysis from the collection of data from 7,000 hacker groups worldwide, providing detailed monthly and year-to-date information on: digital attack hot spots, emerging threats to digital security, economic damage estimates, top hacker groups, most vulnerable operating systems and trends for vulnerabilities, spam, malware and denial of service attacks. Under the economic damage analysis, they include the estimation of the incidence and cost of what they call "overt digital attacks".[6] Fig. 1 shows the cost estimates for those digital attacks, which include hacking, malware and spam, from 1996 to 2003.

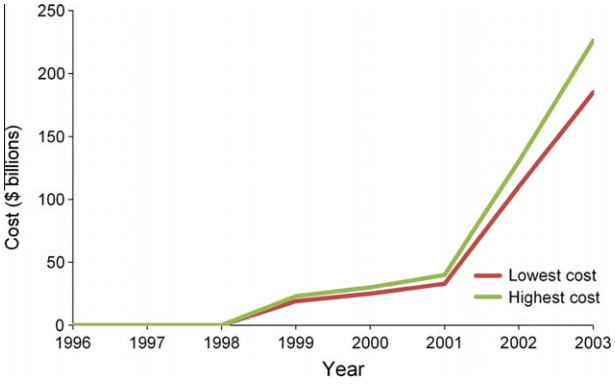

Fig. 1. Economic damage estimates for all forms of digital attacks worldwide, based on business interruption, denial of service, data theft or deletion, loss of sensitive intelligence or intellectual property, loss of reputation, and share price declines. *Source*: Mi2g, frequently asked questions: SIPS and EVEDA, v1.00.

Trying to break down the numbers, in another study, Hansell (2003) states that in 2003 the volume of spam, which was growing rapidly, implied worldwide costs exceeding 20 billion US dollars annually. And that is with "only" an estimated volume of 50% of e-mail being spam. Nowadays more than 86% of received e-mails are spam. In this way, although the numbers correspond to some years back in time, the projections to current days, according the increase of users with access to new technologies and the growth that electronic commerce has experienced, can be overwhelming.

Supporting that theory, in a more recent study, Smith et al. (2011) analyze the impact of cybercrime on marketing activity and shareholder value. Their results indicate that costs of cybercrime go beyond the tangible issues (e.g., stolen assets, business losses or damages on company reputation), having significant negative effect on shareholder value. The explanation to that fact, resides on the worries of users about security of their business transactions with companies that fall prey to cyber criminals. Such vulnerabilities result in a decrease of the trust from the user, causing the company to lose future business and, hence, raising the concerns of financial analysts, investors and creditors.

In a similar vein, other recent studies show the influence and impact of spam in several companies that suffered from considering their e-mail advertisement as a spam (Mostafa Raad et al. 2010) or the plague problem that the, in words of the on-line market research company e-Marketer, "killer-app of the on-line advertising world" (i.e., e-mail) is suffering as a result of spam (Gopal et al. 2011).

3. The problem of disambiguation

The task of disambiguating word sense is the process of identifying the most appropriate meaning of a polysemous word given a specific context. The Word Sense Disambiguation (WSD) problem

[6] Mi2g defines an overt digital attack as one in which a hacker group gains unauthorized access to a computer network and modifies any of its publicly visible components. Overt attacks may include either data attacks, where the confidentiality, authenticity, or integrity of data is violated, or control attacks, where network control or administrative systems are compromised. Overt attacks are those that become public knowledge, as opposed to covert attacks, which are known only to the attacker and the victim.

Download English Version:

https://daneshyari.com/en/article/379788

Download Persian Version:

https://daneshyari.com/article/379788

Daneshyari.com