Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 29 (2008) 19–33
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Time-efficient spam e-mail filtering using *n*-gram models

Ali Çıltık, Tunga Güngör *

Department of Computer Engineering, Boğaziçi University, İstanbul 34342, Turkey

Received 15 September 2006; received in revised form 6 July 2007
Available online 19 August 2007

Communicated by M.-J. Li

Abstract

In this paper, we propose spam e-mail filtering methods having high accuracies and low time complexities. The methods are based on the *n*-gram approach and a heuristics which is referred to as the first *n*-words heuristics. We develop two models, a class general model and an e-mail specific model, and test the methods under these models. The models are then combined in such a way that the latter one is activated for the cases the first model falls short. Though the approach proposed and the methods developed are general and can be applied to any language, we mainly apply them to Turkish, which is an agglutinative language, and examine some properties of the language. Extensive tests were performed and success rates about 98% for Turkish and 99% for English were obtained. It has been shown that the time complexities can be reduced significantly without sacrificing performance.

Keywords: Spam filtering; *n*-Gram model; Heuristics; Agglutinative language; Free word order; Morphology; Turkish

1. Introduction

Spam e-mail (or junk e-mail) messages are the messages that the recipients are exposed to without their approval or interest. We may also use the word ''unsolicited'' to name this kind of messages, since spam concept depends on the person who receives the e-mail. An unsolicited e-mail for a person may be regarded as legitimate (normal) by another person, and vice versa. In today's world where the Internet technology is growing rapidly and thus the communication via e-mail is becoming an important part of daily life, spam e-mail messages pose a serious problem. So it is crucial to fight with spam messages which tend to increase exponentially and cause waste of time and resources.

Past 1994, some spam prevention tools began to emerge in response to the spammers (people sending spam messages) who started to automate the process of sending spam e-mail. The very first spam prevention tools or filters used a simple approach to language analysis by simply scanning e-mail messages for some suspicious senders or for phrases such as ''click here to buy'' and ''free of charge''. In late 1990s, blacklisting, whitelisting, and throttling methods were implemented at the Internet Service Provider (ISP) level. However, these methods suffered some maintenance problems. Furthermore, whitelisting approach is open to forgeries. Some more complex approaches were also proposed against spam problem. Most of them were implemented by using machine learning methods. Naïve Bayes Network algorithms were used frequently and they have shown a considerable success in filtering English spam messages (Androutsopoulos et al., 2000). Knowledge-based and rule-based systems were also used by researchers for English spam filters (Apte et al., 1994; Cohen, 1996). As an alternative to these classical learning paradigms used frequently in spam filtering domain, genetic programming was employed (Oda and White, 2003). It required fewer computational resources, making it attractive for spam filtering application. Case-based reasoning for spam e-mail filtering was discussed in (Delany et al., 2005 and Deepak et al., 2006). Meta data were also taken into account in addition to the content of the e-mail by some researchers (Berger et al., 2005).

* Corresponding author. Tel.: +90 212 3597094; fax: +90 212 2872461.
E-mail address: gungort@boun.edu.tr (T. Güngör).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.07.018

Since spam filtering is thought as a kind of text classification, support vector machines (SVMs) (Moon et al., 2004; Tong and Koller, 2002) and latent semantic indexing (LSI) (Gee, 2003) used in classification tasks were also studied. In (Cardoso-Cachopo and Oliveira, 2003), various text classification methods were compared and it was found that SVM and *k*-nearest neighbor (*k*-NN) supported LSI performed best. Memory-based learning was used in (Sakkis et al., 2003) and the effect of several parameters such as the corpus size was studied.

It is possible to combine several spam filtering techniques within a single filter resulting in a more robust system (Sakkis et al., 2001). An arguable point in spam filtering domain is determining the performances of different systems relative to each other. It is not easy to arrive at a sound conclusion since systems are trained and tested on different and incomparable data sets. There exist only a few efforts for measuring the relative successes of algorithms. Lynam et al. (2006) compared the performances of several filters using Receiver Operating Characteristics (ROC) analysis and attempted to produce combined filters that outperform individual filters.

Besides trying to apply machine learning techniques to the spam problem, the research has also progressed in another direction. The solutions based on some protocols and standards form a different point of view to the problem. Authenticated SMTP (Simple Mail Transfer Protocol) and SPF (Sender Policy Framework) have been developed as tools that restrict the spammers dramatically. SPF has also increased the popularity of Authenticated SMTP. (http://tools.ietf.org/html/rfc2821; http://www.openspf.org/).

In this paper, we propose an approach for spam filtering that yields high accuracy with low time complexities. The research in this paper has several directions. First, we develop methods that work in much less time than the traditional methods in the literature. Previous work on spam filtering has mainly concentrated on performance issues and ignored execution times. In this work, we present two novel methods and consider some variations of each. We show that, despite the simplicity of these methods, the success rates lie within an acceptable range. Second, in relation with the first goal, we develop a heuristics based on an observation about human behavior for spam filtering. The plausibility of the heuristics is tested with different parameter values and we find that the heuristics provides a significant improvement in time. Third, we form two models, one of which follows the traditional classification by dividing the messages into two classes (spam and legitimate), while the other considers each e-mail as a separate class. We test each method, their variations, and the heuristics under these models. We also form a refinement of these models by combining them using a voting strategy. We observe that this combined model decreases the error rate significantly and achieves the best performances in the work. Fourth, we test the effect of some parameters and language properties on spam filtering. For this purpose, the words are subjected to morphological analysis, the words in *n*-grams are reordered, and the size of the data set is changed.

Though the approach proposed and the methods developed in this paper are general and can be applied to any language, we mainly apply them to Turkish, which belongs to the group of agglutinative and synthetic languages and owns a highly complex morphology (Kornfilt, 1997; Lewis, 2002). To the best of our knowledge, the sole research for filtering Turkish spam e-mail was given in (Özgür et al., 2004). By using artificial neural networks (ANNs) and Naïve Bayes, a success rate of about 90% was achieved. In the current work, by using the mentioned methods and the heuristics, we obtain a success rate of about 98% (and a lower time complexity), which indicates a substantial increase compared to Özgür et al. (2004). In addition to Turkish messages, in order to be able to compare the results of the proposed approach with the results in the literature, we tested on English e-mail messages. The results reveal that 99% success rate is possible without the use of the heuristics and nearly 98.5% success can be obtained when the heuristics is used. We thus conclude that great time savings are possible without decreasing the performance below an acceptable level.

The organization of the paper is as follows: Section 2 explains the proposed perception models and methods, the heuristics, the free word order implementation, and some issues specific to e-mail classification. Section 3 combines the models of the previous section. Section 4 gives the details of the data sets used in this work, explains the details of the experiments, and comments on the results. Section 5 is for the conclusions.

2. Perception models and *n*-gram methods

In this work, we aim at devising methods with low time complexities, without sacrificing performance. The first attempt in this direction is forming simple and effective methods. The methods proposed are based on the *n*-gram approach, which is used frequently to model phenomena in natural languages. We develop two simple variations of this approach, which yield high performance ratios for filtering spam messages.

The second attempt in this direction is exploiting the human behavior in spam perception. Whenever a new e-mail is received, we just read the initial parts of the message and then decide whether the incoming e-mail is spam or not. Especially in the spam case, nobody needs to read the e-mail till the end to conclude that it is spam; just a quick glance might be sufficient for our decision. We simulate this human behavior by means of a heuristics, which is referred to as the *first n-words heuristics*. According to this heuristics, considering the first *n* words of an incoming e-mail and discarding the rest can yield the correct class.

In this section, we first explain some issues about the preprocessing phase. This is followed by a detailed explanation of the methods and the underlying models. Then