Chapter 12
Spam Detection Approaches with Case Study Implementation on Spam Corpora

Biju Issac
Swinburne University of Technology (Sarawak Campus), Malaysia

EXECUTIVE SUMMARY

Email has been considered as one of the most efficient and convenient ways of communication since the users of the Internet has increased rapidly. E-mail spam, known as junk e-mail, UBE (unsolicited bulk e-mail) or UCE (unsolicited commercial e-mail), is the act of sending unwanted e-mail messages to e-mail users. Spam is becoming a huge problem to most users since it clutter their mailboxes and waste their time to delete all the spam before reading the legitimate ones. They also cost the user money with dial up connections, waste network bandwidth and disk space and make available harmful and offensive materials. In this chapter, initially we would like to discuss on existing spam technologies and later focus on a case study. Though many anti-spam solutions have been implemented, the Bayesian spam detection approach looks quite promising. A case study for spam detection algorithm is presented and its implementation using Java is discussed, along with its performance test results on two independent spam corpuses – Ling-spam and Enron-spam. We use the Bayesian calculation for single keyword sets and multiple keywords sets, along with its keyword contexts to improve the spam detection and thus to get good accuracy. The use of porter stemmer algorithm is also discussed to stem keywords which can improve spam detection efficiency by reducing keyword searches.

INTRODUCTION

Over the last years, unsolicited bulk mail, better known as spam, has become one of the most annoying problems of the Internet. The increase of spam emails uses bandwidth and fills up databases and therefore the global network becomes more crowded and less useful. Even though spam emails do not damage the data in the way that viruses do, they do harm the business intentions. For example, spam emails wastes user's time since the users devoid of anti-spam protection have to

DOI: 10.4018/978-1-60960-015-0.ch012

check which email is spam manually and then delete it. Sometimes, users can easily overlook or delete important email because of confusing it with spam. Email spamming often contains deceptive, worthless content or even a virus attachment.

Spam emails are getting better in its ability to break anti-spam filters and it would take a great deal of research to get it fully eradicated by coming up with very intelligent anti-spam filters. Spammers are also becoming more innovative, so that the anti-spam research is having a great relevance these days. There are various anti-spam techniques that have been created and implemented since spam started infiltrating user's inboxes. The most popular and direct way to prevent spam is the anti-spam filters. Anti-spam filters are the software tools that block spam messages automatically. These filters vary in functionality from black list (spammer list) and white list (trusted user list) to content-based filters. There are a lot of anti-spam filters or spam detection schemes available in the market.

The spammer's methods of avoiding detection evolve constantly, differing significantly from what has been used in the past. For every techniques created for filtering the emails, a new method to spread spam also comes out, making the battle between the spammers and mail agent even more challenging. We would like to introduce a Bayesian approach to the anti-spam solution, considering the context of keywords found. First we implement a simple Bayesian filter based on single keyword sets. Then we improve that by using multiple keyword sets and assigning a higher weightage to them. Finally, we further refine the anti-spam filter by using context matching technique along with the previous steps. The keywords are mapped to a keyword context, which is a collection of other keywords where the specific keyword is found.

The spam relayed by different countries in second quarter of 2007 is shown as a graph in Figure 1 (E-mail spam, na). This gives a good indication that some selected countries are the top relay points of spam emails. The actual spammer may or may not be sending spam emails from the country of his residence or may use compromised PCs elsewhere, even in other countries.

EXISTING AND RELATED WORKS

A number of research works are happening in the field of spam detection techniques. Some are listed below. Sasaki and Shinnou proposed a new spam detection technique using the text clustering based on vector space model. Their method computes disjoint clusters automatically using a spherical k-means algorithm for all spam/non-spam mails and obtains centroid vectors of the clusters for extracting the cluster description. For each centroid vectors, the label (`spam' or `non-spam') is assigned by calculating the number of spam email in the cluster. When new mail arrives, the cosine similarity between the new mail vector and centroid vector is calculated. Finally, the label of the most relevant cluster is assigned to the new mail (Sasaki & Shinnou, 2005). When classifying emails as spam and ham (which is a valid email), a false positive is the valid email that was erroneously classified as spam and a false negative is the spam email that was erroneously classified as valid email. For email classification as spam or non-spam, naive bayes classification was used in several systems (Kiritchenko & Matwin, 2001; Chan & Poon, 2004; Schneider, 2003; Androutsopoulos et al., 2000). Chiu et al. presents an alliance-based approach to classify, discovery and exchange interesting information on spam mails. The spam filter is built based on the mixture of rough set theory, genetic algorithm and XCS (eXtended Classifier System) classifier system (Chiu, Chen, Jeng, & Lin, 2007). Sirisanyalak et al. uses an email feature extraction technique for spam detection based on artificial immune systems that extracts a set of four features that can be used as inputs to a spam detection model (Sirisanyalak & Sornil, 2007). Dhinakaran et al. collected 400