PhD Thesis Proposal in Computer Science (MAP-i)

Title: Spam Telescope Miner: worldwide unsolicited email detection using data mining techniques

Introduction: Unsolicited email (or "spam") is a severe problem that is motivated by its economics, i.e. there is a low cost to reach a high number of consumers. Therefore, it is expected that unsolicited email will remain a pertinent issue in the next years. As an illustration, in 2004 around 50% of all world email traffic was considered to be spam. This problem affects both people and organizations due to factors such as intrusion of privacy, illegal scams or spread of virus, Internet traffic increase and time spent reading unwanted messages. The most used approach to stop unsolicited emails is based on filters, where spam is blocked at the server or client level. In particular, Data Mining (DM) techniques, such as Bayesian methods have become popular.

Objectives: This project proposes the design and development of a Spam Telescope, a longitudinal controlled study inspired in the Network Telescope concept, where the intention is the collection of a significant slice of the entire world spam traffic. One possibility is to use a high number of spam traps (i.e. fake emails set in free email services) around the world, under different user profiles (e.g. webmaster or normal user). Another strategy is to set several e-mail accounts within one or more controlled servers. Normal email ("ham") could be stored by using email donations in a web page or including messages from public mailing lists/newsgroups. With such rich distributed email database, it will be possible to identify differences between spam and ham and also how spam changes through time. Also, a Spam Telescope Miner evolutionary architecture is proposed, where several DM techniques (e.g. Bayesian methods or Neural Networks) are automatically optimized (evolved) for the best individual (competition) and collaborative performances. This architecture will be tested under a realistic "in vivo" approach, where emails are received in sequence and the spam/ham messages change through time.

The objectives of this thesis are:

1. to design and analyze the Spam Telescope (worldwide distributed email database); and
2. to design, develop and test the Spam Telescope Miner, an evolutionary architecture for collaboration and competition of filtering data mining techniques.

Notes: This thesis is integrated in the R&D project PTDC/EIA/64541/2006 funded by FCT (http://www.fct.mctes.pt) and there are funds for computer equipment and scientific conference traveling expenses. The project team includes 2 PhD researchers from CCTC (Pedro Sousa, Miguel Rocha) and consultants from University of Vigo and University College London.

Supervisor: Paulo Cortez (PhD in Computer Science), pcortez@dsi.uminho.pt, http://www.dsi.uminho.pt/~pcortez, Department of Information Systems, Univ. Minho, Guimarães, Portugal.

R&D Unit: Algoritmi/CCTC, Univ. Minho, Portugal.