Chapter 3
Comparative Study Between a Swarm Intelligence for Detection and Filtering of SPAM:
Social Bees vs. Inspiration From the Human Renal

Mohamed Amine Boudia
Dr. Tahar Moulay University of Saida, Algeria

Mohamed Elhadi Rahmani
Dr. Tahar Moulay University of Saida, Algeria

Amine Rahmani
Dr. Tahar Moulay University of Saida, Algeria

ABSTRACT

This chapter is a comparative study between two bio-inspired approaches based on swarm intelligence for detection and filtering of SPAM: social bees vs. inspiration from the human renal. The authors took inspiration from biological model and use two meta-heuristics because the effects allow the authors to detect the characteristics of unwanted data. Messages are indexed and represented by the n-gram words and characters independent of languages (because a message can be received in any language). The results are promising and provide an important way to use this model for solving other problems in data mining. The authors start this paper with a short introduction where they show the importance of IT security. Then they give a little insight into the state of the art, before starting the essential part of a scientific paper, where they explain and experiment with two original meta-heuristics, and explain the natural model. Then they detail the artificial model.

DOI: 10.4018/978-1-5225-3004-6.ch003

INTRODUCTION AND PROBLEMATIC

The appearance of the Internet and the incredibly rapid development of telecommunication technology have made the world a global village. The Internet has become a major channel for communication. Email is one among the tools for communication that Internet users take advantage of as it is available free of charge and supplies the transfer of files.

According to the most recent report of the Radicati Group (2014), who supplies quantitative and qualitative researches with details on the e-mail, the security, the Instant messaging (IM), the social networks, the archiving of the data, the regulatory compliance, the wireless technologies, the Web's technologies and the unified communications, there was exactly:

- 4.116 trillion Of active emails accounts in the world.
- 2.504 Billion People who use e-mails regularly to over 2.8 billion in 2018.
- 196,3billion is the number of e-mails that are sent to by day in 2014in the world on average. This number will increase to 227,7 billion in 2018.
- 1,6 is the number of accounts detained by each person and which should increase to 1,8 in four years.

According to the same reports of the Radicati Group, unsolicited mail, or SPAM, can reach more than 89,1%; 262 million SPAMS a day. Although the decision "spam / no-spam" is most often easy to take for a human. Messages in circulation, prevents address manually sorting the emails acceptable and others. Spam is a global phenomenon and massive. According to the CNIL (The National Commission of The computing and Freedoms), spam is defined as follows: "The" spamming "or" spam "is to send massive and sometimes repeated, unsolicited electronic mail, to individuals with whom the sender has had no contact and he has captured the email address erratically. ".

From the above statistics, the detection and filtering of spam is a major stake to the Internet community making the detection and filtering of spam a crucial task.

It is only the late 90s that the problem of detection and spam filtering by content, drew attention to three areas of research that were not directly affected by e-mail: the Information Retrieval (IR), the Data Mining (DM) and Machine Learning (ML).

This whole issue leads us to a study as to the representation of data (message corpus) to try to identify sensitive parameters that can improve the results of classification and categorization, around detections and spam filtering. We know very well that supervised learning techniques give the best results, and it is for this reason that we tried to experiment a new meta-heuristic to solve the problem of detecting and filtering spam.

The literature gives two broad approaches for the filtering and the detection of SPAM: The approach based on the machine learning and the approach not based on the machine learning. The first approach is based on feature selection which is an important stage in the systems of classification. It aims to reduce the number of features while trying to preserve or improve the performance of the used classifier. On the other hand, the second approach (not based on the machine learning) is based on many existing techniques and algorithms: content analysis, the block-lists, black-lists and white-lists, the authentication of mailbox and the heuristics and finally meta-heuristics.