TACKLING SPAM

IBM's prototype filter system for the automatic identification of unsolicited email messages

In recent years, electronic mail users around the world have noticed that an ever-increasing amount of unsolicited email – commonly known as spam – reaches their mailboxes. The contents of such email ranges from get-rich-quick schemes and low-priced printer cartridges, to stock tips, illegal substance offers, cheap pharmaceuticals, and information on web sites with pornographic material. Recent estimates place spam traffic to approximately 14 billion messages per day, or an average of approximately 25 messages per user per day! Despite the immensity of these numbers, it is only a relatively small number of people who are responsible for the generation of these unwanted messages.

Following the surge in the amount of spam email circulating, a number of methods have been proposed to address the problem and include the use of white/black-lists, bulk-email detection, various forms of filtering or some combination of the above. Within the filtering category, there are three sub-categories: Bayesian-based schemes, rule-based schemes, and similarity-based schemes. Bayesian methods are very good at identifying spam messages and typically exhibit low false-positive rates. On the other hand, rule-based methods apply heuristic tests on the headers or bodies of messages and can achieve good levels of recognition, but they require the explicit addition of active rules in the collection. Finally, similarity-based methods rely on comparing a candidate message with the ones in the spam message repository to draw conclusions. The performance of filtering methods typically suffers when a newly-arrived spam message is a 'pioneer' of sorts with no counterpart among the messages that are already in the spam repository.

IBM Research is developing an enterprise-class anti-spam filter as part of an overall strategy of attacking the spam problem on multiple fronts. Our anti-spam architecture, SpamGuru, mirrors this philosophy by incorporating several different filtering technologies and intelligently combining their output to produce a single quantitative measure of 'spamminess' for each incoming message. The use of multiple algorithms is expected to improve SpamGuru's effectiveness and make it more difficult for spammers to achieve their goal. SpamGuru's classification technologies include spoof detection, Bayesian filtering, plagiarism detection, automatically generated white - and black-lists, and Chung-Kwei, a novel technique that uses advanced pattern-matching algorithms developed by IBM's bioinformatics group.

Chung-Kwei is a prototype spam filtering system which capitalizes on our earlier pattern-discovery work on problems from computational biology that included protein annotation and gene finding. The name Chung-Kwei is taken from Feng-Shui and refers to a figure traditionally shown carrying a bat and holding a sword behind him, which is associated with the protection of expensive goods. Chung-Kwei follows an earlier IBM project in computer security from several years ago. At that time we developed 'Daemon-Watcher', a system that was also based on pattern-discovery and was able to automatically identify intrusion attempts based on the File Transfer-Protocol. Central to our earlier life science applications (as well as to Daemon-Watcher and Chung-Kwei) was a pattern discovery algorithm called Teiresias, which was first presented in 1998 and has since been used to effectively address a very wide spectrum of problems in the life sciences.

The idea underlying Chung-Kwei can be summarized as follows: given a collection of spam messages, we use Teiresias to discover patterns that appear two or more times in this collection (the instances can appear within messages as well as across messages of the collection). This stage takes place off-line and can be repeated as frequently as needed. Subsequently, new incoming email messages are examined to see if they match any of the collected patterns: the more patterns an email message contains, the more likely that it is bona fide spam (white email).

In order to test the system, a prototype implementation of Chung-Kwei was installed on a 2.2 GHz Intel-Pentium PC. Chung-Kwei was first trained using a knowledge-base of accumulated spam email comprising 65,000 messages. We randomly selected 10,000 white messages and used it for negative training. During testing on 66,967 spam email messages, Chung-Kwei correctly reported 64,665 as spam email with a resulting sensitivity of 96.56% and just a 0.066 % false positive rate. Based on these early findings, combined with the other techniques that make up the SpamGuru anti-spam filter, there is the potential to eliminate virtually all spam with minimal error rates. As of December 2004, the entire training phase (pattern discovery + negative training + formation of final vocabulary + processing of white email to decide thresholds) can be completed in under 15 minutes, while the memory requirements remain below 100 MB.

BIOGRAPHIES – Dr Isidore Rigoutsos and Tien Huynh

Dr Isidore Rigoutsos is Manager of the Bioinformatics and Pattern Discovery Group at the Computational Biology Center of IBM's Thomas J Watson Research Center in Yorktown Heights, New York, and Visiting Lecturer at the Massachusetts Institute of Technology. Dr Rigoutsos is a Fulbright Scholar, a senior member of the Institute of Electrical and Electronics Engineers (IEEE), a member of the International Society for Computational Biology (ISCB), the American Society for Microbiology, the American Association for the Advancement of Science (AAAS), and a Fellow of the American Institute for Medical and Biological Engineering (AIMBE). He is author of numerous peer-reviewed publications, and holds 13 US and 2 European patents.

Tien Huynh is an IBM researcher currently working in the Bioinformatics and Pattern Discovery Group at the Thomas J Watson Research center, Yorktown Heights, New York. He has a BS and MS degrees in computer science from Kent State University, Kent, Ohio.

Taking lessons from computational biology to improve the detection of spam Courtesy of IBM