University of Science and Technology

College for Graduate Studies & Academic Advancement

Features Reweighting and Similarity Coefficient Based Method for Email Spam Filtering

Thesis submitted in Partial Fulfillment of the Requirements for the Degree of Master of Information System

Student Name: Ahmed Osman Ali Elsiddig

Supervisor Name: Dr. Ammar Ahmed E. Elhadi

2013-2014

Abstract

Spam is flooding the Internet with many copies of the same message, in an attempt to force the message on people who would not otherwise choose to receive it.

Anti spam by determining whether or not an incoming email is spam has become an important problem. Various techniques for automatically detecting or filtering spam emails have been proposed. Many practical applications rely on building comprehensive databases for blocking emails whose addresses have been reported as black-lists or whose message bodies contain specific words or phrases defined as threatening terms.

One of the main characters or the problem of Spam filtering is its high dimension of space feature. For this reason, we need a reducing stage of dimensions.

This Thesis tried to cover this side from spam detection techniques by study the effect of re-weight of features.

The objective of this Thesis is to Study the similarity coefficient (Cosine and Dice) and Study the effects of the important feature to other features through the re-weight process.

The works started by applying similarity coefficient in the dataset, and then re-weight the features in the dataset and applying similarity coefficient in the new data set. Finally make a Comparison between the result before and after re-weight and Comparison with feature selection method.

The most important results of this thesis are: Reweighting process did not improve the success rate of any of the two methods (Cosine and Dice). Also, Feature selection method led to improve detection in Cosine, while reweighting method not improve detection any of (Cosine or Dice).

Arabic Abstract

الرسائل غير المرغوب فيها تغرق الإنترنت بسيل من الرسائل المكررة ، وذلك بهدف إجبار مستخدمي البريد الإلكتروني على إستقبالها .

الكثير من تقنيات الإكتشاف الآلي للرسائل غير المرغوب فيها تم إقتراحها . كما أن هنالك العديد من التطبيقات التي تعتمد على بناء قواعد بيانات شاملة لحجب رسائل البريد الإلكتروني التي لديها عناوين مصنفة ضمن القوائم السوداء أو التي تحتوي على كلمات أو عبارات معينةمعرفة مسبقاً بأنها تمثل تهديد .

من أهم المعضلات التي تواجه تقنيات اكتشاف الرسائل غير المرغوب فيها هي الأبعاد المتعددة لفضاء السمة. لهذا السبب برزت الحاجة لتقليل الأبعاد المتعددة لفضاء السمة.

هذا البحث حاول أن يغطي هذا الجانب من تقنيات إكتشاف الرسائل غير المرغوب فيها ، وذلك بدراسة أثر إعادة وزن المميزات على عملية إكتشاف الرسائل غير المرغوب فيها .

يهدف البحث لدراسة معامل التشابه (Cosine and Dice) ودراسة أثار أهم المميزات على المميزات الأخرى وذلك عن طريق عملية إعادة الوزن . لتحقيق هدف البحث بدء العمل بتطبيق معامل التشابه في قاعدة البيانات ، بعد ذلك طبقت عملية إعادة وزن المميزات في قاعدة البيانات ومن ثم تم تطبيق معامل التشابه على قاعدة البيانات الجديدة . أخيراً تمت مقارنة النتائج قبل وبعد عملية إعادة الوزن ومقارنة الناتج مع أسلوب إختيار المميزات .

أهم نتائج البحث هي : تم أثبات أن عملية إعادة الوزن لم يكن لها تأثير إيجابي في تحسين إكتشاف الرسائل غير المرغوب فيها لكل من (Cosine and Dice) . كما تم إثبات أن أسلوب إختيار المميزات أدى إلى تحسين عملية إكتشاف الرسائل غير المرغوب فيها في حالة Cosine ، بينما أسلوب إعادة الوزن لم تحسن نسبة الإكتشاف في كل الحالتين (Cosine and Dice) .

INTRODUCTION

1.1 An Overview

Email spam or junk mail, or unsolicited commercial email is process of sending not required email messages, frequently with commercial content, in large quantities to an indiscriminate set of recipients. Spam in email started to become a problem when the Internet was opened up to the general public in the mid-1990s. It grew exponentially over the following years, and today composes some 80 to 85 percent of all the e-mail in the World, by a "conservative estimate". Pressure to make email spam illegal has been successful in some jurisdictions, but less so in others. The efforts taken by governing bodies, security systems and email service providers seem to be helping to reduce the onslaught of email spam. According to "2014 Internet Security Threat Report, Volume 19" published by Symantec Corporation, spam volume dropped to 66% of all email traffic. Spamming users exploit this fact, and frequently outsource parts of their malpractices to countries where spammers could avoid legal pursuits. (Spamming, Wikipedia, 2015)

Spam causes over stacking that is both time-consuming to handle and resource intensive. Apart from that, a large number of organizations have been victims of spam that has an effect similar to a Distributed Denial of Service on the email system.

Spamming keeps economically tolerant because advertisers have no operating costs beyond the management of their mailing lists, and it is difficult to hold senders accountable for their mass mailings. Because the barrier to entry is so low, spammers are numerous, and the volume of unsolicited mail has become very high. In the year 2011, the estimated figure for spam messages is around seven trillion. The costs, such as lost productivity and fraud, are borne by the public and by Internet service providers, which have been forced to add extra capacity to cope with the deluge. Many countries could draft legal frames o spamming crises. (Spamming, Wikipedia, 2015)

There a release from European Union's Internal Market Commission guessing that in 2001 "junk email" cost Internet users €10 billion per year worldwide. The California legislature found that spam cost United States organizations alone more than $13 billion in 2007, including lost productivity and the additional equipment, software, and manpower needed to combat the problem. Spam's direct effects include the consumption of computer and network resources and the cost in human time and attention of dismissing unwanted messages. Prominent companies who are mainly spam targets utilize numerous techniques to check and combat spam. (Spamming, Wikipedia, 2015)

Computer viruses dispersion is one of spam characteristics, Trojan horses or other malicious software. The objective may be identity theft, or worse (e.g., advance fee fraud). Some spam attempts to capitalize on human greed, while some attempts to take advantage of the victims' inexperience with computer technology to trick them (e.g., phishing).

On May 31, 2007, one of the world's most prolific spammers, Robert Alan Soloway, was arrested by US authorities. Described as one of the top ten spammers in the world, Soloway was charged with 35 criminal counts, including mail fraud, wire fraud, e-mail fraud, aggravated identity theft, and money laundering. Prosecutors allege that Soloway used millions of "zombie" computers to distribute spam during 2003. This is the first case in which US prosecutors used identity theft laws to prosecute a spammer for taking over someone else's Internet domain name. This shows too late response of the world community against spam problems. (Spamming, Wikipedia, 2015)

In an attempt to assess potential legal and technical strategies for stopping illegal spam, a study from the University of California, San Diego, and the University of California, Berkeley, "Click Trajectories: End-to-End Analysis of the Spam Value Chain", cataloged three months of online spam data and researched website naming and hosting infrastructures. The study concluded that: (Spamming, Wikipedia, 2015)
this treatise unveiled that half of all spam programs have their domains and servers distributed over just eight percent or fewer of the total available hosting registrars and autonomous systems, with 80 percent of spam programs overall being distributed over just 20 percent of all registrars and autonomous systems;

also I disclosed that 76 purchases for which the researchers received transaction information, there were only 13 distinct banks acting as credit card acquirers and only three banks provided the payment servicing for 95 percent of the spam-advertised goods in the study; and,

We can indicate that a "financial blacklist" of banking entities that do business with spammers would dramatically reduce monetization of unwanted e-mails. Moreover, this blacklist could be updated far more rapidly than spammers could acquire new banking resources, an asymmetry favoring spam controlling potentials.

To indicate reality of spam legal control, thereupon, in June 2007, two men were convicted of eight counts stemming from sending millions of e-mail spam messages that included hardcore pornographic images. Jeffrey A. Kilbride, 41, of Venice, California was sentenced to six years in prison, and James R. Schaffer, 41, of Paradise Valley, Arizona, was sentenced to 63 months. In addition, the two were fined $100,000, ordered to pay $77,500 in restitution to AOL, and ordered to forfeit more than $1.1 million, the amount of illegal proceeds from their spamming operation. The charges included conspiracy, fraud, money laundering, and transportation of obscene materials. The date June 5 revealed the first registered charges under the CAN-SPAM Act of 2003, according to a release from the Department of Justice. The specific law that prosecutors used under the CAN-Spam Act was drafted to crack down on the transmission of pornography in spam activities. (Spamming, Wikipedia, 2015)

So, Spam is no more garbage but risk since it recently includes virus attachments and spyware agents which make the recipients' system ruined, therefore, there is an emerging need for spam detection.

Many spam detection techniques based on machine learning algorithms have been proposed. As the amount of spam has been increased tremendously using bulk mailing tools, spam detection techniques should deal with it. For spam detection, parameters optimization and feature selection have been proposed to reduce processing overheads with guaranteeing high detection rates. (Lee et al, 2010)

1.2 Problem Definition

One of the main characters or the problem of Spam filtering is its high dimension of space feature. The feature space that contains words or phrases in the documents has more than ten thousands features, which is a great preventive problem for many of the machine learning algorithms. (Beiranvand, Osareh & Shadgar, 2012) For this reason, we need a reducing stage of dimensions.

The previous approaches have not taken into account the importance of weights of features and there are no previous studies discuss this topic. So, in this Thesis, we tried to cover this side from spam detection techniques by study the effect of re-weight of features.

1.3 Research Objective

The objective of this Thesis is:

A. Study the similarity coefficient (Cosine and Dice).
B. Study the effects of the important feature to other features through the re-weight process.

1.4 Scope of Research

The data used in this Thesis, are the dataset available in Hewlett-Packard Labs. It was generated in June-July 1999. The dataset is available at ftp://ftp.ics.uci.edu/pub/machine-learningdatabases/ spambase/.

The number of Instances in the data set are: 4601 (1813 Spam and 2788 Non-Spam), while the number of Attributes are: 58 (57 continuous, 1 nominal class label).

The works will start by applying similarity coefficient in the dataset, and then re-weight the features in the dataset and applying similarity coefficient in the new data set. Finally make a Comparison between the result before and after re-weight and Comparison with feature selection method.

1.5 Thesis Structure

After this introductory section, starting **Chapter II** Literature Review, it's include introduction to information security, Spam, techniques or spam filter, similarity coefficient, feature selection, then we will review the related work. **Chapter III** is about

Research Methodology; it's including Preprocessing Data, and the experiment phases and tools. **Chapter IV** contains Experimental Results and Analysis. **Chapter V** contains Conclusions and recommendations.