Chapter 11
Enhanced Bio-Inspired Algorithms for Detecting and Filtering Spam

Hadj Ahmed Bouarara
University of Dr. Tahar Molay, Algeria

ABSTRACT

The internet era promotes electronic commerce and facilitates access to many services. In today’s digital society, the explosion in communication has revolutionized the field of electronic communication. Unfortunately, this technology has become incontestably the original source of malicious activities, especially the plague called undesirables email (SPAM) that has grown tremendously in the last few years. This chapter unveils fresh bio-inspired techniques (artificial social cockroaches [ASC], artificial haemostasis system [AHS], and artificial heart lungs system [AHLS]) and their application for SPAM detection. For the experimentation, the authors used the benchmark SMS Spam corpus V.0.1 and the validation measures (recall, precision, f-measure, entropy, accuracy, and error). They optimize the sensitive parameters of each algorithm (text representation technique, distance measure, weightings, and threshold). The results are positive compared to the result of artificial social bees and machine-learning algorithms (decision tree C4.5 and K-means).

1. INTRODUCTION

In recent years, engineers and decision makers are confronted daily to NP-difficult problems where the classical techniques are unable to find effective solutions. They affect generally all sectors (the design of mechanical systems, image processing, information retrieval, clustering ... ... etc.).

The current scientific world was considerably built up with the inaugural appearance of novel concepts and paradigms. Our nature aged more than 5 billion years represents the largest reserve of solutions and ideas for many kinds of problems. The human being started to explore them recently (initiated from the most minor thing as bacteria to the complex systems collection of human body).

Nowadays, for each encountered problem we must observe the nature; it may already have the same problem where it had found solutions, long years ago. In the digital community, the world celebrates the birth of new interesting paradigms known under the name of bio-inspired techniques. They have demonstrated theirs strength face to different challenges. The foremost part of our work is the modelling of three fresh bio-inspired algorithms:

1. Artificial social cockroaches (ASC) inspired from the lifestyle of cockroaches and their movement as a decentralised system without the conductor. They use the interaction between them and with their environment in order to be grouped under the most secure and attractive shelter for hiding.
2. Artificial haemostasis system AHS inspired from the mechanism of stopping the loss of blood (external haemorrhagic).
3. Artificial heart-lungs system AHLS mimicked from the functioning of blood oxygenation. Recently, the e-mail service has become enormously used, and the principal vector of communication in our digital society despite the emergence of social webs and the web 2.0 tools. Moreover, it permits to users with a mailbox (BAL) and address mail to exchange messages (picture, files, and text documents) from anywhere in the world via internet. Regrettably, among all the messages received by an individual in his mail box, we recognize two cases:
   a. Regular (HAM): The email (Ham message) sent by friends or by websites subscribed in.
   b. Irregular (SPAM): The unsolicited emails (junk e-mail) sent in bulk by malicious people (spammers).

According to the most recent report of the Radicati Group (2013), who supplies quantitative and qualitative researches with details on the e-mail, the security, and the social networks, has exhibited that 70-80% of email traffic is composed of spam. It is a rigorous problem in the electronic life, which presents the main challenge for Mail server administrators, and responsible of information organizations.

For that matter, several spam detection systems have seen the light based on learning machine algorithms and probabilistic techniques including Bayesian classification, artificial neural networks and text compression. It represents a supervised classification task, which has witnessed a burning interest from companies and particles. Merely, the spammer techniques were dramatically evolved where the conventional systems are inefficacious face to several limits:

- **Choice of Parameters**: The majority of spam detection techniques are based on the parameters (distance measure and messages pre-processing techniques). A poor choice of these parameters may cause a degradation of detection performance.
- **False Positive Problem**: The classification of spam messages as ham.
- **False Negative Problem**: The classification of ham messages as spam.
- **The Messages Language**: The language of messages may be written in (English, French, Chinese ... .etc.).
- **Ambiguity of Natural Language**: It is a major problem caused by the difference in the discourse of users. This is reflected by the difference of vocabulary used for expression of the contents of emails.
- **Representation of Spam Detection Results**: The majority of current spam detection systems are based on a primary visualization and interaction model. They present the spam message as an ordered list which makes the visualisation of all spam messages difficult.