A RULE BASED APPROACH FOR SPAM DETECTION

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
In
Computer Science & Engineering

By:
Ravinder Kamboj
(Roll No. 800832030)

Under the supervision of:

Dr. V.P Singh
Assistant Professor
Computer Science & Engineering

Mrs. Sanmeet Bhatia
Assistant Professor
Department of SMCA

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

JULY- 2010

Certificate

I hereby certify that the work which is being presented in the thesis entitled, **"A Rule Based Approach for Spam Detection "**, in partial fulfillment of the requirements for the award of degree of Master of Engineering in Software Engineering submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of Dr. V.P Singh and Mrs. Sanmeet Bhatia, and refers other researcher's works which are duly listed in the reference section.

The matter presented in this thesis has not been submitted for the award of any other degree of this or any other university.

(Ravinder Kamboj)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. V.P Singh)
Assistant Professor
Computer Science and Engineering Department
Thapar University
Patiala

(Mrs. Sanmeet Bhatia)
Assistant Professor
SMCA
Thapar University
Patiala

Countersigned by

15/07/10

(RAJESH BHATIA)
Head
Computer Science & Engineering, Department
Thapar University
Patiala

24/7/10

(R.K.SHARMA)
Dean (Academic Affairs)
Thapar University,
Patiala.

Acknowledgement

I would like to express my sincere gratitude to my supervisors **Dr. V.P Singh** and **Mrs. Sanmeet Bhatia** for their immense help, guidance, stimulating suggestions and full time encouragement. They always provide a motivating and enthusiastic atmosphere to work with, it was a great pleasure to do thesis under their supervision.

I am thankful to **Dr. Rajesh Bhatia,** Head of Department, Computer Science & Engineering Department and **Mrs. Inderveer Channa**, P.G. Coordinator, for the motivation and inspiration that triggered me for the thesis. I would also like to thank all the staff members (Kalam Singh, Jasleen Kaur etc.) who were always there at the need of the hour and provided with all the help and facilities, which was required for the completion of the thesis.

I am also thankful to my friends like: Yogesh Kumar, Vandana Ladha, Tikka Singh, Virender Kumar, Bhisham Sharma, Mudit Kumar, Nitesh Kumar Singh, Pawan Kumar, and all others. They provided me all the help which was required for the completion my thesis work.

Last but not the least, I express my heartfelt thanks to my parents for their blessings encouragement which helped me to stay calm during hours of frustration.

Ravinder Kamboj
(800832030)

ii

Abstract

Spam is defined as a junk Email or unsolicited Email. Spam has increased tremendously in the last few years. Today more than 85% of e-mails that are received by e-mail users are spam. The cost of spam can be measured in lost human time, lost server time and loss of valuable mail. Spammers use various techniques like spam via botnet, localization of spam and image spam. According to the mail delivery process anti-spam measures for Email Spam can be divided in to two parts, based on Emails envelop and Email data. Black listing, grey listing and white listing techniques can be applied on the Email envelop to detect spam. Techniques based on the data part of Email like heuristic techniques and Statistical techniques can be used to combat spam. Bayesian filters as part of statistical technique divides the income message in to words called tokens and checks their probability of occurrence in spam e-mails and ham e-mails. Two types of approaches can be followed for the detection of spam e-mails one is learning approach other is rule based approach. Learning approach required a large dataset of spam e-mails and ham e-mails is required for the training of spam filter; this approach has good time characteristics filter can be retrained quickly for new Spam. But has very less space characteristics. Knowledge obtained from this method is tough to share with other users and mail servers. Second approach is rule based approach. It is used as direct approach by implementing rules for various kinds of Spams.

For thesis work rule based approach has been followed. The intent is to implement rules for the various kinds of spam like: health, adult, educational and product offering Spam. Pattern analysis is performed on set of ham (legitimate) e-mails and spam e-mails. Corresponding probabilities for occurrences of various words/tokens has been calculated to design rules for selected tokens.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

We are living in the age of information and technology. IT has great impact on our life. Now people are going away from the traditional ways of information and communication. One prefers to use e-mail, instant messages and mobile messages to communicate over the traditional methods of communication like postal mails. These new ways of communication are very fast and interactive. Now people prefer reading news from the newsgroups or news websites rather than reading news papers. These ways of information and communication are facing a serious and irritating problem known as spam.

1.1 Definition of Spam

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. Mostly spam is commercial advertisements, often for dubious products, get-rich-quick schemes, or dating services. Spam is one of the major threats to network security. Spam costs the sender very little to send but most of the costs are paid by the recipient or the service providers rather than by the sender. User lost time while following the spam mail and ISPs lost bandwidth for carrying spam. While most recognized form of spam is e-mail spam but other methods of communication and information technology are facing same problem.

1.2 Types of Spam

Basically, spam can be categorized into the following four types:

1. Usenet Spam
2. Instant messaging Spam
3. Mobile Spam
4. E-mail Spam

1.2.1 Usenet Spam

Usenet spam is posting of some advertisement to the newsgroups. Spammers target the users those read news from these newsgroups. Spammers post advertisement to large amount of newsgroups at a time. Usenet spam robs users of the utility of the newsgroups by overwhelming them with a barrage of advertising or other irrelevant posts [11].

1.2.2 Instant Messaging Spam

Instant messaging systems, such as Yahoo! Messenger, AIM, Windows Live Messenger, Tencent QQ, ICQ, XMPP, and MySpace chat rooms, are all targets for spammers. Many IM systems offer a directory of users, including demographic information such as age and sex. Advertisers can gather this information, sign on to the system, and send unsolicited messages, which could include commercial scam-ware, viruses, and links to paid sites. As instant messaging tends to not be blocked by firewalls; therefore, it is an especially useful channel for spammers. It targets the users when they join any chat room to find new friends. It spoils enjoy of people and waste their time also.

1.2.3 Mobile Phone Spam

Mobile phone spam is directed at the text messaging service of a mobile phone. This can be especially irritating to customers not only for the inconvenience but also because of the fee they may be charged per text message received in some markets. This kind of spam usually contains some schemes and offers on various products. Sometimes service providers also make use of this to trap the user for activation of some paid service.

1.2.4 E-mail Spam

Email spam is the most recognized form of spam. E-mail spam targets the individual users with direct mails. Spammers create a list of e-mail users by scanning Usenet postings, stealing internet mail lists, search web for e-mail addresses. E-mail spam costs money to user of e-mail because while user is reading the e-mails meter is running. E-mail spam also costs the ISPs because when a bulk of spam mails are sent to the e-mail users its wastes the band width of the service providers these costs are transmitted to

users. All unwanted e-mails are not spam e-mails as figure1.1 shows a scenario about incoming e-mails [11, 2].

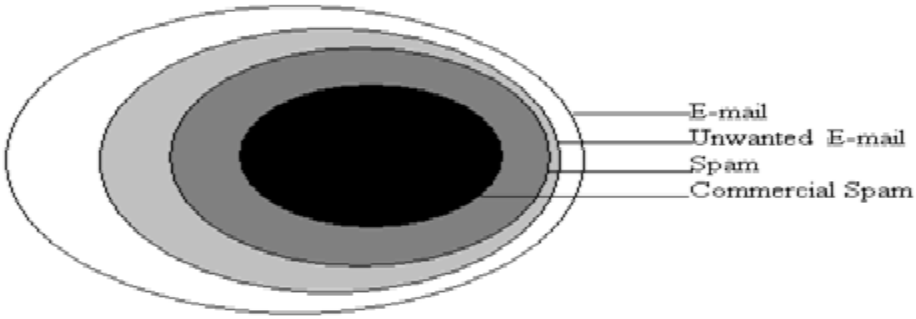

Figure 1.1: Set of Spam E-mails and Other E-mails [11]

1.3 History of E-mail Spam

Spamming began in 1978. At that time spam was sent manually so it was not able to reach millions of user [11]. This era of spam lasted till mid of 1994. These were the tolerable years. In 1994 "Green card lottery" a spam posted by Canter and Siegel, two lawyers from phoenix. Same year programmers were hired to write a spammer program. In 1995, spamming became business, an e-mail list was offered for sale with more than 2 million addresses. Nevertheless, spam turned to be totally out of control by 1997. Now, amount of spam received has become very large and uncontrollable. Out of all spam, e-mail spam is growing repeatedly; growth of e-mail spam in recent years is shown in Figure 1.2.

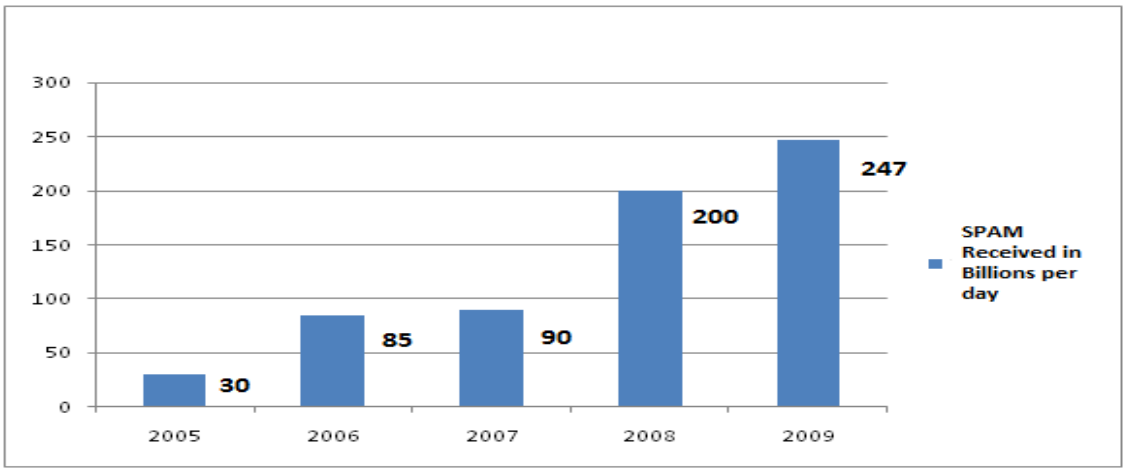

Figure 1.2: Growth of E-mail Spam [15, 16]

1.4 Areas of Spam

It is well understood by all, that in the competitive world of business nothing can survive if it is not efficient and doesn't give enough profit. Therefore, spamming is definitely a profitable business and apparently quite lucrative. Spam is like TV advertisements. Commercial Ads also do not have many fans and in fact most of the people try to avoid them. But repeating same ad so many times, one must notice. Years after years, the product is the same and only the commercials are slightly changed. Finally, doing shopping every day we can realize that among the range of products in fact we will prefer one that "we know" meaning the product, which is more often and more actively advertised on TV. These advertisements belong to the various products of different use. Spamming is little different from TV advertising, the goal of spam is to make viewers to consider its offer immediately so that we either order the promoted product/service or at least visit the promoted web site.

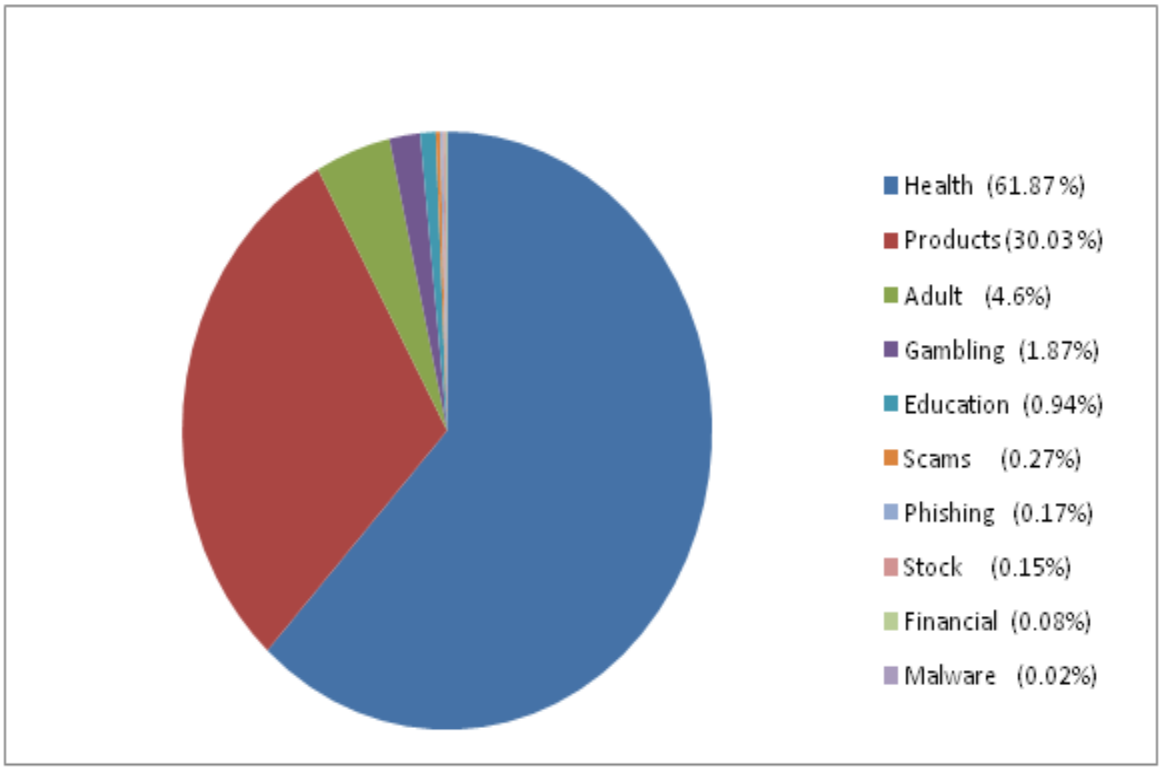

Figure 1.3: Spam in Various Areas [26]

Spam categorized in various areas, as shown in Figure 1.3 and some of these are briefly discussed in this chapter.

1.4.1 Health

Mostly, spam is sent regarding weight problems, skin problems, heart problems, and sex problems. These e-mails are written in such a manner that attracts the reader of e-mail. One can see those e-mail spam ads that tout the latest herbal remedy that promises dramatic weight loss. Health spam is a major part of the spam flowing on the internet as it can be seen in figure 1.3 that it contains 61.8% of all spam e-mails.

1.4.2 Products

Mostly, spam is product advertisements. One can see Online Pharmacy spam: Spam promoting different versions of Viagra, Calais, and spam regarding anti-depressant pills that can be purchased online. Stock-encouraging spam, encouraging people to buy cheap stocks also can be seen. There are some spam mails offering pirate software, usually much cheaper than the official prices.

1.4.3 Adult

One of the earliest types of spam came bearing adult and pornographic content. To the current day, it remains one of the most common forms a user will find in their inbox. Content may range from explicit dating services to enhancement products. This type of spam is often difficult to prevent, as clicking on a single link within it will typically bring more from different senders.

1.4.4 Gambling

So many mails are flowing on Internet containing links to gambling sites. These mails are used to promote gambling sites by offering some schemes to the receiver of mail. Various e-mails have been received in past regarding lotteries. At one time, Australian lottery was famous e-mail spam, other similar kind of lottery spam also exists. One of the lotteries spam is shown in Figure 1.4.

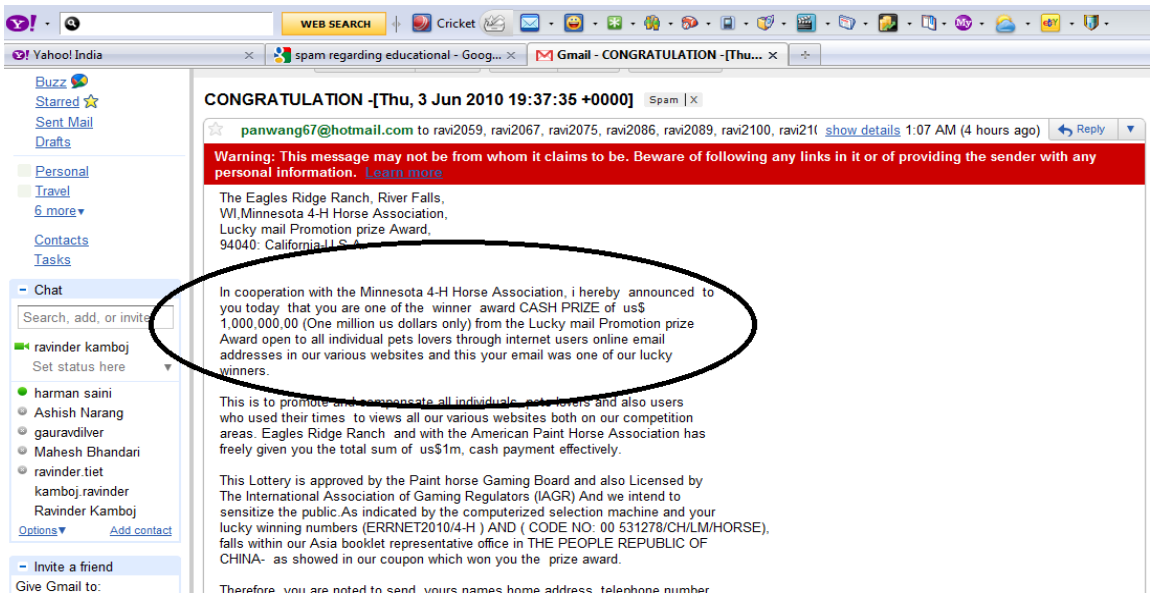

Figure 1.4: Spam Regarding Lottery

1.4.5 Education

This category includes offers of seminars, training, and online degrees. Spam regarding offers of scholarship also can be seen. Spam regarding higher education is very famous; it contains offers of degree along with scholarships and funds for education. Advertisements regarding online courses are also coming in the mail boxes of the users.

1.4.6 Phishing

In the field of computer security, phishing is the criminally fraudulent process of attempting to acquire sensitive information such as usernames, passwords and credit card details by masquerading as a trustworthy entity in an electronic communication. Communications purporting to be from popular social web sites, auction sites, online payment processors or IT administrators are commonly used to lure the unsuspecting public. Phishing is typically carried out by e-mail or instant messaging, and it often directs users to enter details at a fake website whose look and feel are almost identical to the legitimate one. Even when using server authentication, it may require tremendous skill to detect that the website is fake. Phishing is an example of social engineering techniques used to fool users, and exploits the poor usability of current web security technologies [5].

1.5 Operating Techniques of Spammers

Many spam mailers use tricks to get users to read their messages. They use the subject line in such a way that it seems interested to the receiver. The worst thing about spam is that the spammers use tricks that help disguise the origin of their messages. One of the spammer's most common tricks is to relay messages through the e-mail server of an innocent third party. This tactic doubles the damages: both the receiving system and the innocent relay system are flooded with spam. And for any mail that gets through, often the flood of complaints goes back to the innocent site because it was made to look like the origin of the spam. Many spammers send their spam from a free account from a large ISP such as Yahoo!, or Hotmail, then abandon the account and open a new one to use for the next assault. Spammers operate by making use of following ways:

- Spam via Botnets
- Localization of Spam
- Image Spam

1.5.1 Spam via Botnet

A botnet is a collection of autonomous computers, which can be controlled remotely via a specific application. Botnets get their origin from criminals who are very tech-savvy and well-versed in computer programming and software creation. The criminals that perpetrate botnets are known as "bot herders" because they control the computers that have been compromised from a remote location. Once the computers are compromised they can communicate over the Internet, which means a botnet can be a group of "zombie" computers that is formed anywhere in the world. The botnet master communicates with individual bots through a commonly used protocol, such as Internet Relay Chat (IRC), HTTP, and P2P. Bots receive commands from their master and carry out attacks as instructed without the knowledge or consent of the machine's owner [2, 14, 19].

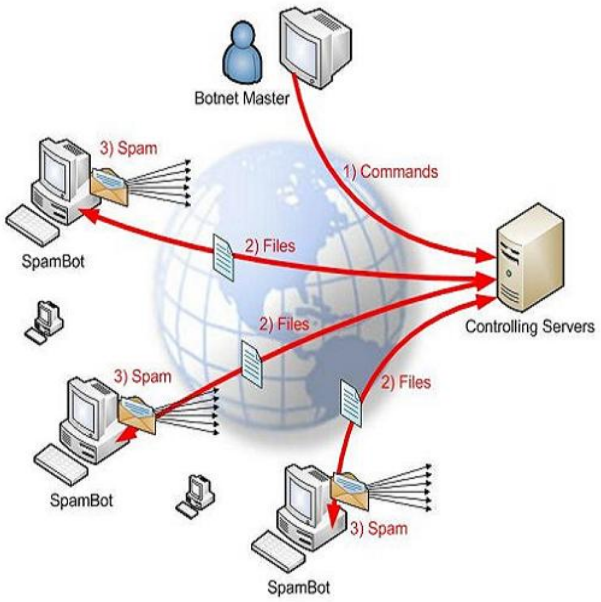

Figure 1.5: Spam via Botnet [19]

1.5.2 Localization of Spam

Earlier spam were sent usually in English, non-English speaking users filtered out those e-mails, either manually or applying content filters those score English e-mail as a spam. Now Spammers make use of some mechanism to convert spam to local languages like: - Russian, French, Spanish, German etc. it has following two advantages:

1) Recipient understand the message
2) Most Spam filters miss the spam

1.5.3 Image Spam

Since majority of text spam is blocked by content filters, then spammers started a new way for spamming called image spam. Image spam is a kind of spam where text of the message is represented in GIF, JPEG or other kind of image. Often, image spam contains nonsensical, computer-generated text which simply annoys the reader. However, new technology in some programs tries to read the images by attempting to find text in these images. They are not very accurate, and sometimes filter out innocent images of products like a box that has words on it. It is difficult to detect by spam filters those are designed to detect text spam.

Figure 1.6: Image Spam [14]

1.6 Problems by Spam

More recently, spam has been spreading at an increasingly rapid rate. Today more than 85% of Email traffic following on the Internet is spam Emails. Following subsections describes a number of reasons which shows why spam has become a serious problem.

1.6.1 Problem Related to Cost

Spam imposes costs on all Internet users. There are three types of cost: capital, staffing and business. The users loose time and various ISP's lose money, trust, working hours or even operation of their servers due to spam . Spam cost USD 20.5 billion in 2003, USD 198 billion in 2007 and it is estimated that spam might cost the much more billions of dollars in the near future [1]. The other reason why spam is costing money is because of the loss of valid mail, by losing it in the flood of spam. This can result in the loss of business, as Email is an important form of communication for various businesses.

1.6.2 Problem Related to Privacy

The Collection of Email addresses are made by spammers usually without the knowledge of users of Email; it is a kind of major loss of privacy [1].

1.6.3 Problems Related to Spam Content

The content of spam may create a problem due to fraud and deception. Spammers disguise the origin of Email because they know their message is being blocked or filtered and they aim to entice individuals to open their Emails. Commonly spammers forge the headers of messages. Some spam messages contain pornographic content and promote adult entertainment products and services [1].

1.7 Overview of Thesis

The rest of the thesis is organized in the following order:

Chapter 2 - Provides literature review. In this chapter various anti-spam techniques have been discussed. These techniques have been classified into two categories some those can be applied on e-mail envelope and others those can be applied on e-mail data. Legalization of spam and user tips to avoid spam has been discussed. Two approaches which can be followed like learning approaches and rule based approaches has been discussed. The way in which SpamAssassin implement these two approaches is also has been discussed.

Chapter 3 - Gives the problem statement which comes out after the literature survey. Problem statement focuses on the requirement of design and implementation of custom rules.

Chapter 4 - In this chapter design and implementation are being discussed, this chapter includes the custom rules to detect the problem of spam with focus of decreasing false positives and false negatives. Rules have been categorized in to body rules, header rules, Meta rules and URI rules.

Chapter 2

Literature Review

2.1 Introduction

Nowadays the implementation of a reliable spam filter has become more and more important. Since e-mail users have to face growing amount of uninvited e-mails. The spam is mainly a cheap and illegal form of advertisement exploiting the thousands of users are easily reachable on the Internet. Although it is illegal, most legislation enforcements fail due to the inability to identify the spammer. The deficit due to the user's time-loss can be measured in thousands of dollars.

2.2 Legal Action against Spam

CAUCE, or the Coalition Against Unsolicited Commercial E-mail, is a non-profit advocacy group that works to reduce the amount of unsolicited commercial e-mail, or spam, via legislation. CAUCE was founded in 1997. Various countries have joined this group, India has joined this group and CAUCE India established in 1999.

In U.S.A, an act was established in 2003 known as CAN-SPAM act. [4] According to this act, following things are not allowed while sending e-mail:

- Use of false or misleading "from" address.
- A subject line that masks the purpose of e-mail.
- Harvesting of e-mail addresses or use of dictionary attacks.
- Use of open relays to send e-mail.

2.3 User Guidelines for Avoiding Spam

Mostly e-mail addresses are harvested by spammers for Internet; spammers create a list of harvested addresses and send spam mails to this list. Following are some guidelines for e-mail user so that spam can be avoided to some extent:

- Don’t share your e-mail address on newsgroups, chat rooms and any other website.
- Check the privacy policy when you submit your e-mail address to a website.
- If you receive a suspicious e-mail do not follow the link given in its content.
- Do not share your e-mail addresses with many people.
- One should use two separate e-mail address, one for personal e-mails and one for newsgroups and Chat rooms.
- Use a unique e-mail address; it should not like simple dictionary words because mostly spammers use dictionary attack while sending spam.
- E-mail filters can be used to filter spam one can set filters in mail program like Gmail.

2.4 Anti-Spam Methods and Techniques

On the basis of transport protocols anti-spam methods can be categorized into two categories:

1. E-mail Envelop Analysis.
2. E-mail Data Analysis.

These two methods use various techniques to detect spam. These techniques are discussed here:

2.4.1 E-mail Envelop Analysis

Once an SMTP connection has been established, following information arrive before the data part:

- IP address is given by TCP/IP dialog.
- The sender’s domain which is easy to forge.
- The sender’s E-mail addresses which can be easily forged.
- Arbitrary recipient addresses**.**

Blocking spam at this part of the SMTP dialog is most efficient; possibility of connection can be cancelled before receiving the mail if sender of e-mail is dubious. This method makes use of following techniques:

- Blacklisting
- Grey Listing
- White Listing

2.4.1.1 Blacklisting

Anti-spam blacklisting describes the process of blocking upcoming SMTP connections from spammers, which are contained within a list of IP addresses (blacklist). Identify sender addresses where spam is known to originate by placing such addresses or domains on a blacklist. Blacklisting needs very few resources and protects against resource misuse, since e-mail delivery is denied beforehand. Blacklisting is independent from e-mail's content, i.e. no liability to weakness of content filtering. Most commonly known blacklists are DNSBLs (Domain Name System Blacklists). Two different kinds of DNS blacklists are used:

- IP-based Blacklists
- Domain-based Blacklists

a) IP-based Blacklists

Majority of DNSBLs are IP-based, which look at the IP (Internet Protocol) address of server sending the mail. When an attempted e-mail is delivered to mail server then mail server or anti-spam software running on that mail-server examines the IP address from the header of e-mail and checks either it is blacklisted or not. If IP address is blacklisted then e-mail is tagged as spam otherwise it is directly sent to the recipient.

b) Domain-based Blacklists

This kind of blacklists are rarely used but they are also very important in some cases like if multiple domains are hosted on same address then only IP address based blacklists are

not of use. Domain-based DNSBLs are also called right-hand side blacklists. These lists look only at right -hand side of @ sign. For example a sender sending e-mail from manjeet4u@example.com, then these kind of lists will look at only example.com and checks either domain name is blacklisted or not. If domain name is blacklisted then e-mail from this domain will be tagged as spam, otherwise it will be directly sent to the recipient of e-mail.

Blocking SMTP connections without looking into the e-mails might be dangerous, because no quarantining and in this way no recovery of false positives is feasible.

2.4.1.2 Grey Listing

Grey listing is defined on the assumption that a legal e-mail sender does more effort to send his e-mail than a spammer. Therefore the server stores cookies for each connection attempt within a defined time span in the past. When sender sends e-mail its attempt is stored in cookie if it is in blacklist. This allows the sender to submit e-mails in a second try after a specific embargo time. Simply put, grey listing waits for a second attempt after a particular blocking time span before accepting an unknown sender's connection. There are some improvements of this technique to increase the sender's efficiency. High benefit can be achieved with very small effort, because many spammers often do not try a second time. Spammers might adopt their methods if grey listing becomes more regular [11, 2].

2.4.1.3 White Listing

White lists are opposite to blacklists, they contain users that are verified contacts. All the e-mails from these senders will be treated as ham (legitimate mail). A majority of ham e-mails is sent by well known sources, which do not need to be checked against grey- and blacklists. However, one could consider a solution allowing everything from this a sender is white, listed. White listing disburdens the resources which are needed for requesting blacklists and allows abstaining from interference of legitimate connections. Disadvantage of white lists are once a white listed server begins sending spam, the trust to this server must be proofed if misuses occur.

2.4.1.4 Sender Authentication

Earlier it was possible that anyone can send e-mail to anyone. But this causes so many problems in the form of spam. So it became necessary to identify the sender whether he/she is authorized to send e-mail [2]. Two different kinds of authentication proposals can be found. These kinds are given here:

- Path Based Sender Authentication
- Signature Based Authentication

a) Path-based Sender Authentication

Authentication methods like Sender ID and SPF (Sender Policy Framework) can be used to test whether an e-mail server is authorized to send on behalf of a given domain. The fundamental for this is publication of DNS records that lists all authorized e-mail servers for a domain. Now on the receiver's site can be checked, if the domain of the given e-mail address may be used by the sending server. The Figure 2.1 explains two different situations. First considering the good case, i.e. an authenticated sender owns the domain "example.com" and sends an e-mail from this domain. The recipient checks via a DNS resolver whether the sending IP is authenticated to send mails on behalf of "example.com". The DNS server, which is responsible for this domain returns special DNS records that express which IP addresses are authenticated to send mails from this domain.

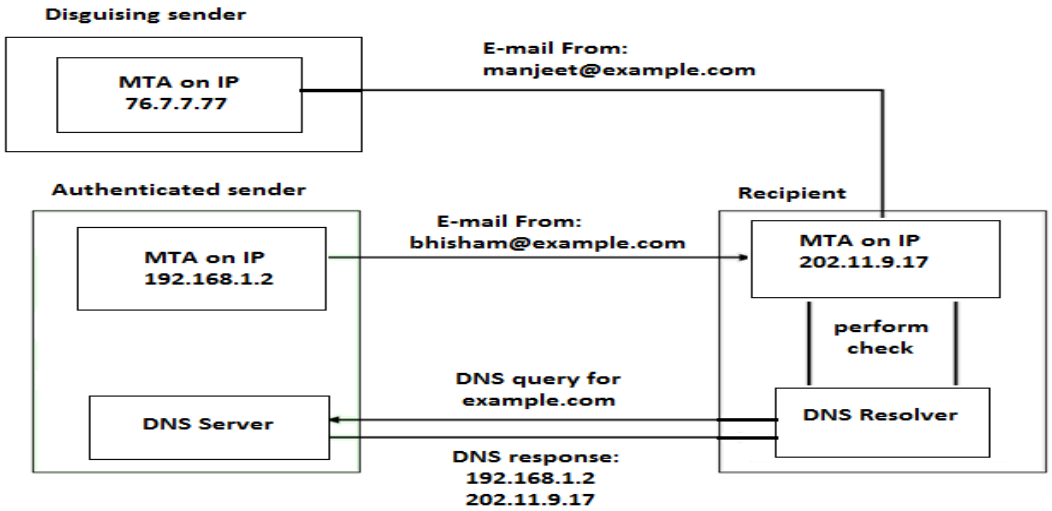

Figure 2.1: Path Based Sender Authentication [2]

In the first case the sender uses one of the granted IP addresses (192.168.1.2) and is well authenticated. On the other hand, senders mailing from different IP addresses (76.7.7.77) is not authenticated. It is easily manageable for senders using existing technology, since only a DNS record has to be published. Receivers have to adopt new software in order to check the DNS records. Disadvantage of this technique is that if two or more domains are running on the same IP address, an e-mail sender from this IP can use every of those domains to send authenticated e-mails. It is Prone to domain tasting, because spammers could use domains with legitimate DNS records [2].

b) Signature Based Authentication

Signature based authentication used to authenticating a sender by using the technology of digital signature via asymmetric encryption. Similar to the path based sender authentication signature based authentication assures the use of correct domain only. An authenticated sender signs an e-mail with a digital signature using private key provided to user by his/her domain "example.com". On receiving MTA when an e-mail is received, MTA retrieves the public key of domain to check whether the signature of sender is correct or not. Using this technique differentiation between several domains hosted with an IP address is possible.

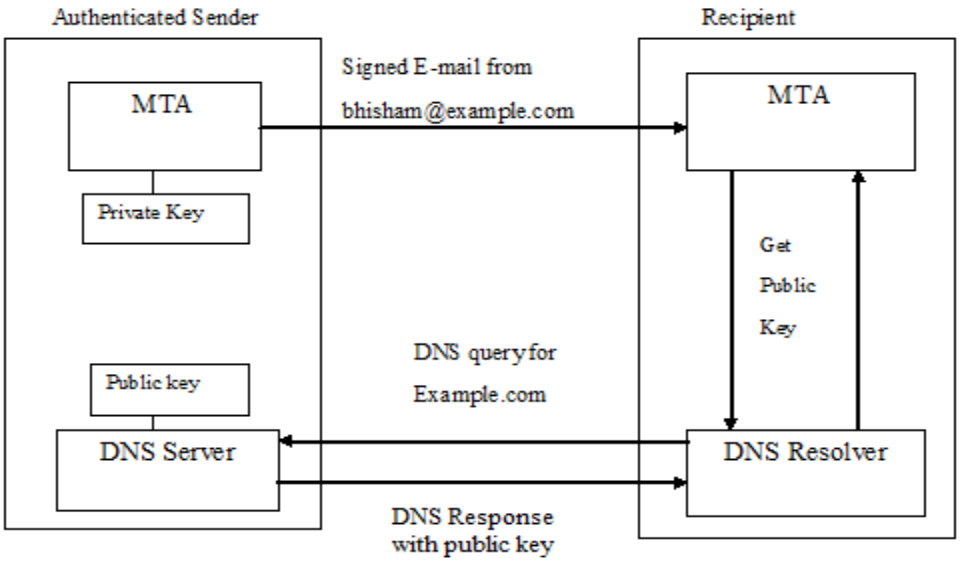

Figure 2.2: Signature Based Authentication [2]

The Figure 2.2 shows how signature based authentication works. Using this modification of e-mails is no longer possible, i.e. some software implementation must prevent this (e.g. mailing lists that put unsubscribe information to the end of every e-mail). But disadvantage of this technology is that senders as well as receivers have to implement new technologies to sign e-mails and/or to check these signatures. It is also prone to domain tasting, because spammers could use domains with legitimate DNS records [2].

2.4.1.5 Sender Address Verification (SAV)

SAV is a mechanism used to check whether an e-mail address exists or not. Spammers often use mythical e-mail addresses that usually do not exist. SAV helps only to verify if the sender's given e-mail address exists. On the other hand, it does not help to verify if the sender is authorized to use this specific e-mail address or domain. Since SAV helps only to block spam due to wrong e-mail addresses and does not increase the reliability when receiving "verified" e-mails. Technically, the receiving MTA performs SAV with the given sender address during the SMTP dialog with the sending MTA. In order to do so the receiving MTA establishes an SMTP dialog to the MTA accepting e-mails for the domain stated in the sender address and tries to deliver a bounce message to this address. If the bounce message was accepted, the sender's address (probably) exists and the receiving MTA should accept this e-mail address. If the bounce message was rejected, the e-mail address is likely to be invalid [2].

2.4.2 E-mail Data Analysis

Once an SMTP connection is accepted and the e-mail's DATA is delivered, this data can be analyzed for spam-like patterns. Usually, the following methods are applied after the SMTP dialog. Some implementations act during the data delivery.

2.4.2.1 Heuristic Techniques

Often also known as rule-based content filtering, heuristic techniques usually aim at finding specific words, regular expressions or misuse related styles in e-mails to classify them as spam or ham. Once suspicious e-mails have been found, e.g. an outstanding

expression of a spam e-mail, it will be added as a new policy. A rulebook contains these checks and must be managed manually. A policy can either work on the e-mail body or on the e-mail header.

a) Header Analysis to Detect Spam

Kobkiat Saraubon and Benchaphon Limthanmaphon from Department of Computer and Information Science (King Mongkut's University of Technology North Bangkok) have given a way extract the e-mail header to find botnet spam. Spam is classified by comparing the locations of these elements: the sender's IP address, the sender's MX hosts, and the sender's e-mail address. If the location of the sender server and the location of the MX hosts are different, then the e-mail will be identified as a spam [14]. This technique works similar for text spam and image which is difficult to detect with content based spam filtering methods, technique is given below in Figure 2.3:

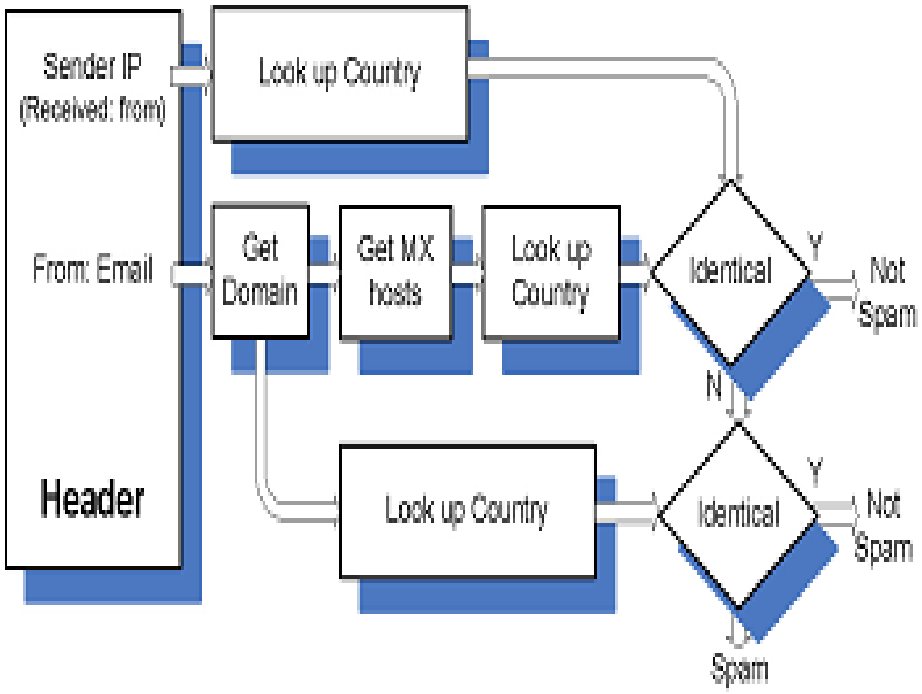

Figure 2.3: Botnet Spam Detection by Header Analysis [14]

Figure 2.3 given above shows the technique and explanation steps of this technique are given next.

1. Extract Mail Header
   - Sender IP
   - Sender's e-mail Address
2. Lookup Country (Sender IP)
3. Analyze Sender's e-mail Address
   - Obtain Sender Domain
   - Obtain Sender MX Hosts
4. Lookup Country (Sender Domain)
5. Lookup Country (MX Host)
6. If the result from step 2 = the result from step 4 Then
   not a spam
   Else if the result from step2 = the result from step 5 Then
   not a spam
   Else identify as a spam

MX records can be checked using IP-Location database contributor such as ARIN, APNIC, LACNIC, RIPENCC, AFRINIC and IANA.

b) Heuristic E-mail Content Analysis

Because spammers tend to disguise the real word into multiple patterns, usually regular expressions are used to detect them. The following example of the most common word contained in spam e-mail "Free" shows some possibilities to modify the word:

Free => Fr_ee
F_r_e_e
Freeeeee
Fr<>ee

All forms make it more difficult to read the word, but nevertheless almost every human should be able to read it. It is much work finding a proper regular expression finding most of all possible cases. In addition to this the list of bad words and/or regular expressions might get very long. This causes several checks for each mail and leads to a leak of performance. It is not recommended to quit the check after the first hit and

classify the e-mail as spam, because ham e-mails could contain some words or wrong headers, too. Hence first a combination of some positive checks should lead to a final classification. The method does not need a training phase. Heuristic analyses are very efficient with a well-managed policy database. This method has some disadvantage like a leak of performance might occur with at high mail volume or huge policy databases and managing the policy list is very time-consuming [2].

2.4.2.2 Statistical Techniques

A statistical filter automatically splits e-mails into several tokens (e.g. words) and looks these tokens up in a database. The database contains common tokens with a classification whether or not it is a common token in spam e-mails. This requires a training phase of statistical techniques, where lots of messages must be classified as spam/ham in order to build up the database. Most of the statistical techniques based on the Bayesian Filtering. Bayesian filtering will be discussed in detail later on.

2.5 Exploring Bayesian Filtering

Bayesian filters work on the basis of that particular words have particular probabilities of occurring in spam e-mail and in legitimate e-mail. For instance, most e-mail users will frequently encounter the word "Free" in spam e-mail, but will seldom see it in legitimate e-mail. The filter doesn't know these probabilities in advance, and must first be trained so it can build them up. To train the filter, the user must manually indicate whether a new e-mail is spam or not. For the words in each training e-mail, the filter will adjust the probabilities of each word according as it appears in spam or legitimate e-mail in its database. For instance, Bayesian spam filters will typically have learned a very high spam probability for the words "Free" and "Lottery", but a very low spam probability for words seen only in legitimate e-mail, such as from friends and family members. Each word in the e-mail contributes to the e-mail's spam probability, or only the most interesting words. This contribution is called the posterior probability and is computed using Bayes' theorem. Then, the e-mail's spam probability is computed over all words in the e-mail, and if the total exceeds a certain threshold, the filter will mark the e-mail as a spam. A

database is being kept out by Bayesian filters which contain probabilities of words according to their previous appearances in spam e-mails and legitimate e-mails.

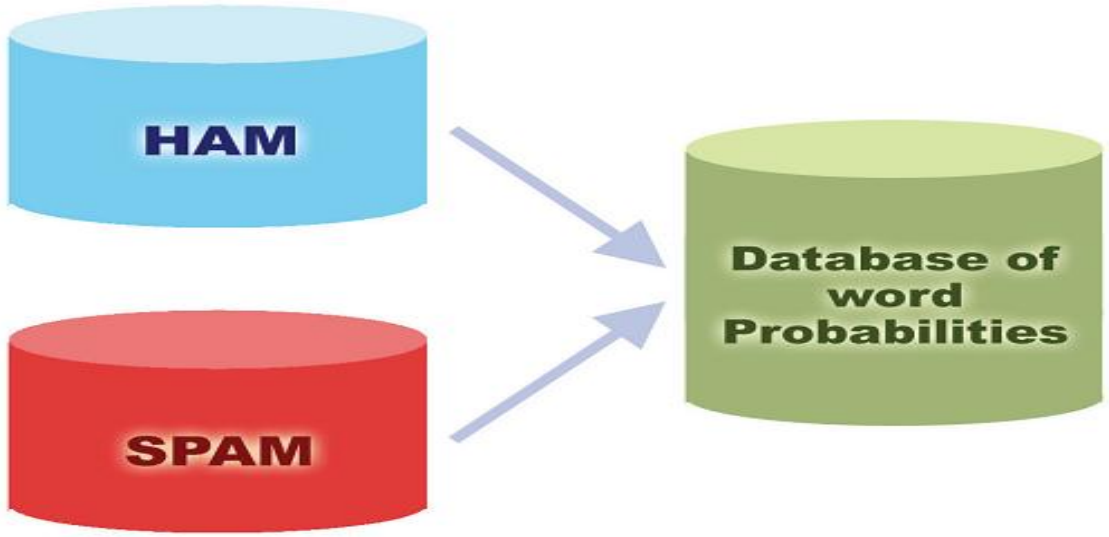

Figure 2.4: Database for Bayesian Filter

2.5.1 Steps of Bayesian filtering

While Bayesian filters categorize the incoming e-mail in to specific category this whole procedure goes through the following steps as shown in figure 2.5.

2.5.1.1 Loading

An e-mail is loaded. Most of the filers are combined with e-mail system and load the e-mail in real time.

2.5.1.2 Pre-filtering

Pre-filtering is required for every text based search system. Without pre-filtering, the tokens kept in the token-dictionary could reach an unlimited size, since the spammers realized in a very short time, that writing the words in their dictionary-form leads to an easy filtering, and so they began to change some letters or put extra characters in (e.g.: L_u_c_k_y). This has an effect not only on the keyword searching methods but also makes statistical filtering more difficult. The pre-filters role is to try changing the

words back to their original form and to make the search engines able to find them. An interesting newer method does not try to change the words back, rather tries to find similar words.

2.5.1.3 Tokenization

The Bayesian filter works with the individual small parts of the text, the so-called tokens. This very simple hierarchic model fits to the naive Bayesian model and it does not count with the dependency between tokens. Many filters work with simple word-by-word tokenizing, where the text is separated into words and the words have their own values in the token-dictionary. In this case, one word will be one token. The dictionary should be updated with all the new tokens from the processed e-mail. After the tokenization, the tokens are handled separately, what means that the order of the tokens is taken into consideration.

Table 2.1: Values for Tokens

N1	No. of spam letters, where the word has existed
N2	No. of legitimate letters, where the word has existed

When a new e-mail is received, the token dictionary is searched for all of the words, included in the e-mail. Two sets of words are handled: one is the set of words matching the dictionary (an update will be necessary at value N1 or N2), other is the set of words not included in the letter (no update will be necessary).

2.5.1.4 Calculation

After the pre-filtering and tokenization the values of the tokens are looked up and a decision matrix is built from the most relevant token's values. Usually, the filters use only a fixed number of values, the ones, which are farthest from the neutral value. In an environment supported by unlimited hardware resources surely all the tokens could take part into the calculation. With the limited number of tokens, only the significant tokens are taken into consideration. If a token does not exist in the dictionary, it is added as a new token to the token-dictionary with a neutral value. Using the

values of the tokens and other statistical data (like number of all letters, number of spam letters, etc.) makes available to calculate the final probability of being a spam. The result is a number, but it could be easily turned into a binary value as well. Bayesian filter performs following calculations to categorize an incoming e-mail as spam or legitimate e-mail.

ALL = SPAM + HAM, No. of all E-mails.

- Let us call a word “matching word”, if the word has existed both in the letter and in the token dictionary.

- P (“matching words” | “letter is spam”) = for all matched word (N1 value of the current word / SPAM)

- P (“matching words” | “letter is legitimate”) = for all matched word (N2 value of the current word / HAM)

- P (“letter is spam”) = SPAM / ALL

- P (“letter is legitimate”) = HAM / ALL

- P (“letter is spam” | “matching words”) =

= P (“letter is spam”) * P (“matching words” | “letter is spam”)

- P (“letter is legitimate” | “matching words”) =

= P (“letter is legitimate”) * P (“matching words” | “letter is legitimate”)

- Final result: P (“letter is spam” | “matching words”) /

P (“letter is legitimate” | “matching words

2.5.1.5 Feedback

After getting the final result, whether the e-mail was a spam or not, the stored values in the token-dictionary have to be changed. Note that a change is necessary only for the

tokens that existed in the letter. This gives the continuous learning capability of the method, since tokens, which are often in spam letters, will get higher and higher values, while tokens related to legitimate mails are getting smaller values.

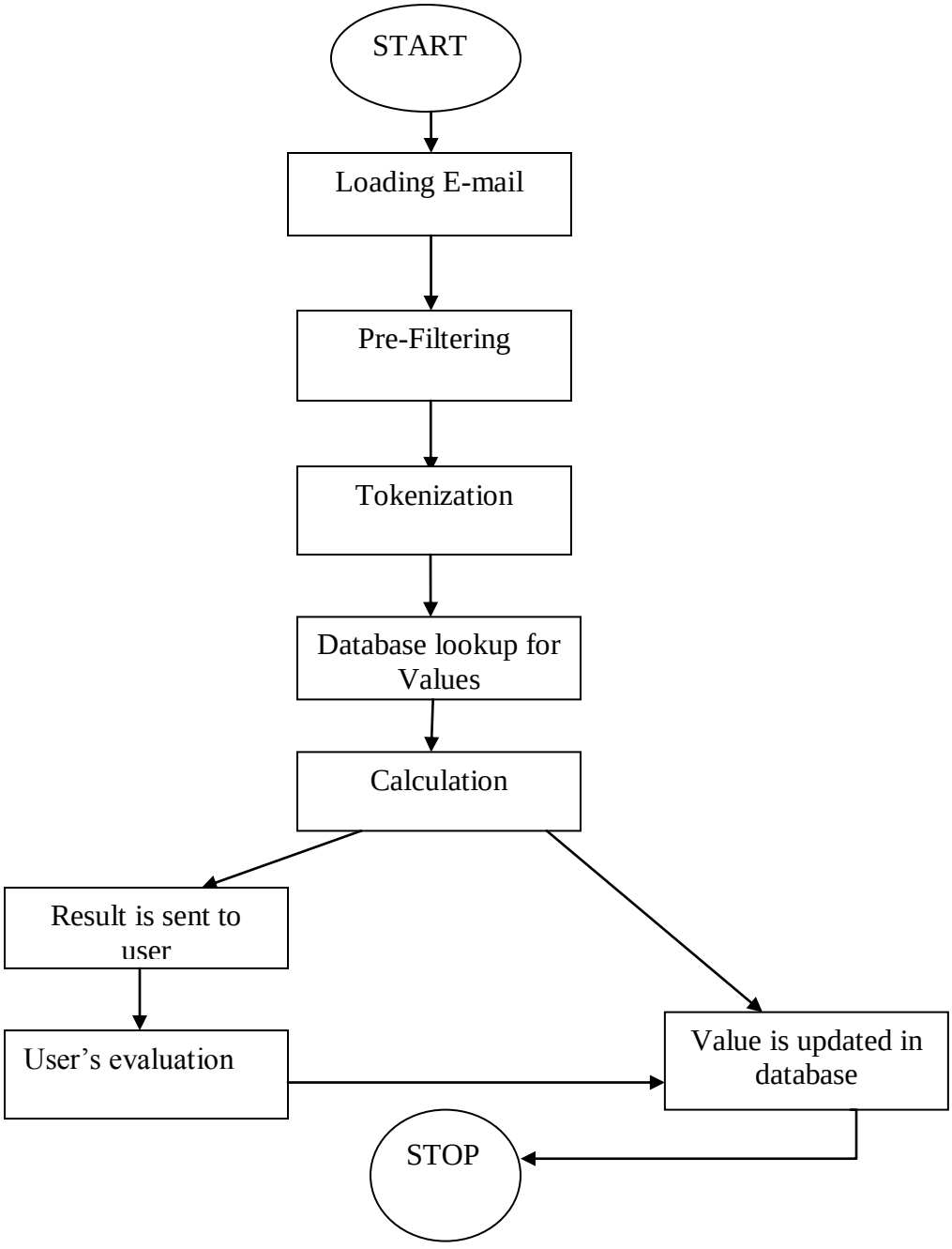

Figure 2.5: Flow Chart of Bayesian Filtering

2.6 Approaches for Spam Detection

Techniques discussed above can be implemented by using following two approaches:

- Learning Approach
- Rule based Approach

2.6.1 Learning Approach

This approach deals with training of Spam filter. A large set of ham e-mails and spam e-mails is used to train the spam filter. In training filter reads tokens from e-mails and adjust the values of tokens/words in the database according to their category whether they are from spam e-mail or ham e-mail. If e-mail has been read as spam than N1 values of the corresponding words will increase in the database, and if e-mail has been read as ham e-mail then N2 value will be increase in database. After training with a large set of spam and ham e-mails filter will be ready to deal with new incoming e-mails [20, 21].

Didier Colin, Catherine Roucairol and **Ider Tseveendorj** from, Prism Laboratory, Versailles University France had proposed a selective learning approach for the detection of spam to improve the learning efficiency. According to them training data may contain more than useless information. Indeed, some may hold destructive knowledge, *i.e.* knowledge that will decrease the performances. In order to reach maximum accuracy and generalization capabilities, classifiers must extract only pertinent information from the training data [24].

2.6.2 Rule -based Approach

Rule-based approach is used by creating rules to categorize the incoming e-mails. It is known as direct approach. It does not require any training phase. Rules cover different threats, suspicious format, and weak origin prone to sending spam means sender is confirmed as open relay [20]. While using this approach we have to be careful because rules generated by us can lead the incoming e-mails in to misclassification [21]. There is risk of false negative and false positive. Following steps are performed while we are using rule base approach for spam detection:

- Pattern Analysis

- Pattern Selection
- Score Assignment

2.6.2.1 Pattern Analysis

Pattern of spam and ham e-mails are analyzed from the set of database of spam e-mails and ham e-mails. For pattern analysis header and body of e-mails are searched for dubious keywords which are known to come in spam or ham e-mails. Analysis of spam e-mails is also depends on the type of e-mails whether it is Adult, Product, Educational or Gambling Spam [21].

2.6.2.2 Pattern Selection

Pattern is selected with the combination of several features. Selection of words/tokens with higher probabilities is done during this step. For the selection of spam like pattern tokens with higher **W/S** (probability that the word is in spam e-mails) value are selected. Along with probability of words header of the e-mail is also considered [21]. Overall pattern selection includes selection of words, Subject and From: field of the e-mail.

2.6.2.3 Score Assignment

Score can be assigned using two ways, one it can be assigned by calculating by score learning tool and other rule maker can assign score of own choice. When we are assigning score with our own choice then we must be more careful. Wrong assignment of scores can lead to misclassification. Total score of the e-mail is calculated as follows:

$$f(x) = \sum_{i=1}^{N} w_i x_i$$

Where W_i is the score for rule i and X_i is whether or not rule i is activated by given e-mail or not [21].

2.7 SpamAssassin

SpamAssassin was developed under Apache license and now it is part of Apache foundation. SpamAssassin is a tool designed to detect spam mails. SpamAssassin is used for e-mail filtering using content matching rules. **SpamAssassin was awarded with Linux New Media Award 2006 as the best Linux-based Anti-spam Solution**. SpamAssassin uses a variety of spam-detection techniques that includes DNS-based and checksum-based spam detection, Bayesian filtering, external programs, blacklists and online databases like Razor and Pazor [18]. It works on the basis of score. Whenever a particular word is found in e-mail it adds score to the e-mail header. A threshold value is fixed to filter spam mails, when this threshold value by an incoming e-mail then it will be categorized as spam. In windows by default threshold value for spam filtering in SpamAssassin is 6.0, but in linux it is 5.0. It makes use of heuristic techniques as well as Bayesian filtering to analyze e-mail for special phrases or words those occur in spam mails received by user. SpamAssassin make use of above given both kinds of approaches.

2.7.1 Learning Approach of SpamAssassin

Learning is required for ever Bayesian based spam filters, so for this to train SpamAssassin **sa-learn** command is used. 'sa-learn' used to tech filter to which kind of mails should be classified as spam and which kind of mails should be classified as ham (legitimate mails). **sa-learn** can be used with following options.

--ham

Learn the input message(s) as ham. If message has been previously learnt as spam, SpamAssassin will forget them first, and then re-learn them as ham. Alternatively, if message has been previously learnt them as ham, it will skip them this time around. If the messages have already been filtered through SpamAssassin, the learner will ignore any modifications SpamAssassin may have made.

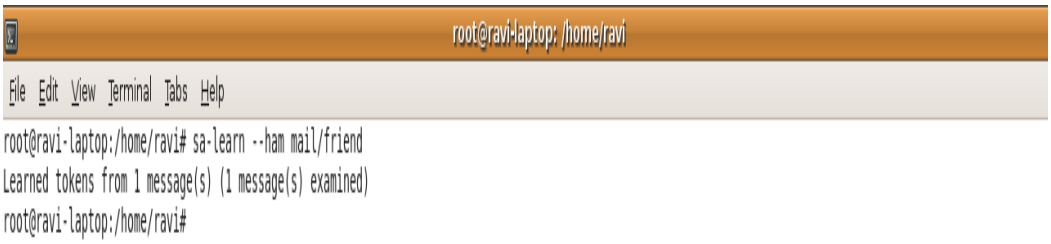

Figure 2.6: E-mail Learned as Ham

--spam

Learn the input message(s) as spam. If message has been previously learnt any as ham, SpamAssassin will forget them first, and then re-learn them as spam. Alternatively, if message has been previously learnt them as spam, it will skip them this time around. If the messages have already been filtered through SpamAssassin, the learner will ignore any modifications SpamAssassin may have made.

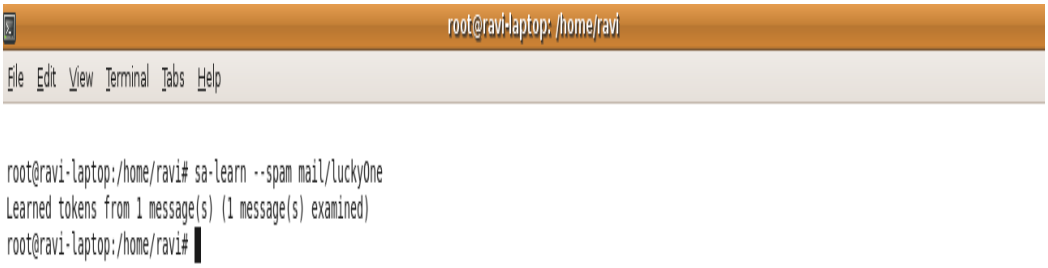

Figure 2.7: E-mail Learned as Spam

--folders=*filename*, -f *filename*

sa-learn will read in the list of folders from the specified file, one folder per line in the file. If the folder is prefixed with **ham**:type: or **spam**:type:, sa-learn will learn that folder appropriately, otherwise the folders will be assumed to be of the type specified by --**ham** or --**spam**.

--mbox

sa-learn will read in the file(s) containing the e-mails to be learned, and will process them in mbox format (one or more e-mails per file).

--use-ignores

Don't learn the message if a From: address matches configuration file item bayes_ignore_from or a To: address matches bayes_ignore_to. The option might be used when learning from a large file of messages from which the hammy spam messages or spammy ham messages have not been removed.

--forget

Forget a given message previously learnt.

--dump *option*

Display the contents of the Bayes database. Without an option or with the *all* option, all magic tokens and data tokens will be displayed. *Magic* will only display magic tokens, and *data* will only display the data tokens.

--clear

Clear an existing Bayes database by removing all traces of the database. But users have to be carefull this is destructive.

2.7.2 Rule -based Approach of SpamAssassin

Rule-based approach is used by creating rules to categorize the incoming e-mails. Rules can be written in **user_prefs** or **local.cf** file of SpamAssassin. Following are types of rules which can be implemented under SpamAssassin:

- Header Rules
- Body Rules
- Meta Rules
- URI Rules

2.7.2.1 Header Rules

Header rules allow checking of a message/mail header for a string matching. Most commonly these rules check the Subject:, From: or To: field of the header of the message. If a particular string match is found then corresponding score is assigned to

message/mail and it will make effect on the **X-Spam-Status** of the mail. Following is the syntax for header rule:

```
header NAME_OF_THE_RULE  /\bword\b/i
score NAME_OF_THE_RULE  any values from 0.0-1.0
describe NAME_OF_THE_RULE description of the rule
```

Here (b) is used for to check anything placed in between the world that is not an alpha numeric character.
Here (i) is used to make the scanning case insensitive. This will check word/WORD and any letter in between the word is other case.

2.7.2.2 Body Rules

These rules search the body of the message with a regular expression and if it matches, the corresponding score is assigned. A sample body rule is written ahead.

```
body LOCAL_RULE /test/
score LOCAL_RULE 0.1
describe LOCAL_RULE this is a sample rule
```

The rule given above will search body of e-mail for the word **test**.

2.7.2.3 Meta Rules

Meta rules are rules that are boolean or arithmetic combinations of other rules. This allows us to create a single rule that fires on header and the boy part of the body. Following is a sample Meta rule:

```
header _LOCAL_RULE_SUBJECT Subject =~ /congratulation/i
header _LOCAL_RULE_FROM       From =~ /donotreply\@mail\.biz/i
body _LOCAL_SCHOLARSHIP /scholarship/i
body _LOCAL_LUCKY /lucky/i
uri      _LOCAL_URI_EXAMPLE  /www.example.com\/OrderProduct\//
uri      _LOCAL_URI_OURDOMAIN  /www.ourdomain.com\/getservice\//
```

```
meta _LOCAL_HEADER_BODY_URI ( LOCAL_RULE_SUBJECT &&
LOCAL_RULE_FROM && LOCAL_SCHOLARSHIP && LOCAL_LUCKY &&
_LOCAL_URI_EXAMPLE && _LOCAL_URI_OURDOMAIN)
score LOCAL_HEADER_BODY_URI 0.5
```

Above rule will search header of the mail for subject and from field, body of mail for the two words scholarship and lucky and for specified URIs in the body of mail, after that it will assign combined score if both words occurs in the e-mail.

2.7.2.4 URI Rules

URI rules very simple, they only match text in the URI's in HTML sections of mail. This is very useful for searching links containing spam advertised sites or adult sites. Following is an example URI rule:

```
uri LOCAL_URI_EXAMPLE  /www.example.com\/OrderProduct\//
score LOCAL_URI_EXAMPLE 0.5
```

Above rule will search body of the e-mail for URI www.example.com.

2.7.2.5 Compilation of rules

Before testing rules with incoming e-mails, they should be compiled for any kind of error. **spamassassin** command is used with **--lint** option. Syntax checks (lint) the rule set and configuration files, reporting typos and rules that do not compile correctly. Exits with 0 if there are no errors, or greater than 0 if any errors are found.

2.7.2.6 Creating Configuration File

A separate file can be created for a particular kind of spam. This file may contain all kinds of rules like: header rules, body rules, Meta rules and URI rules and also can contain list of blacklisted and white listed senders. This file can be saved on **/etc/mail/spamassassin** location with **.cf** extension. This file needs to be checked for any kind of error while creating rules. All the rules will fire automatic when similar kind of

pattern will be seen in incoming e-mails and corresponding scores will be added to the score of e-mail. **Quang-Anh, Haixin Duan, Xing Li** from Network Reseach Centre, Tsinghua University, and Beijing, China have generated Chinese_rules.cf for the detection of Chinese spam. These types of various **.cf** files can be found for the various categories of spam [21].

There is no doubt that sender address verification approaches are getting better day by day, but rule based approach will be always there. There is always scope of new rules which can be added to make the filter to work better according to the requirements of the user.

Table 2.2: SpamAssassin's Release Description [23]

Release	Year	Rules	Added	Removed
2.4.3	2002	655	655	-
2.6.4	2003	591	283	347
3.0.0	2004	463	187	315
3.1.0	2005	500	159	122
3.2.0	2007	636	367	231
3.3.0	2009	475	92	253

From 2002 to 2010, 25 SpamAssassin versions were released. Each version includes new features, and updated rules. Table given next shows some of these versions. Above given table shows that rules for SpamAssassin are always in continues process. So many rules have been added and removed from each and every release of SpamAssassin.

Chapter 3

Problem Statement

As noticed during the literature survey, there are so many techniques to deal with the Spam problem; all these techniques are being used in so many different kinds of Spam filters. Most filters use Whittling, blacklisting and grey listing as a part of Email envelops analysis and combination of heuristic and Bayesian as Email content analysis. Basically, all these filters classify the Emails in to the category of Spam and non-Spam. Most of the Spam filters decide faith of an incoming Email on the basis of some words in data part or Subject part of the Email or from the source which it is coming, and categorize the Email as Spam. But this is not good enough to categorize Email as Spam.

There should be some group of rules that should be followed to categorize Emails. During the literature survey, we have noticed that the most dangerous problems that occur with the Spam filters are False Positives and False Negative. Even mail server of yahoo and Google are facing these problems. Because of these problems, user lost important Emails which can in turn bring loss in business and user will be certainly dissatisfied with the services of mail service provider. It is hard fact that Spam cannot be tackled 100 percent, but there is need to design some good custom rules, which can be used to identify spam on the basis of content (tokens/words) of Email and the source of Email. For the design and implementation of good rules, following steps are required to be followed:

- Pattern Analysis
- Pattern Selection
- Score Assignment
- Compilation of Rules

Objectives

The objective is to employ the above described process for the implementation of good rules which must be able to detect following categories of Spam:

- Health
- Product offering
- Education
- Adult
- Gambling

Rules will be implemented on SpamAssassin which is an open source filter from Apache. Rules can be implemented in user_prefs, local.cf files of SpamAssassin or these can be implemented by creating separate **.cf** file for each category of Spam.

Chapter 4

Implementation and Experiment

Rule based approach is used for the detection of spams. Rules have been implemented by using SpamAssassin on Ubuntu. 3.2.5 Version of SpamAssassin has been installed on ubuntu 9.04. Analysis of spam mails is done from 3 of Email accounts those includes 2 gmail accounts and 1 yahoo account. These accounts consists 1777 mails in inboxes and 105 mails in spam boxes. Analysis and rules are applied on the content of the mails. This content analysis is performed on of the header mails body of the Emails and URIs contained by the body of an Email. Rules have been written by making configuration files on location **/etc/mail/spamassassin** under **root** all configuration files have **.cf** extension**.** Testing of the rules has been done locally.

4.1. Steps Performed while Implementation

1. Selection of tokens/words done from spam mails from the above given 3 Emails accounts, values are assigned to **N1** and **N2**.

   N1 is occurrence of word in no. of spam mails

   N2 is occurrence of word in no. of ham mails

2. Comparative probability of each token has been calculated regarding their occurrence in spam mails or ham (legitimate) mails.

   W/S is the probability of the word in spam mails

   W/H is the probability of the word in ham mails

3. The rules are written for the tokens with higher probability of spam by creating **.cf** files for each kind of spam at following location:

   /etc/mail/spamassassin

4. Rules are compiled by using following command:

   spamassassin --lint

5. Emails have been read by using Alpine (command line program to read mails). These mails are Exported to the home directory using export feature of Alpine and saved with name of own choice.
6. Rules are fired against the imported Email to detect either it is spam or ham (legitimate mail). It is done by using following command:

   Spamassassin --local name_of_file

4.2 Rules for Blacklisting and White Listing

Some senders are known for sending spam mails they should be marked as blacklisted. Mostly we receive Emails from our near and dear ones or people with whom we are officially connected all sender can be white listed.

4.2.1 Blacklisting

Blacklisting describes the process of blocking upcoming SMTP connections from spammers, which are contained within a list of blacklist. Identify sender addresses where spam is known to originate by placing such addresses or domains on a blacklist. Following rule is an example of how a sender of e-mail can be added to blacklist:

blacklist_from **niteshkumar.maths@gmail.com**

describe **This sender always send SPAM mails**

This rule will see the From: field of the e-mail and if it founds that sender is blacklisted, then a score of **100** will be added to score of e-mail, which is much more than the default threshold value **5.0** of SpamAssassin. It means incoming e-mail from this sender will be automatically declared as spam. Following Figure 4.1 shows the same fact.

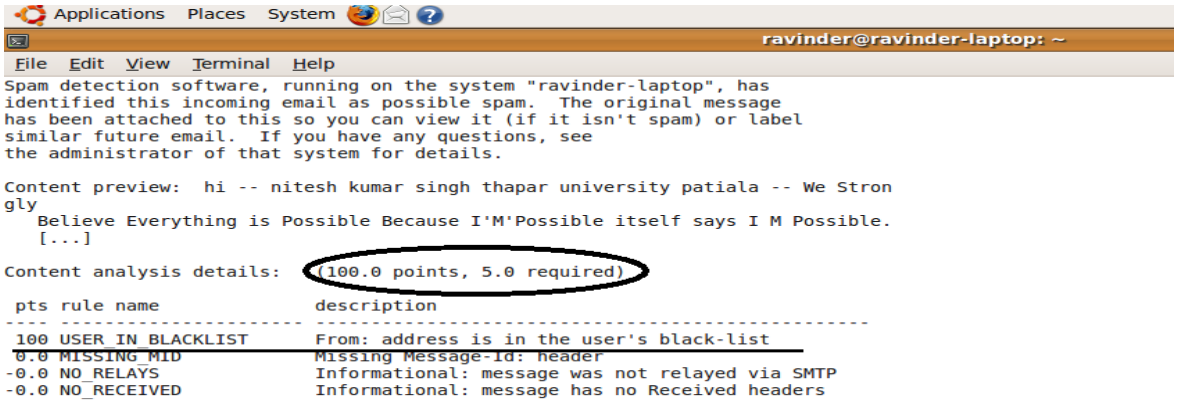

Figure 4.1: Sender Blacklisted

4.2.2 White listing

White lists are opposite to blacklists, they contain users that are verified contacts. All the E-mails from these senders will be treated as ham (legitimate mail). A majority of ham e-mails is sent by well known sources, which do not need to be checked against blacklists. Following rule shows how to add a sender in white list.

whitelist_from **kamboj.ravinder@yahoo.com**

describe **This sender sends legitimate mails**

This rule will add -100 to the score of e-mail, when it will find the same sender in the From: field of the Email. Following figure shows the above given sender as the legitimate sender.

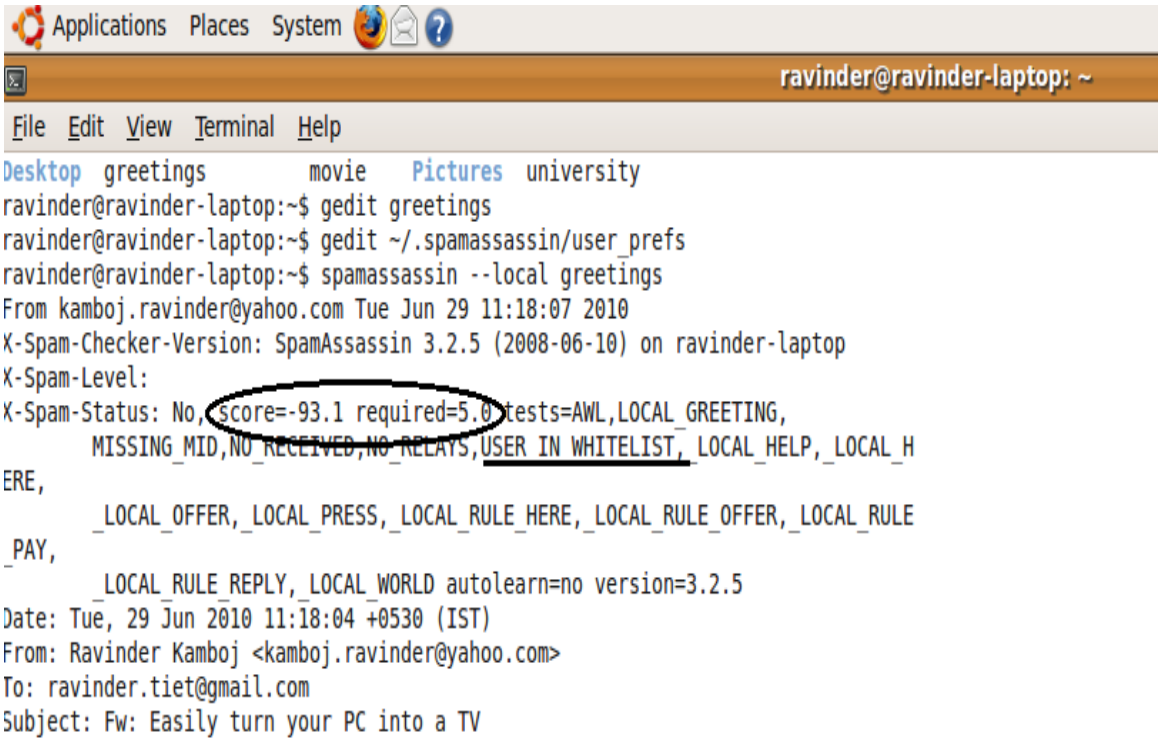

Figure 4.2: White Listed Sender

Above given figure shows that sender of the Email is a legitimate sender, score of **-100** is added but mean while some other rules are applied on same mail then there some deduction from the score is done that is why score of mail is **-93.1**.

4.3 Rules on Content of E-mail

Main focus of our thesis work is on the on the content analysis of the Emails and building rules on the basis of that. These rules are categorized as following:

- Header Rules
- Body Rules
- Meta Rules
- URI Rules

4.3.1 Rule set for Detection of Spam

Various rules are made to detect various kinds of Spam. Rules have been made by using Meta rule feature of **SpamAssassin.** Meta rules are very interesting and useful feature of SpamAssassin, there is an advantage of Meta rules that it can be applied on both part of the Email header part and the body part of the e-mail. Meta rule can be designed for any kind of spam. Following are categories for those rules are created:-

- Products Offering
- Adult
- Health
- Education

4.3.2 Detecting Product Offering Spam

There various spam those offers various kinds of products, these products can be anything starting from greeting cards to credit cards. There are some common words those occur in all these kind of messages and rules can be applied on these common words to detect these kinds of spam. Following is a sample Meta rule to detect these kinds of spam. There an advantage of Meta rules that it can be applied on multiple part of the message like body of the message along with header of the message.

4.3.2.1 Credit Card Spam

So many spam mails can be seen regarding the offering of credit cards. All these mails contain some common words. Rules can be implemented to detect credit card spam. Following are some words those are taken from credit card spam mail.

Table 4.1: Comparative probability of words of Credit Card Spam

Word	N1	N2	W/S	W/H
Credit	5	15	0.05	0.01
Card	6	43	0.06	0.02
Fee	5	32	0.05	0.02
Save	7	95	0.06	0.05
Everytime	1	4	0.01	0.00
Surcharge	2	1	0.02	0.00
Points	3	2	0.04	0.00
Benefit	2	37	0.03	0.02
Profit	0	5	0.00	0.00

Table 4.1 given above shows the occurrence of words in Spam/Ham e-mails and their probability for future occurrences. Following rules are created in **CreditCard.cf** to detect credit card spam.

```
#****************Rules_to_detect_credit_card_SPAM**********

header _LOCAL_FROM_DONOTREPLY From =~
/donotreply\@mail1\.way2sms.biz/i

body _LOCAL_CREDIT  /credit/i

body _LOCAL_CARD    /card/i

body _LOCAL_FEE          /fee/i

body _LOCAL_SAVE    /save/i

body _LOCAL_EVERYTIME    /everytime/i
```

```
body _LOCAL_SURCHARGE    /sucharge/i

body _LOCAL_POINTS  /points/i

meta  _LOCAL_HEADER_WORDS2 (_LOCAL_FROM_DONOTREPLY &&
_LOCAL_CREDIT || _LOCAL_CARD &&
_LOCAL_FEE && _LOCAL_SAVE && _LOCAL_EVERYTIME &&
_LOCAL_SURCHARGE && _LOCAL_POINTS)

score _LOCAL_HEADER_WORDS2 0.5
```

After applying rules for the detection of credit card Spam, result is shown in Figure 4.3.

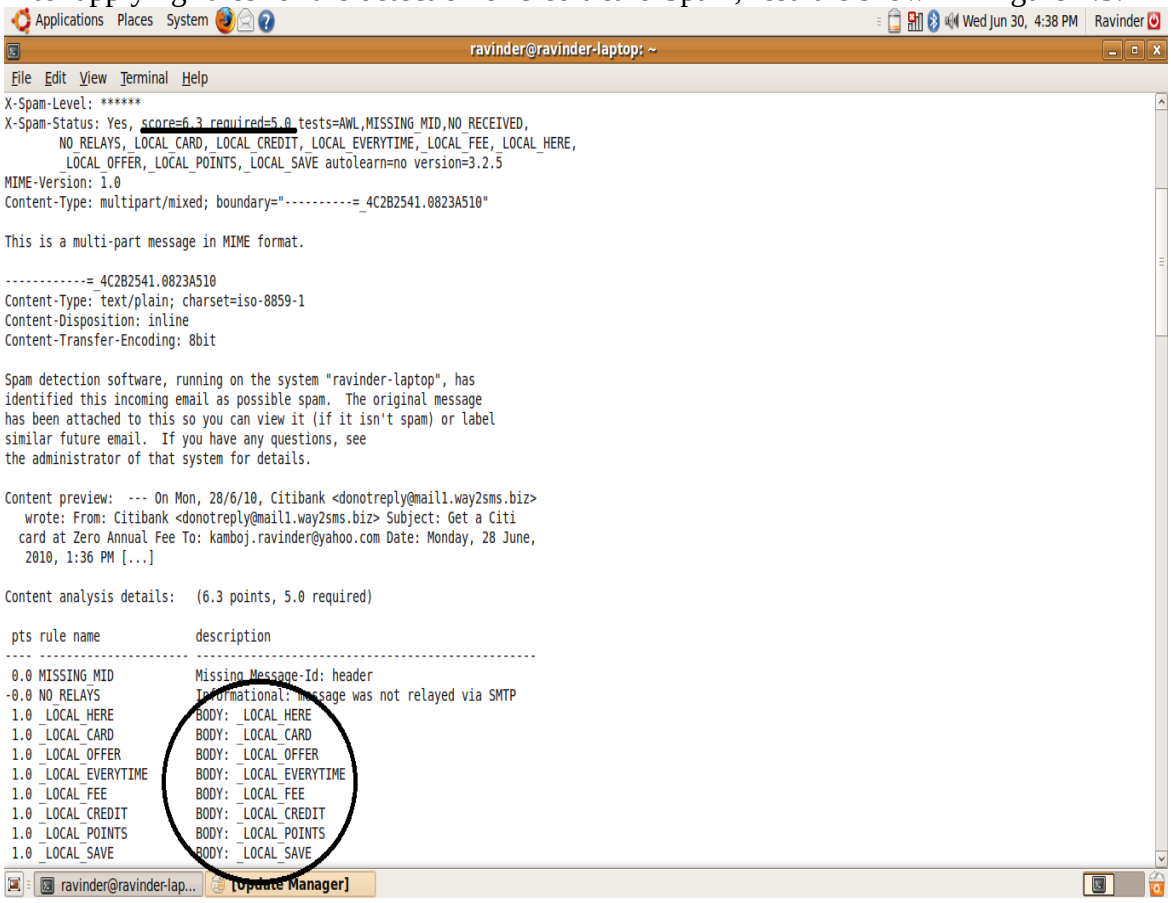

Figure 4.3: Detection of Credit Card Spam

4.3.3 Detection of Adult Spam

Adult spam is one of the major spam those come into the inboxes of e-mail users. These spams are mostly about the joining of some group to see live adult cams; these can be

also online dating etc. There are some common words those comes in most of adult Spams body. It is easy to recognize these words and rules can be made to detect adult Spams on the bases of the words. Table given below shows some common words those comes in adult spam

Table 4.2: Comparative probability of Words of Adult Spam

Word	N1	N2	W/S	W/H
Adult	5	0	0.05	0.00
Click	45	290	0.43	0.16
Horny	0	0	0.00	0.00
Webcam	0	2	0.00	0.00
Membership	4	58	0.04	0.03
Sex	1	7	0.01	0.00
Reply	18	296	0.17	0.17
Access	5	82	0.05	0.05
Accept	5	49	0.05	0.03
Ignore	1	25	0.01	0.01
Apologies	2	8	0.02	0.00

Table 4.2 shows the corresponding N1 and N2 are calculated and probabilities are also calculated according to their occurrences. Following rules are created in **Adult_Spam_Detect.cf** to detect adult spam.

```
#************************Body_Rules**********************
body LOCAL_ADULT    /adult/i
score LOCAL_ADULT   0.5
describe This word can be found in most of the adult Spam
body LOCAL_CLICK    /click/i
score LOCAL_CLICK   0.4
describe This word can be found in Spam mails
body LOCAL_HORNEY   /horney/i
score LOCAL_HORNEY  0.2
describe This word can be found in Spam mails
body LOCAL_SEX      /sex/i
score LOCAL_SEX          0.1
body LOCAL_IGNORE   /ignore/i
```

```
score LOCAL_IGNORE  1.0
body LOCAL_ACCEPT   /accept/i
score LOCAL_ACCEPT  0.4
#**********************HEARDER_RULES**********************
header    LOCAL_FROM_ADULT    From =~
/yahoogroups\@yahoo\.com/i
score     LOCAL_FROM_DONOTREPLY    1.0
header    LOCAL_SUBJECT_OFFER Subject =~     /adult
webcam/i
score     LOCAL_SUBJECT_OFFER 1.0

#******META_RULE_TO_DETECT_ADULT _SPAM*********************
header    _LOCAL_ADULT_SUBJECT     Subject =~
/horneyadultwebcam/i
header    _LOCAL_ADULT_FROM   From =~
/yahoogroups\@yahoo\.com
body _LOCAL_RULE_HORNEY  /horny/i
body _LOCAL_RULE_ADULT   /adult/i
body _LOCAL_RULE_WEBCAM  /webcam/i
body _LOCAL_RULE_MEMBER  /membership/i
body _LOCAL_RULE_CLICK   /click/i
body _LOCAL_RULE_SEX          /sex/i
body _LOCAL_RULE_APOLOGIES    /aplogies/i
body _LOCAL_RULE_REPLY   /reply/i
body _LOCAL_RULE_ACCEPT  /accept/i
body _LOCAL_RULE_IGNORE  /ignore/i
uri  _LOCAL_RULE_URI1    /www\.groups\.yahoo\.com/
meta _LOCAL_ANTI_ADULT_RULE    (    _LOCAL_ADULT_SUBJECT    &&
_LOCAL_ADULT_FROM     &&     _LOCAL__LOCAL_RULE_HORNEY     &&
_LOCAL_RULE_ADULT       &&       _LOCAL_RULE_WEBCAM       &&
_LOCAL_RULE_MEMBER  &&  _LOCAL_RULE_CLICK  &&  _LOCAL_RULE_SEX
&&    _LOCAL_RULE_APOLOGIES    &&    _LOCAL_RULE_REPLY    &&
```

```
_LOCAL_RULE_ACCEPT        &&        _LOCAL_RULE_IGNORE        &&
_LOCAL_RULE_URI1)
score      _LOCAL_ANTI_ADULT_RULE 1.0
```

this meta rule will help to detect most of the adult spams those are coming to the users of mail service. Figure given next shows the detection of adult SPAM on the basis of above rule. **X-Spam-Status** is shown as yes and score of the mail is calculated as **7.8** which is greater than the standard threshold value.

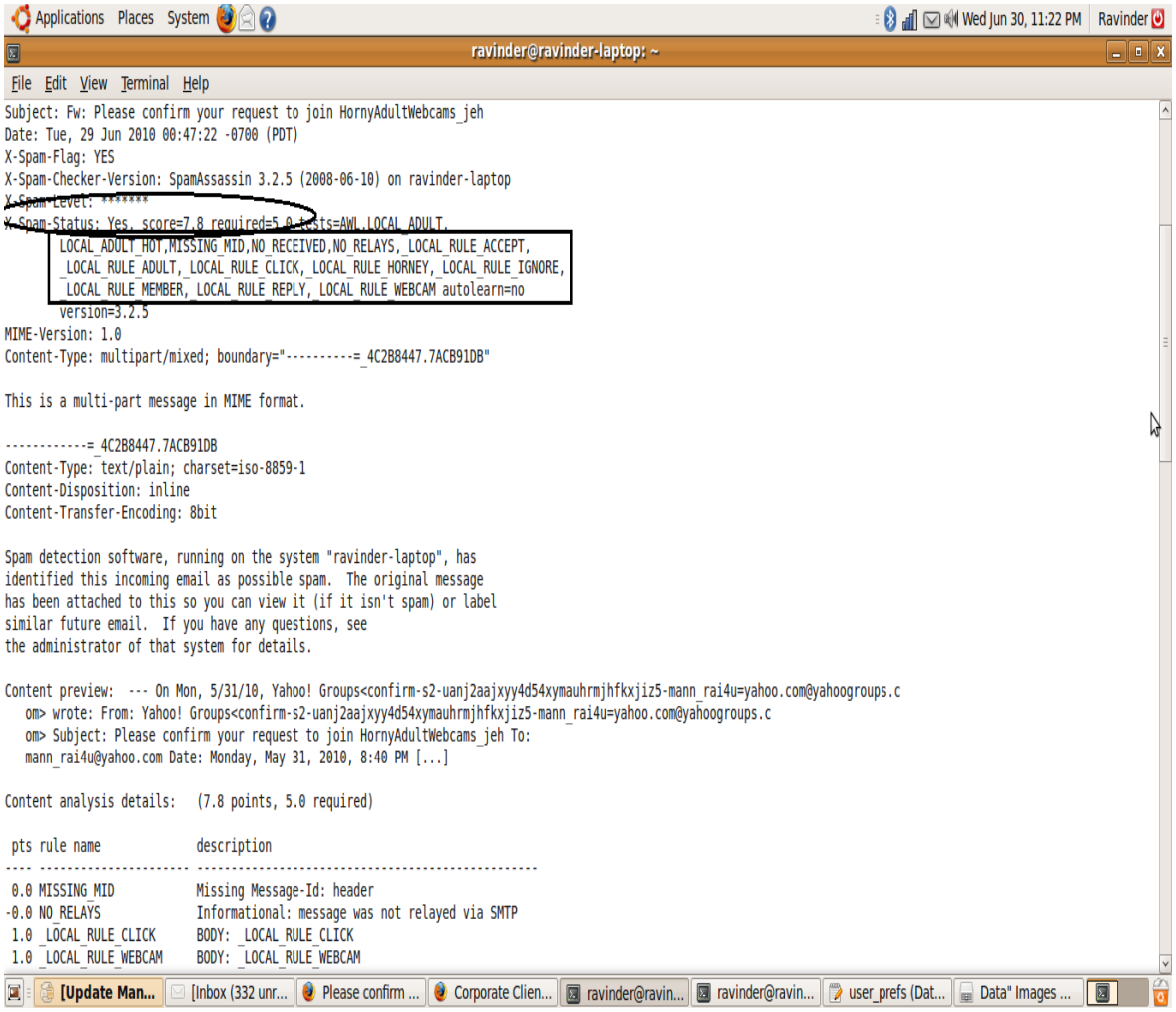

Figure 4.4: Detection of Adult Spam

4.3.4 Detection of Health Spam

Mostly SPAM is send regarding weight problems, skin problems, heart problems and sex Problems. These E-mails are written in such a manner that attracts the reader of E-mail and force them to click over the link of given **URI** in the body of Email. One can see those Email SPAM ads that about the latest herbal remedy that promises dramatic weight

loss. One of the most common SPAM falls under this category is **Viagra Spam.** Table 4.3 shows some words which are very common in **Viagra Spam**.

Table 4.3: Comparative probability of words of Viagra Spam

Word	N1	N2	W/S	W/H
Offer	6	51	0.06	0.03
Click	45	290	0.43	0.16
Dear	10	316	0.10	0.18
Here	50	512	0.48	0.29
Site	12	326	0.11	0.18
Local	3	42	0.03	0.02
Powerful	2	12	0.02	0.01
Amazing	3	21	0.03	0.01
Energy	1	31	0.01	0.02

Following rule are created in **Viagra_Spam_Detect.cf** to detect Viagra Spam.

```
#***************META_RULE_TO_DETECT_VIAGRA_SPAM************

header    _LOCAL_SUBJECT_VIAGRA    Subject =~     /viagra/i

header    _LOCAL_SUBJECT_OFFICIAL Subject =~ /official/i

body _LOCAL_RULE_VIAGRA  /viagra/i

body _LOCAL_RULE_OFFER   /offer/i

body _LOCAL_RULE_CLICK   /click/i

body _LOCAL_RULE_DEAR    /dear/i

body _LOCAL_RULE_HERE    /here/i

body _LOCAL_RULE_SITE    /site/i
```

```
meta META_VIAGRA_SPAM_DETECTION_RULE (
_LOCAL_SUBJECT_VIAGRA && _LOCAL_SUBJECT_OFFICIAL &&
_LOCAL_RULE_VIAGRA && _LOCAL_RULE_OFFER &&
_LOCAL_RULE_CLICK && _LOCAL_RULE_DEAR && _LOCAL_RULE_HERE
&& _LOCAL_RULE_SITE)

score      META_VIAGRA_SPAM_DETECTION_RULE 1.0
```

After implementation of this rule it becomes easy to detect Viagra spam. Figure given next shows the detection of Viagra SPAM after applying the rule given above.

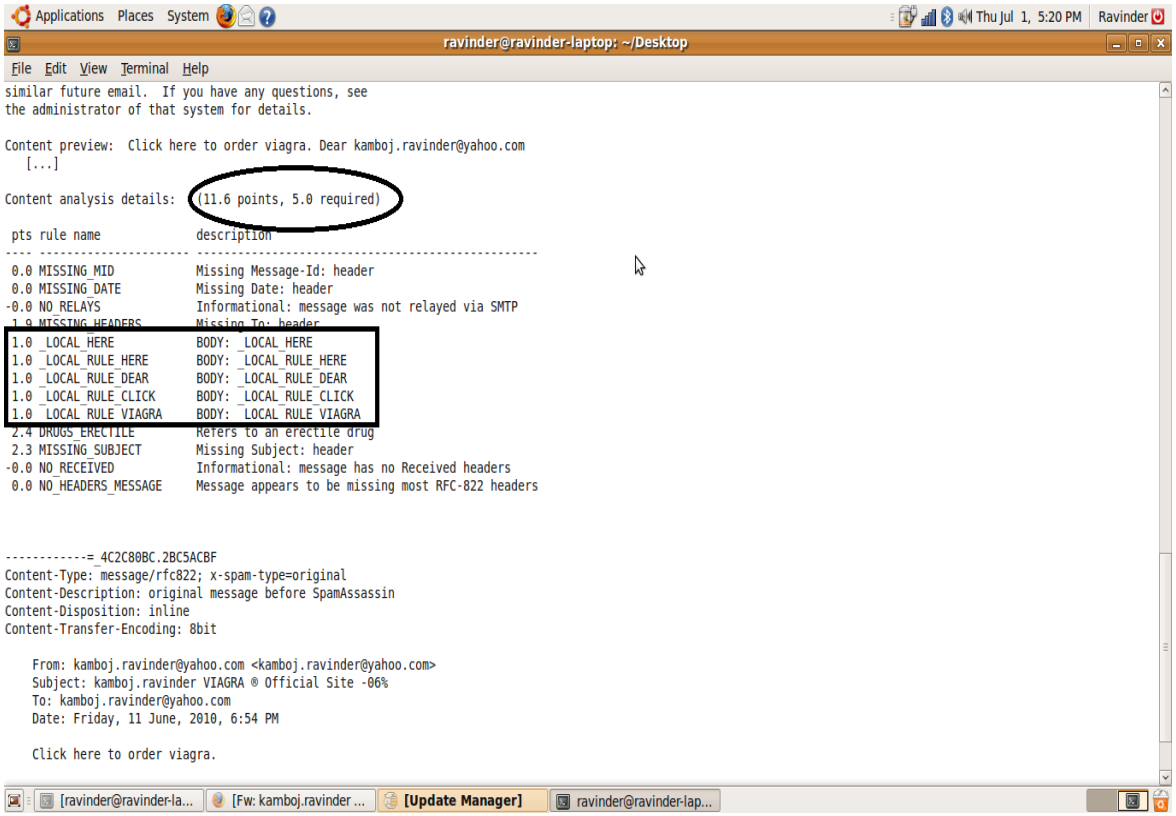

Figure 4.5: Detection of Viagra Spam

4.3.5 Detection of Gambling Spam

So many mails are flowing on Internet those are containing links to gambling sites. These mails are used to promote Gambling sites by offering some schemes to the receiver of mail. Various e-mails have been received in past regarding lotteries. At one time Australian lottery was famous e-mail spam; other similar kind of lottery spam also exits. All lottery spams have some common key words those cannot be forged. This is the

weakness of lottery spam, and content based filters can take advantage of this weakness by a making rule upon these words. The table given next shows some words which are very common in the Lottery Spam.

Table 4.4: Comparative Probability of Words of Lottery Spam

Word	**N1**	**N2**	**W/S**	**W/H**
Lottery	5	4	0.05	0.00
Award	6	20	0.06	0.01
Pay	4	67	0.04	0.04
International	6	78	0.06	0.04
Ticket	5	7	0.05	0.00
Lucky	4	16	0.04	0.01
Congratulation	2	2	0.02	0.00
Claim	5	6	0.05	0.00
Prize	5	11	0.05	0.01

After analyzing the values of table rule has been made for the words with higher SPAM probabilities. Following rules are created in **Lottery_Spam_Detect.cf** for the detection of Lottery SPAM.

```
#**********************Body_Rules************************
body LOCAL_TICKET   /ticket/i
score LOCAL_TICKET 0.7

body LOCAL_CHOOSEN  /choosen/i
score LOCAL_CHOOSEN 1.0
describe This word can be found in Spam mails
body LOCAL_LUCKY    /lucky/i
score LOCAL_LUCKY   0.5
describe This word can be found in Spam mails
body LOCAL_CLAIM         /claim/i
score LOCAL_CLAIM   1.0
body LOCAL_SURCHARGE          /surcharge/i
```

```
score LOCAL_SURCHARGE    1.0
body LOCAL_MONEY    /money/i
score LOCAL_MONEY   0.5
#**********************HEARDER_RULES**********************
header    LOCAL_FROM_DONOTREPLY    From =~
/donotreply\@mail1\.way2sms.biz/i
score     LOCAL_FROM_DONOTREPLY    1.0
header    LOCAL_SUBJECT_CONGRATULATION  Subject =~
     /congratulation you are winner/i
score     LOCAL_SUBJECT_CONGRATULATION  1.0
#**********************META_RULES************************

header    _LOCAL_SUBJECT_CONGRATULATION Subject =~
     /congratulation/i
header    _LOCAL_SUBJECT_WINNER    Subject =~     /winner/i
body _LOCAL_LOTTERY /lottery/i
body _LOCAL_AWARD   /award/i
body _LOCAL_PAY          /pay/i
body _LOCAL_TICKET  /ticket/i
body _LOCAL_INTERNATINAL /internatinal/i
body _LOCAL_LUCKY   /lucky/i
body _LOCAL_RECIEVE /recieve/i
meta LOCAL_META_LOTTERY_SPAM
(__LOCAL_SUBJECT_CONGRATULATION && _LOCAL_SUBJECT_WINNER &&
_LOCAL_LOTTERY && _LOCAL_AWARD && _LOCAL_PAY &&
_LOCAL_TICKET || _LOCAL_INTERNATINAL && _LOCAL_LUCKY &&
_LOCAL_RECIEVE)
score  LOCAL_META_LOTTERY_SPAM 1.0
```

After implementation of this rule detection of lottery SPAM becomes easy, that is shown in the figure given next.

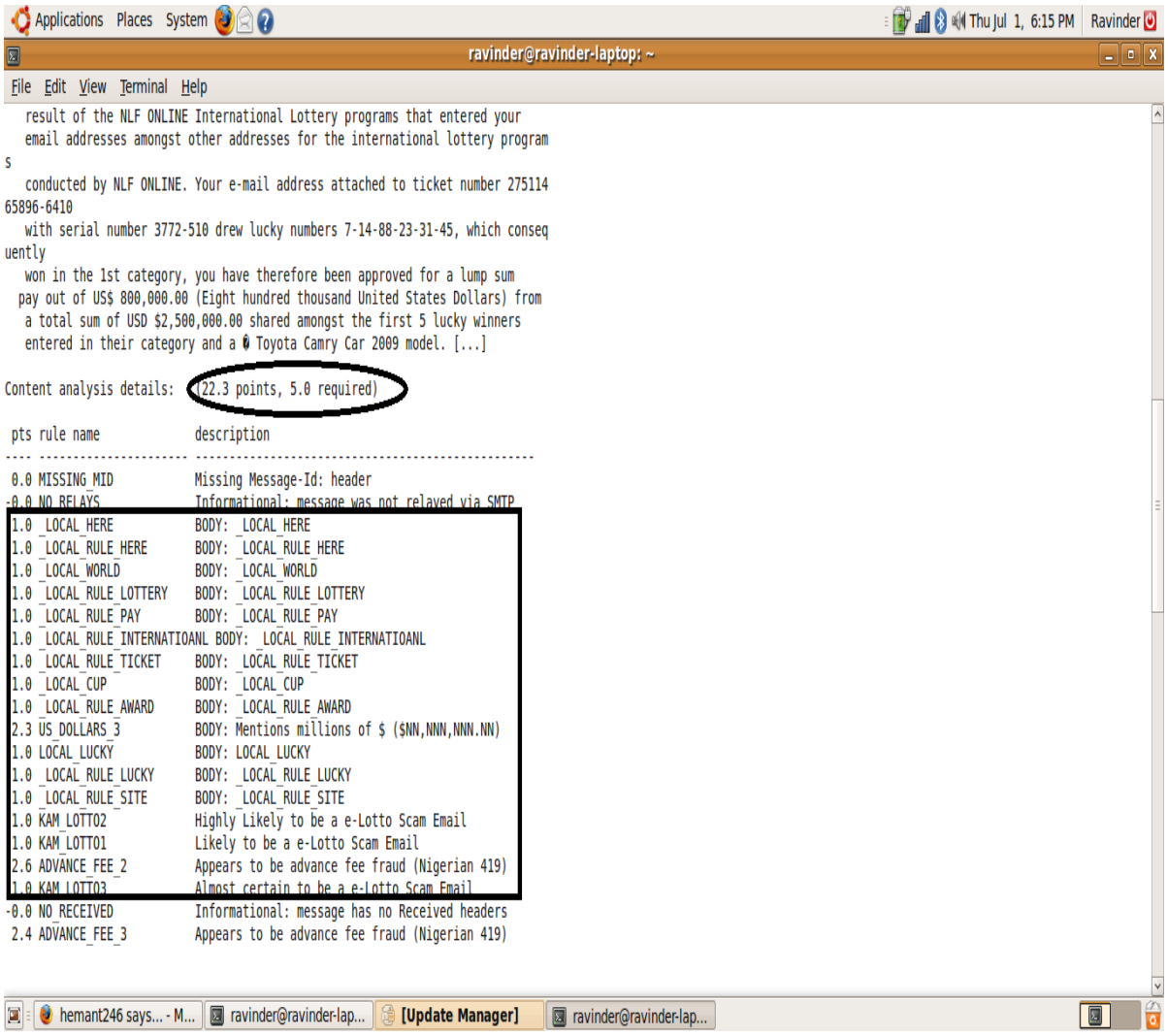

Figure 4.6: Detection of Lottery Spam

4.3.6 Detection of Educational Spam

Educational spam comes mostly in the inboxes of young generation, educational institutes, schools and universities target them to promote their institutes by offering scholarship schemas and various facilities. Even some time Spam from some institute can be seen which is offering degree at a handsome cost. To this they make use of Spam. All these kind of spams contains some common words those cannot be forged by spammer to keep the readability of the mail. Table contains a list of few words which are very common in these types of spams.

Table 4.5: Comparative Probability of Tokens of Educational Spam

Word	N1	N2	W/S	W/H
Scholarship	5	10	0.05	0.01
Opportunity	7	6	0.07	0.00
Offer	15	54	0.14	0.03
Facilities	6	12	0.06	0.01
Study	5	37	0.05	0.02
Lucky	3	15	0.03	0.01
Fee	22	32	0.21	0.02
Refund	2	1	0.02	0.00

Following are rules to detect educational Spam, written in **Edu_Spam_Detection.cf** by including words chosen with higher probability of those occurrences in spam mails.

```
#************************Body_Rules************************
body LOCAL_STUDY /^study/i
score LOCAL_STDUY 0.4

Describe this word is very common in Educational Spams

body LOCAL_INTERVIEW /^interview/i
score LOCAL_INTERVIEW 0.4

body LOCAL_WELCOMES /welcomes/i
score LOCAL_WELCOMES     1.0

body LOCAL_UNIVERSITY    /university/i
score LOCAL_UNIVERSITY   1.0

body LOCAL_INFORMATION   /information/i
```

```
score     LOCAL_INFORMATION   0.5

body LOCAL_AWARD    /award/i
score     LOCAL_AWARD    0.4

#**********************Hearder_Rules************************
header LOCAL_SUBJECT_CONRATULATION Subject=~
/congratulatio/i
score     LOCAL_SUBJECT_CONGRATULATION  0.5

header LOCAL_SUBJECT_STDUY_IN_UK   SUBJECT=~ /STUDY IN UK/i
score  LOCAL_SUBJECT_STUDY_IN_UK   1.0

#************************Meta_Rules************************
header _LOCAL_FROM_MARKETING  From =~
/marketing\ecindia\.net/i
body _LOCAL_STUDY   /study/i
body _LOCAL_INTERVIEW    /interview/i
meta _LOCAL_STUDY_INTERVIEW (_LOCAL_FROM_MARKETING &&
_LOCAL_STUDY || _LOCAL_INTERVIEW)
score _LOCAL_HEADER_WORDS4 1.0
body _LOCAL_STUDENT /Student/i
body _LOCAL_CAN          /can do/i
body _LOCAL_PART    /part time/i
body _LOCAL_JOB          /job/i
meta      LOCAL_STUDY_IN_UK (_LOCAL_STUDENT && _LOCAL_CAN
&& _LOCAL_PART && _LOCAL_JOB)
score     LOCAL_STUDY_IN_UK   1.0
```

This Meta rule will help to detect educational spam. Screen Shot given in Figure 4.7 shows the detection of educational spam after applying these rules.

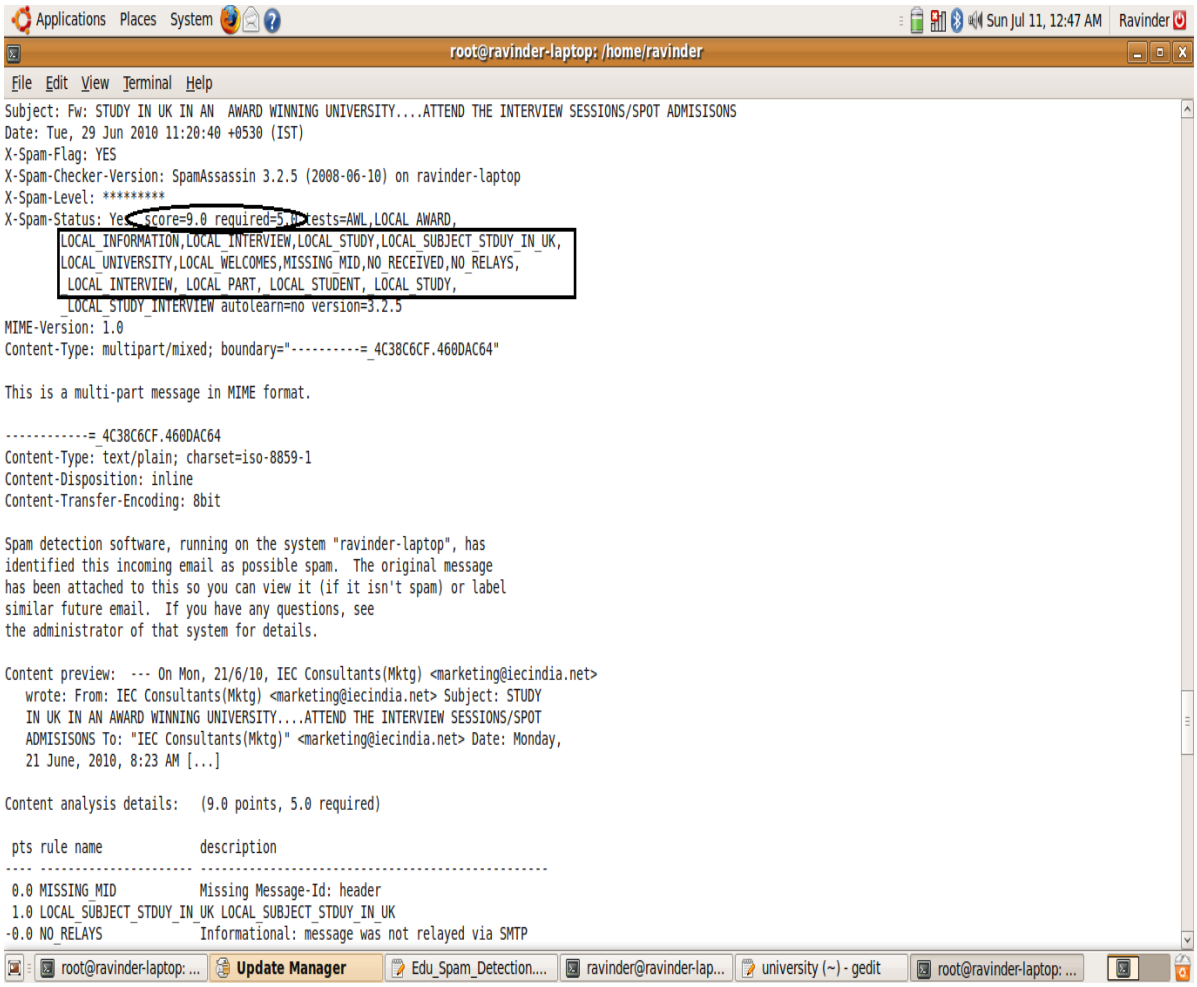

Figure 4.7: Detection of Educational Spam

Chapter 5
Conclusion and Future Work

5.1 **Conclusion**

These days E-mail system is facing a serious and irritating problem like Spam. Spam is a cheap and easiest way of advertisements for the spammers. Spam causes the problems like loss of time, money and bandwidth. Problem of spam is needed to be solved. Spam filters are used to solve this problem. These filters make use of various techniques which are based on envelop of the E-mail and data of the E-mail. Spam filters follow two approaches to tackle spam, one is learning approach and other is rule-based approach. So many Spam filters are there, SpamAssassin is most popular and effective Spam filter which can be used to detect various kind of Spams.

In this thesis work Rule based approach has been followed by designing various rules for the detection of Spam on SpamAssassin. Blacklisting, White listing and Bayesian filtering have been used. Meta rules, body rules and header rules have been implemented for the detection of Spam. All the Rules have been fired on the Emails imported from Alpine. Various kinds of Spam have been detected like:

- Health spam mostly flows on internet; Spam comes under this category like Viagra Spam has been detected.
- Product offering Spam contains offers of like software, computers and credit cards etc. Credit card Spam has been detected by use of custom rule.
- Adult like invitation regarding joining of some adult group or dating site has been detected.
- Rules have been implemented in .cf file for the detection of educational Spam and Spam has been detected regarding abroad university offering various courses and scholarship schemes.
- Lottery Spam is one of the famous Spam; it has been detected by implementing rules for this kind of Spam.

5.2 Future Work

There is an arms race between spammers and anti-spammers, so possibility for the improvement of rules and building new kind of rules for new kind of Spams is always there. Combination of learning and rule based approach can provide better results, because rule based approach not adoptable to the changing Spams.

In this thesis work rules could not be implemented for some kinds of spam like: Stock, Phishing and Finance, because dataset related to these kinds of spams could not be collected. So requirement is there for building of new rules to detect above given kind of spams.

References

[1]. Md. Rafiqul Islam, Morshed U.Chowdhury, “Spam Filtering Using ML Algorithms”, IADIS International Conference on WWW/Internet, 2005

[2]. Christain Rossow “Anti-spam measures of European ISPs/ESPs”, August 2007

[3]. “You’ve got Spam: how to “can” unwanted E-mail, Federal Trade Commission, Bureau of Consumer Protection, 2002.

[4]. “What the Federal CAN-SPAM Act Means for Commercial E-mail Marketers”, Return Path

[5]. Christine E. Drake, Jonathan J. Oliver, and Eugene J. Koontz, “ Anatomy of a Phishing E-mail”

[6]. J. Stewart, "Spam Botnets to Watch in 2009," 2009, http://www.secureworks.com/

[7]. “Greylisting – Simple Concept, Highly Effective Technique”, GFI White Paper

[8]. Ing.Shahzihan A.Choudhary “Anti-spam masters project,” Centrum Voor Wiskunde en Informatica, University van Amsturdum. Augest 2004

[9]. “Intrusion detection system”, Wikipedia, the free encyclopedia

[10]. “Network Security: A guide for small and medium businesses,” a Star Technology White Paper, http://www.star.net.uk

[11]. Subhjeet Choudhary, Bisvanath Dey “Spam a Threat to Network Security in Digital Library and Information Centers” , INFLIBNET center,Ahemdabad , 3rd convention PLANNER-2005,Assam Uni, Nov 2005,.

[12]. Taughnnock Networks “Technical Responses to Spam”, http://www.info@taug.com

[13]. Tuomas Auro “Network Security: Security and threats” Microsoft research,UK

[14]. Kobkiat Saraubon Benchaphon Limthanmaphon “Fast Effective Botnet Spam Detection”, Fourth International Conference on Computer Sciences and Convergence Information Technology, 2009.

[15]. J. Carr, "TRACE: Six botnets generate 85 percent of spam," in SC Magazine, 2008.

[16]. http://royal.pingdom.com/2010/01/22/internet-2009-in-numbers/

[17]. http://www.circleid.com/posts/20081217_spam_200_billion_per_day_2008_cisco/

[18]. SpamAssassin “The Apache SpamAssassin Project” Date of visit: May09,2010 URL: http//spamassassin.apache.org/

[19]. Areej Al-Bataineh, Greogry White, “Detection and Prevention Methods of Botnet-generated Spam”, University of Texas San Antanio.

[20]. Pablo Daniel Aguero, Jorge Castineiera Moreira, Monica Liberatori, Juan Carlos Tulli, “Improving The Performance of Anti-Spam Filters Using Out-of-Vocabulary Statics,” Ingeniare. Revista Chilena de ingenieria, vol. 17 No. 3, pp. 386-392, 2009

[21]. Quang-Anh Tran, Haixin Duan, Xing Li, “Real-Time Statistical Rules for spam detection,” IJCSNS International Journal of Computer Science and Network Security, VOL.6, No.2b, February 2006

[22]. Edoardo Airroldi, Bradley, Malina, Latanya Sweeney, “Technologies to Defeat Fraudulent Schemes Related to Email Requests,” American Association for Artificial Intelligence, Spring Syposium on AI Technologies for Homeland Security, 2005

[23]. Pedro H. Calais Guerra, Dorgival Guedes, Wagner Meira Jr. Cristine Hoepers, Marcelo H. P. C. Chaves, Klaus Steding-Jessen, “Exploring the Spam Arms Race to Characterize Spam Evaluation”, Redmond, Washington USA, CEAS 2010

[24]. Didier Colin, Catherine Roucairol, Ider Tseveendorj, “Selective Learning Model for Spam Detection”, March 2009

[25]. Jonathan A. Zdziarski, “Ending Spam - Bayesian Content Filtering and the Art of Statistical Language Classification”

[26]. http://www.securitysoftwarezone.com/types-of-spam-review81-4.html

List of Paper Published/Communicated

1) Ravinder Kamboj, Dr. V.P Singh, Mrs. Sanmeet Bhatia, "Spam a Major Threat to Network Security: A Review of Spam and Anti-spam Techniques" is communicated in National Conference on *ADVANCES AND RESEARCH IN TECHNOLOGY, "ART – 2010",*19-21 August, 2010, organized by Yamuna group of Institutions, Yamuna Nagar, Haryana.