IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 4, Issue 3, October 2015)

Survey of Email Spam Analysis System

Aastha D. Shah
Dwarkadas J. Sanghvi College of Engineering, Mumbai, India
aasthashahdjsce14@gmail.com

Aashana D. Shah
Dwarkadas J. Sanghvi College of Engineering, Mumbai, India
aashanashahdjsce14@gmail.com

Sindhu Nair
Dwarkadas J. Sanghvi College of Engineering, Mumbai, India
Sindhu.Nair@djsce.ac.in

Abstract: **E-mail is an effective tool for communication as it saves a lot of time and cost. But e-mails are also affected by attacks which include Spam Mails. The problem of email spam has become so popular and wide spread that it has been the subject of a Data Mining Cup contest as well as numerous class projects. Hence in this paper we have presented the problem of email spam, how spams are dangerous and ways of preventing or avoiding it. We have also covered the importance of classification using datasets and clustering along with the description of the spam filters. The criteria and the need for detecting spam mails have been elaborated in this paper. A survey of email spam analysis which includes email spam detection and correction is made.**
Keywords: **Spam, ham, classification, clustering, criteria, spam filters, learning approaches.**

I. INTRODUCTION

Emails today are a fast and inexpensive mode of sharing personal and business information in a convenient way. They are delivered at once around the world. No other form of written communication is as fast as an email. But its simplicity and ease of use has also made it a hub of scams. Often we find our inbox full of undesirable mails. These kind of unwanted and unimportant mails are better known as SPAM Mails. So it has become essential to have reliable tools to detect spam and ham mails. Spam Mail is the practice of frequently sending unwanted data or bulk data in a large quantity to some email accounts. Spam is an unfortunate problem on the internet. SPAM is also called as Unsolicited Commercial Email (UCE) and Unsolicited Bulk.
Email (UBE) [1]. It is the bane of email communication. Email spam targets individual users with direct mail messages [1].

1.1 Statistics

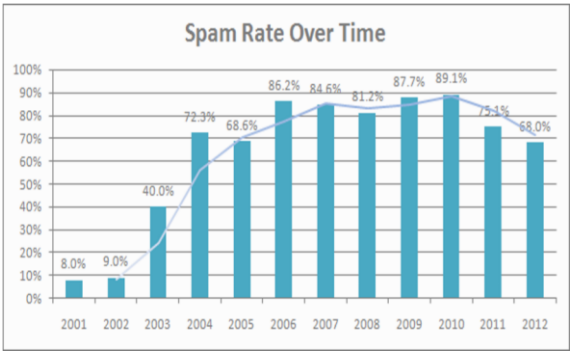

Figure 1: Spam rate over time

1.2 Definitions

Spam: In computing terms, spam means something unwanted.
Ham: The opposite of spam email that is wanted [20].

1.3 Cost of a spam

Spam is very cheap to send, the cost are insignificant as compared to conventional marketing techniques, so marketing by spam is very cost-effective, despite very low rates of purchases in response. But it translates into major costs for the victim [6].

1.4 Criteria for an email spam

Email spam is any email that meets the following criteria:

Unsolicited:
Unsolicited emails refer to unwanted, uninvited or not asked for emails. These mails are not requested by the recipients.
Anonymity:
No revealing of the identity and address of the sender.
Mass Mailing:
It refers to an act of sending the same email message to a large number of people at the same time.

With the expansion of web usage, the problem of spam mails is also expanding. It is an expensive problem that costs billions of dollars per year to service providers for loss of bandwidth [2].

II. SPAM FILTERS

When it comes to fighting spam, there are two different approaches that should be both considered and applied: Spam **prevention** and **filtering**. Spam prevention refers to prevention of the spam before it has actually taken place whereas spam filtering refers to detecting the spams after its occurrence by trying to automatically differentiate between good and bad emails. A spam filter is applied on emails [11]. Like other types of filtering programs, a spam filter looks for certain criteria on which it bases judgments. For example, the simplest and earliest versions (such as the one available with Microsoft's Hotmail) can be set to watch for particular words in the subject line of messages and to exclude these from the user's inbox [3]. Most often

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 4, Issue 3, October 2015)

the term email filtering refers to the automatic processing of incoming messages. The block diagram of a spam filter is as shown below.

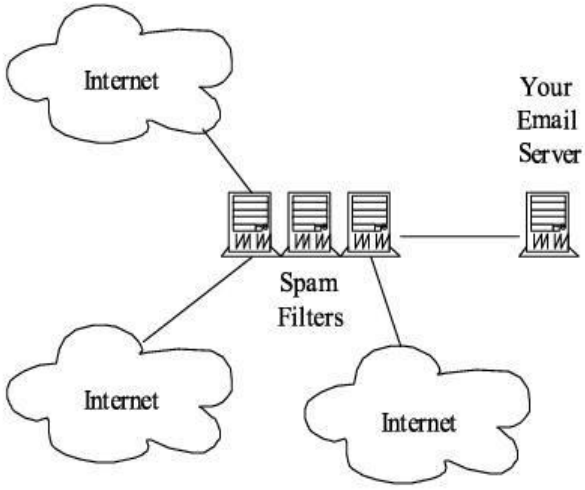

Figure 2: Use of Spam filters

III. CLASSIFICATION OF SPAM MAILS

Classification is the separation of objects into classes. If the classes are created without looking at the data then the classification is known as apriori classification. The significance of classification of mails is to gather similar emails at one place for easy access [2]. So in case of spamming due to replication of mails, classification proves as an effective method for clearing spam emails. If classes are created by looking at the data then the classification method is known as posterior classification.

Data Classification

Data classification is a two-step process:

Step 1: A model is built describing a predetermined set of data classes which is constructed by analyzing database tuples described by the attributes. Each tuple is assumed to belong to one of the existing class, as determined by the class label attribute.

Step 2: This model is used for classification. First the predictive accuracy of the model is estimated and then this accuracy of a model on a given test data set is the percentage of test set samples that are correctly classified by the model. For each test sample the known class label is compared with the learned model's class prediction for that sample [2].

IV. CLUSTERING OF SIMILAR EMAILS

What do u mean by clustering?

Clustering is assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters. It is one of the most important unsupervised learning problems. Clustering improves the system's availability to users, its aggregate performance, and overall tolerance to faults and component failures. Clustering is a data reduction step i.e. after the clusters of email documents are formed, we select only the 'spammy' clusters and then the subsequent steps are applied only to the selected clusters [4]. Kmeans [5] is one of the simplest clustering algorithms. It attempts to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

V. SOLUTIONS TO SPAM

Several solutions have been proposed to the spam problem. Several solutions involve detection and filtering of the spam emails on the client side.

Machine learning approaches have been used in the past for this purpose. Some examples of machine learning approach includes : Bayesian classifiers as Naive Bayes, Rippe and SVM [4]. In many of these approaches, Bayesian classifiers were observed to give good results, so they have been widely used in several spam filtering software.

5.1 Learning Approaches

Let us learn more about the learning approaches:

1. Navïe Bayes:

The navïe Bayes technique is based on Bayesian approach hence it is a simple, clear and fast classifier. A navïe Bayes classifier is a simple probabilistic classifier with strong independence assumptions [8]. A navïe bayes classifier assumes that the presence or absence of a particular feature of a class is not related to the presence or absence of any other feature, given the class variable depending on the nature of probability model. It only requires a small amount of training data to estimate the parameters required for classification. Bayesian classifiers are basically statistical classifiers which means they can predict the class membership probabilities [2].

Algorithm:

- In Bayesian classification we have a hypothesis that the given data belonging to a particular class.
- We then calculate the probability for the hypothesis to be true.

  $$P(A/B) = P(B/A)\, P(A)\, P(B)$$
- Once these probabilities have been computed for all the classes, we simply assign X to the class that has highest probability.

  $$P(Ci/X) = [P(X/Ci)\, P(Ci)] / P(X)$$

Where,

P (Ci/X): probability of the object X belonging to a class Ci

P(X/Ci): probability of obtaining attribute values X if we know that it belongs to class Ci P (Ci): probability of any object belonging to a class Ci (eg. C1, C2, C3 and so on) without any other information

P(X): probability of obtaining attribute values X whatever class the object belongs to.

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 4, Issue 3, October 2015)

2. Support Vector Machine:

Support Vector Machine abbreviated as SVMs, are based on the concept of decision planes that define decision boundaries. They are supervised learning models with associated learning models that analyse data and are mainly used for classification purpose [7]. As an input, (SVM) takes a set of data and output the prediction that data lies in one of the two categories that is it classify the data into two possible classes. A support vector machine performs classification by constructing an N-dimensional hyper plane that optimally categorizes the data in two categories. SVMs are set of related supervised learning methods used for classification [2].

Example of a linear classifier:

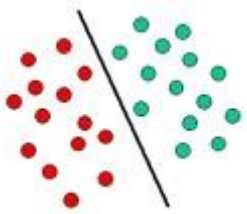

Figure 3: A Linear classifier

Let us consider two classes i.e. class c1 for RED and c2 for GREEN. Hence the objects belong to either class c1 or class c2. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object falling to the right is classified as GREEN or as RED should fall to the left of the separating line. This was a simple example of a classic linear classifier [10]. However, most classification tasks are not that simple, and hence often more complex structures are needed in order to make an optimal separation.

Spam email detection:

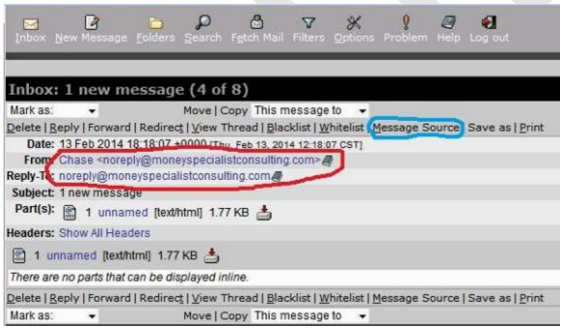

Figure 4: An Example of spam detection

The image above shows the normal headers for the email. Note that the From: header says Chase, but the email address, which is circled in red, is noreply@moneyspecialistconsulting.com.noreply@moneyspecialistconsulting.com. Hence this is a clear indication! It's not from chase.com, nor is it even a good attempt to cover this up, like chasebanking.com or bankofchase.com.

VI. ANTI-SPAM TECHNIQUES

There is no predefined method or technique as such used to prevent email spams. No technique is a complete solution to the problem of spam. Anti-spam techniques can be broken into four broad categories and they are, those that require actions by individuals, those that can be automated by email administrators, those that can be automated by email senders and those employed by researchers and law enforcement officials [12].

6.1 Detecting Spam

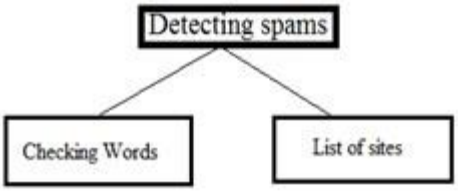

Figure 6: Detecting spam categorization

6.1.1 Checking Words: False positives

This technique is based on detecting the spam based on the content of the email. Content based statistical means or detecting keywords can be very accurate. They can be accurate only if they are correctly tuned to the types of legitimate (Valid or justifiable) email that an individual gets, but they can also make mistakes. For example: detecting the keyword "cialis" in the word "specialist". Hence if the keyword is such as “viagra”, then if anyone sends you a joke that mentions “Viagra”, the content filters can hence easily mark it as being spam thought it was not unsolicited nor sent in bulk. Non-content base statistical means can help lower false positives. It looks at statistical means vs. blocking based on content/keywords so you will be able to receive a joke that mentions "viagra".

6.1.2 List of Sites

The most popular DNS Blacklists are lists of domain names of known spammers. A DNS based Blackhole List (DNSBL) or Real-time Blackhole List (RBL) is an effort to stop email spamming [13]. DNSBL is a "blacklist" of locations on the Internet reputed to send email spam. The locations consist of IP addresses or networks linked to spamming. The term Blackhole List is sometimes interchanged with the term "blacklist" or "blocklist". This is used to block mail coming from systems that are not compliant with the RFC standards.

VII. END-USER TECHNIQUES

There are various techniques a user can use to avoid email spams. Some of them are as follows:

a. Discretion: It focuses on sharing an email address only among a limited group of correspondents to limit spam. This method relies on the discretion of all members of the group that is as disclosing email

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 4, Issue 3, October 2015)

addresses outside the group to maintain the trust relationship of the group [12].

b. **Avoid responding to spam:** Spammers often regard responses to their messages. Even the responses like "Don't spam me" as confirmation that an email address is valid and so one must avoid responding to the spam.
c. **Ham passwords:** Systems that use ham passwords ask unrecognized senders to include a password in their email that demonstrates that the email message is a "ham" message. Ham passwords are often combined with filtering systems and it would be included in the "subject" line of an email message [15].
d. **Contact forms:** Contact forms allow users to send email only after filling out forms in a web browser [12]. However, such forms are at times inconvenient to the users.
e. **Disposable email addresses:** Disposable email address means a temporary address. The user can disable or abandon the temporary address which forwards email to the real account.
f. **Disable HTML in email:** The display of HTML, URLs, and images can easily expose the user to the offensive images in spam.
g. **Reporting spam:** One must not ignore a spam because reporting of spam can lead to tracking down a spammer's ISP. Also reporting the offense can lead to the spammer's service being terminated.
h. **Address munging:** Email address munging is an act of using ASCII, JavaScript and scrambling of letters in your email address in order to hide your email address and avoid email address harvesting. It makes use of fake name and address. Care must be taken by users in ensuring that the fake address is not valid [14].

Other methods include use of honeypots, Hybrid filtering and Pattern detection to stop spams before it gets to the user.

VIII. EMAIL DATA CLEANING

Data cleaning is an important area in data mining. Many text mining applications need to take emails as inputs. Email filtering, email routing, email analysis, information extraction and newsgroup analysis are few such applications. But unfortunately, Email data can be very noisy. Hence the first step that needs to be followed is data cleaning from the emails. Data cleaning is important in order to attain high quality email mining. Data cleaning services include the process of detecting and correcting errors and inconsistencies from a data set in order to improve its quality. Several products have email cleaning features. For example: eClean 2000 and WinPure ListCleaner Pro are Data cleaning products. eClean 2000 is a tool that can clean up emails by removing extra spaces between words, removing extra line breaks between paragraphs, removing email headers, and re-indenting forwarded mails. The email data cleaning is performed by the rules defined by the user [16].

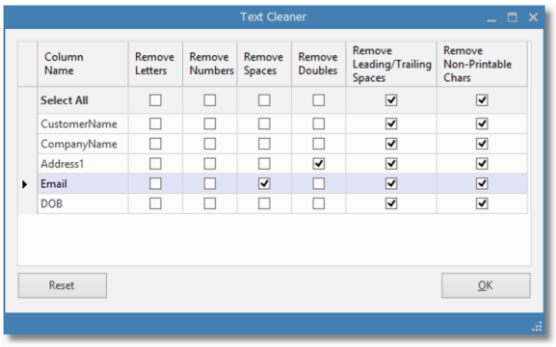

Figure 7: Text Cleaner

When we think of email data cleaning, questions that arise for email data cleaning are:

1. Data involves various factors like different levels, different factors and hence is very complex so the question that arises here is how to formalize that problem.
2. Once the problem is formalized, how to solve the problem in a principled approach is the next query.
3. Ultimately, how to make an implementation.

Answers to the above questions:

1. Here, we formalize email data cleaning as that of non-text filtering and text normalization. Irrelevant non-text data includes header, signature, quotation and program code filtering. Such irrelevant non-text data is eliminated in data cleaning.
2. Proposed to conduct email cleaning in a "cascaded" fashion.
   Cascade: A succession of stages, processes, operations, or units. That is we clean up the email using several phases or stages [19]. The phases includes, non-text filtering which is at email body level and text normalization which is at word, paragraph level.
3. It differentiates the tasks that can be accomplished using the existing methodologies and those which cannot be.

8.1 Tabular Data

At its simplest tabular data is data that is stored in rows and columns, either in a flat file or a database, and is usually comprised of simple alphanumeric values.

8.2 Tabular Data Cleaning

Tabular data cleaning aims at detecting and eliminating duplicate information or data when data comes from

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 4, Issue 3, October 2015)

different sources. It is not similar to text data cleaning. It involves removing duplicate rows, duplicate columns, finding and replacing text in rows or columns, changing the case of text, merging and splitting columns, transforming and rearranging rows and columns [17]. Reconciling table data by joining or matching and so on. Tabular data cleaning has been investigated at both schema level and instance level. Some products like SQL Server 2005 provides a tool for tabular data cleaning called Fuzzy Grouping [16;18].

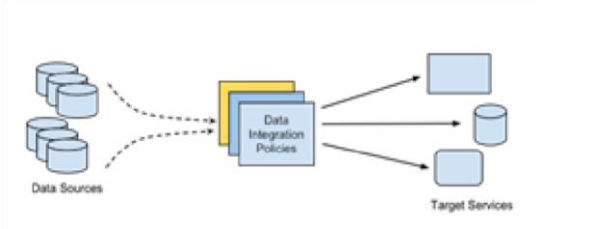

Figure 8: Tabular data cleansing process

IX. CONCLUSION

In this paper we have presented what is spam, an email spam, how spams are dangerous and ways of preventing or avoiding it. Thus we have studied detection methods for spam emails and their recovery process. Each term spam, clustering, clustering, properties of filtering is explained in this paper.

REFERENCES

[1] In vivo spam _ltering:A challenge problem for data mining, Tom Fawcett Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA USA; Available: tom.fawcett@hp.com
[2] Spam Mail Detection through Data Mining – A Comparative Performance Analysis, I.J. Modern Education and Computer Science, 2013, 12, 31-39 Published Online December 2013 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijmecs.2013.12.05.
[3] Comparison and analysis of spam detection algorithms, In International Journal of Application or Innovation in Engineering & Management (IJAIEM), Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com Volume 2, Issue 4, April 2013, ISSN 2319 – 4
[4] Detecting E-mail Spam Using Spam Word Associations, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 4, April 2012)
[5] MacQueen, J. 1967. ―Some Methods for Classification and Analysis of Multivariate Observations.‖ Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66) Vol. I: Statistics, pp. 281–297.
[6] L. F. Cranor and B. A. LaMacchia. Spam! CACM, 41(8):74{83, August 1998.
[7] Vapnik V N. Statistical learning theory [M]. John W iley &Sons, NewYork, N Y, 1998.
[8] Ian H, Witten and Eibe Frank.Data Mining: Practical machine learning tools and techniques", 2nd Edition. San Fransisco: Morgan Kaufmann; 2005.
[9] Learning Mixtures of Linear Classifiers, Dept. of Electrical Engineering & Dept. of Statistics, Stanford University, 350 Serra Mall, Stanford, CA 94305
[10] Survey on Spam Filtering Techniques, Communications and Network, 2011, 3, 153-160 doi:10.4236/cn.2011.33019 Published Online August 2011 (http://www.SciRP.org/journal/cn)
[11] From Wikipedia, the free encyclopedia: "Antispam techniques"
[12] From Wikipedia, the free encyclopedia: "Spam blacklist"
[13] Email Address Harvesting: How Spammers Reap What You Sow, Federal Trade Commission, 24 April 2006.
[14] David A. Wheeler, (May 11, 2011) Countering Spam by Using Ham Passwords (Email Passwords).
[15] Email Data Cleaning, Department of Computer Science, Tsinghua University 12#109, Tsinghua University Beijing, China, 100084
[16] M. A. Hernández and S. J. Stolfo. Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. Publisher: Kluwer Academic Publishers Hingham, MA, USA. Vol. 2, Issue 1, January 1998, pages 9-37.
[17] Fuzzy Lookup and Fuzzy Grouping in Data Transformation Services for SQL Server 2005. http://msdn.microsoft.com/library/default.asp?url=/li brary/en-us/dnsql90/html/FzDTSSQL05.asp.
[18] The Free Dictionary by Farlex: "cascade".
[19] Ham v Spam: what's the difference?, by Christine Barry – Chief Blogger, Oct 2nd, 2013.