NTNU

Norwegian University of
Science and Technology

An Investigation of Spam Filter Optimaltiy

based on Signal Detection Theory

Kuldeep Singh

Master in Security and Mobile Computing

Submission date: June 2009

Supervisor: Øivind Kure, ITEM

Co-supervisor: Audun Jøsang (Professor), UNIK, Oslo
Sasu Tarkoma (Professor), TKK, Finland

Norwegian University of Science and Technology
Department of Telematics

Problem Description

Spam can be described as useless messages that pollute peoples email inboxes. The purpose of spam from the sender's point of view can e.g. be for marketing, for spreading malware or it can be an element in criminal phishing attacks. It is estimated that about 97% of all email traffic passing through the Internet is spam. Spam is increasingly sent by botnets that currently are infecting millions of computers worldwide. Spammers can obtain mailing list e.g. by trading/exchanging Internet mailing list, by stealing such lists or by searching the Internet for the email addresses. A single email spam message has virtually no cost to the sender, but a real and noticeable cost to the recipient. This is in contrast to e.g. normal advertisement sent through the mail, which has a real cost to the sender.

In order to eliminate spam, organisations apply spam filters in their handling of incoming email. A problem with spam filters is that they always produce false negatives, false positives or both. No spam filter is 100% effective. The ratios of false negatives and false positives can often be tuned in spam filters. The optimal tuning of spam filters will be a function internal variables of an e-mail.

The objective of this Masters project is to investigate the optimality of spam filters from the specific user's point of view. This will be done by determining the ratios of false negatives and positives in specific spam filters, and by estimating the cost of false negatives and false positives.

Assignment given: 24. January 2009
Supervisor: Øivind Kure, ITEM

Acknowledgements

I would like to thank all people who have helped and inspired me during my master thesis.

I especially want to thank Prof. Audun JØsang, for his guidance during my research at Graduate Research Center(Unik). His perpetual energy and enthusiasm in research had proved to be a great motivation in completing this thesis. In addition, he was always accessible and willing to help with the research. As a result, research life became smooth and rewarding for me.

My sincere thanks go to Mona Nordaune, Executive officer (Department of Telematics-NTNU) and Eija Kujanpää -Planning Officer(NordSecMob); TKK, for helping in various ways to clarify the things related to my academic works in time with excellent cooperation and guidance.

Lastly, I would like to thank my friends and colleagues Md. Sadek Ferdous and Ravishankar Borgaonkar with whom it was nice to share thoughts and I thank for their support during the thesis.

Kjeller, June 27th 2009

Kuldeep Singh

Abstract

Unsolicited bulk email, commonly known as spam, represents a significant problem on the Internet. The seriousness of the situation is reflected by the fact that approximately 97% of the total e-mail traffic currently (2009) is spam. To fight this problem, various anti-spam methods have been proposed and are implemented to filter out spam before it gets delivered to recipients, but none of these methods are entirely satisfactory. This thesis analyzes the properties of spam filters from the viewpoint of Signal Detection Theory (SDT). The Bayesian approach of Signal Detection Theory provides a basis for determining the tuning of spam filters from the particular user's point of view and helps in determining the utility which the spam filter provides to the user.

Abbreviations and Acronyms

E-mail	Electronic Mail
ISP	Internet Service Provider
CO2	Carbon Dioxide
SDT	Signal Detection Theory
MUA	Mail User Agent
Cc	Carbon copy
Bcc	Blind Carbon copy
SMTP	Simple Mail Transfer Protocol
POP3	Post Office Protocol Version3
MDA	Mail Delivery Agent
ARPANET	Advanced Research Projects Agency Network
DNS	Domain Name System
HTML	Hyper Text Markup Language
MIME	Multipurpose Internet Mail Extensions
URL	Uniform Resource Locator
IP	Internet Protocol
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative
FA	False Alarm
CI	Correct Identification
CR	Correct Rejection
TPR	True Positive Rate
FPR	False Positive Rate
TNR	True Negative Rate
FNR	False Negative Rate
ROC	Receiving Operating Characteristics
LR	Likelihood Ratio
STI	Subjective Tuning Index
SFRG	Spam Filter Rationality Graph

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Spam: An unsolicited bulk email.

Spam is a huge and growing problem. The amount of spam that circulates through the Internet and that gets delivered to email clients is increasing day by day, and is affecting everyone on the Internet, ranging from network providers to Internet Service Providers (ISPs) and end users. Viewing spam everyday in the in-box is annoying and time consuming for all Internet users. In Nielsen (2008)[1] it was found that approximately 97% of the total email traffic consists of spam. The increasing amount of spam has attracted the attention of Internet and security experts. As a result many anti spam strategies have been proposed and implemented. Current work also investigates methods to completely block spam. The reason behind getting attracted to spam is that spam messages are viewed as a serious threat to the internet, leading to flooding users' in-boxes, costing users and ISPs with extra time and money, and becoming a means of doing fraud. Therefore, it becomes very important to contain the spam messages over the internet.

1.1 Motivation

The motivation for carrying out this thesis work has been mentioned in the following sub-subsection.

1.1.1 Fraud

Since large numbers of spam messages are received by internet users every day, therefore, spammers employ different fraudulent methods to encourage users to open the spam messages in-order to obtain users private information. The simplest way to encourage user is to alter the subject line of the email in such a way that implies that the message is not a spam.

There are different types of frauds which are carried out by spammers. For example, phishing attack and 419 Scams. Phishing is a criminally fraudulent method that makes an attempt to acquire private information like, credit card details, passwords by pretending to be an authentic and trustworthy entity on the internet. Phishing can be also viewed as a social engineering technique used to fool users. And 419 Scam is a trick used to take the recipient in to confidence and persuade the recipient to transfer a sum of money in hope realizing a significant larger profit [14, 8].

1.1.2 Recipient bearing the cost

The main reason behind the increasing amount of spam lies in the cost imbalance between senders and recipients. Sending large amounts of spam has a very small cost compared to the relatively high cost of viewing and deleting a single spam message. Millions of emails can be sent per hour with just 56 kbps of bandwidth[7]. According to[20], if even one among 500,000 spam messages of direct-mail print campaigns attracts a recipient to buy the product then the whole cost incurred in sending 500,000 spams is covered. On the other hand the recipients and the ISPs have to carry significant costs. The most obvious cost is the bandwidth consumed for processing spam. In large organization the charging for Internet connections is based on traffic, and because of spam traffic these firms end up paying significant amounts for non-productive traffic. On the ISP side the cost comes from wasted bandwidth and CPU time. If the consumption of the bandwidth is significantly large due to spam messages (which is generally the case) then the scenarios that are faced by the ISP can be categorized as follows:

1. Increasing the internet usage charges in order to compensate the bandwidth getting wasted by spam messages.
2. Continuing to provide the internet service with a slower speed because of the spam messages.
3. Absorbing the cost of the wasted bandwidth.

With these scenarios, ISPs generally prefer to go with the 1st i.e. increasing the internet usage charges with directly effects the subscribers. This scenario can also be seen as a cost shift scenario where recipients of the spam messages are paying instead of the ISPs.

1.1.3 Wastage of resources

Large numbers of spam messages are causing a severe problem of traffic congestion over the network. This leads to significant level of resource wastage. Routers in the network are forced to handle unwanted traffic sent to millions of users. Therefore, apart from user end, resources are also getting consumed in the network. It is problematic to filter spam messages at the router level. Filtering at the router level also has undesirable impact on throughput. Since, spam messages get delivered to respective recipients it is regarded as the wastage of network resources because spam messages are normally deleted as they reach their destination.

In addition to this considering the time as a resource it has been found that significant amount of time is wasted around spam. For example in a survey, conducted in 2006 among employees of 500 large companies in US and Finland, it was found that on an average an employee spends 13 minutes of his daily working time in reading, deleting or replying to spam messages[18].

1.1.4 Spam Produces Carbon Dioxide

In [2] report it has been found that 62 trillion spam e-mails are sent over the internet every year. This results in the emission of more than 17 million tons of carbon dioxide (CO2). It has been found that CO2 related to spam amounts to 22% of 131 kg, which is the total CO2 generated by an average business user. The report says that spam filtering would result in the reduction of spam by 75% which is equivalent to taking 2.3 millions of cars off the road. Report is based on the extra energy use spent dealing with spam.

1.1.5 Losing a solicited mail

Some of the mail servers provide limited space for email in the inbox. In such case if the quota may get exceeded on the daily or weekly basis resulting in the solicited mail getting rejected by the spam filter and ending up in the spam folder. This scenario may prove to be very expensive where the cost of

losing a solicited email is significantly high.

1.2 Research Problem

This thesis explores different aspects of spam filters, describing how the performance of a spam filter can be analyzed. In addition to this the thesis will find if the spam filter being used is optimal from particular user point of view and whether the filter provides positive or negative utility to the user.

It is important to understand, analyze and measure the effectiveness and efficiency of the spam filters in order to improve their quality so that problems like mentioned in section 1.1 may be avoided or at some extent reduced. In the context of spam filters, *"effectiveness"* means the degree to which genuine spam is detected and removed. On the other hand, *"efficiency"* means the degree to which genuine email messages are correctly delivered. A filter that removes most spam messages will have high effectiveness, but if it removes many genuine email messages together with spam messages it will have poor efficiency.

1.3 Methodology of the research

The methodology of the research in this thesis is based on Signal Detection Theory (SDT). Spam filters are investigated on the basis of SDT.

SDT [12, 4] is a model that is suitable for analyzing the effectiveness and efficiency of the spam filters and finding their optimality. SDT provides a rational basis for decision making under conditions of uncertainty. For example, the question "Is this my dog barking, or is it just the television?" is a typical situation where SDT can be applied to guide the dog owner to the most optimal action, i.e. to ignore the sound, or to go and look after the dog. Visualization used in SDT makes the decision making even simpler in situations of uncertainty.

A survey has been conducted among students to get the data e-mail data. This data has been used to calculate the tuning of the spam filter and utility provided by it.

1.4 Organization of the report

- Chapter 2- Background: This chapter starts with the definition of spam and ham and then describes about the electronic mail system, spamming statistics, techniques of spamming, measure against spamming and signal detection theory.
- Chapter 3- Related work: This chapter describes about the previous work done in order to analyze the spam filters.
- Chapter 4- Investigation of spam filters: This chapter analyzes the effectiveness and efficiency of spam filters (Yahoo mail, Gmail, Hotmail, MS Outlook) using signal detection theory.
- Chapter 5- Discussion and Comparison among Spam filters: This chapter comparison of the spam filters has been done on the basis of the results obtained in chapter 4. It also deals with the discussion based on the analysis of the spam filters.
- Chapter 6- Conclusion and Future work: This chapter concludes this thesis report with along with the description of the future work.

Chapter 2

Background

This chapter will cover the literature behind this thesis. It will explain the working of the internet mailing system and loophole which is the root cause for spamming along with the needed terminologies to properly understand the topics covered. This chapter will also cover the methods adopted by the spammers for spamming and countermeasures against spamming. In addition to this, it also covers the literature about the Signal Detection Theory (SDT), which is used to analyze the spam filters.

2.1 Definition: Spam and Ham

The word spam has been derived from a popular sketch of Monty Python [10]. Spam and Ham (non-spam/genuine mail) has been defined in many ways but the shortest, simple and convincing definition for each of them is as follows:

1. Spam: Unsolicited email sent indiscriminately in bulk.
2. Ham: Genuine email or email which is not a spam.

2.2 Electronic Mail System (e-mail)

Email is a method of receiving electronic messages over the internet. This exchange of messages is done with the help of Simple Mail Transfer Protocol (SMTP). The first SMTP was published in 1982 as an internet standard 10 (RFC 2821)[3].

2.2.1 Creation of an e-mail

An e-mail is composed using Mail User Agent (MUA). An e-mail has two main sections:

- Header: It is composed of different fields such as sender, receiver, Carbon copy, Blind Carbon Copy, Date and subject.
- Body: It is the actual unstructured text message.

Below is a sample e-mail.

```
from :<mmMicha@gmail.com>
to :ppPeter@hotmail.com
cc :ssSmith@gmail.com
bcc :rrRoshan@gmail.com
date :Sat, May 9, 2009 at 5:54 PM
subject :Example
mailed-by :gmail.com

This is a sample E-mail.  (BODY)
```

In the above example of the e-mail, in the header, `from` field shows the e-mail address of the sender of the message, `to` field shows the e-mail address of the person to whom the message is sent, `cc` stands for *Carbon Copy*, this field shows the e-mail address of those person who receive the copy of the e-mail apart from the main recipient i.e. the recipient mentioned in the `to` field. The field `bcc` stands for *Blind Carbon Copy*, it shows the e-mail address of the third type of recipient of the e-mail. In this case no other recipient is aware that bcc'd recipient had also received the copy of the e-mail. The `date` field shows the when the e-mail has been sent. The `subject` and `mailed-by` fields shows about what the message is and which server is involved in sending the message, respectively.

2.2.2 Transmission of an e-mail

When the sender presses then send button after composing the header and the body of the e-mail using MUA, the e-mail client on the sender's machine connects to the e-mail server (SMTP server) at the sender's side using port 25. After the connection the sender client interacts with the SMTP server and sends the receiver's and sender's address along with the body of the

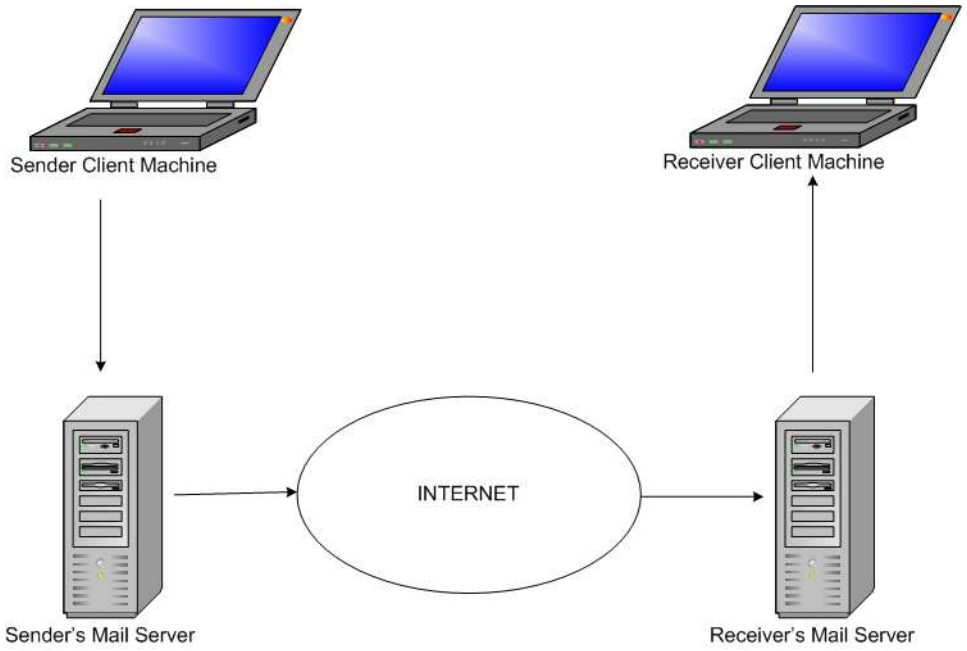

Figure 2.1: Generic architecture of e-mail transmission

message to the SMTP server. Fig.(2.1) shows the general architecture of the transmission of the e-mail over the internet.

SMTP server at this stage takes the receiver's address and breaks it into two parts- the receiver's name `ppPeter` and the domain name `hotmail.com`(refer example e-mail). If the receiver's address had been at `gmail.com` then the sender's SMTP server would have simply handed over the e-mail to POP3 (Post Office Protocol version 3) for `gmail.com` using Mail Delivery Agent (MDA) program (MDA is a software that delivers an e-mail just after the e-mail has been accepted by the server) but the receiver, in our example, is at different address (`hotmail.com`) therefore, SMTP server first communicates with that domain then transfers the e-mail.

The SMTP server converses with the Domain Name System and asks for the IP address of the SMTP server for `hotmail.com`. The DNS replies with IP address(es). After getting the IP address the sender SMTP server connects with the receiver's SMTP server and transfers the e-mail. The hotmail server after receiving the e-mail sends the e-mail to hotmail's POP3 server which ultimately puts the e-mail in the receiver's mail box.

Usually the header of an e-mail indicates the address of the sender and the receiver. Therefore, an e-mail can be tracked back to it's root i.e. from where

it originated. In case of fake header it becomes difficult to track the e-mail.

2.2.3 SMTP: Cause of Spamming

Spammers don't want to reveal their identity as well as the address from where the spam originated [17]. The main reason behind the origin of the spam is the improper design of the SMTP protocol. SMTP protocol was developed when the internet was quite new and was not so widespread, therefore, spamming was not a problem at that time. So, these may be the reasons of not implementing any proper anti spam method in SMTP protocol. Though, theoretically it is quite easy to change the SMTP protocol to deal with spams but practically it is very difficult. The reason for this difficulty is the millions of users who are using this protocol daily and this change cannot happen in very short period of time. Therefore, many solutions of anti spam strategies have been proposed which could work with the SMTP protocol and not within it.

In addition to this, the other problem with the SMTP protocol is still a system based on trust. Anyone submitting the message can claim to be anyone else with little or no accountability and there is no way to track back the original sender of the message [20].

2.3 Spam: Past and Present

The first spam was sent by Gary Thuerk in 1978 over ARPANET. He sent a message advertising new model of DEC computers to 396 people out of around 2600 people who were on the ARPANET at that time[22].

The first major commercial spamming was done in 1994, by two lawyers Laurence Canter and Martha Siegel. By using Usenet posting they advertised for immigration law services. The major explosion of spam happened between 2002 to 2004. Spammers in order to improve their financial fingerprints sent lot of spam. So much so that by 2004 the level of spam increased to more than 90%, as shown in Fig.(2.2) and after slight decrease it again went up to 97% in 2009.

After major rise in the number of spam messages various anti spam laws were formed but in 2003 US enacted CAN-SPAM law [16]. Under this law the first successful suit was in June 2007 against Jeffrey A. Kilbride and he was sentenced to 6 years of prison. In 2004, MY DOOM virus was formed which is a mass mailing trojan that gave birth to spam sending botnets.

Situation became so worse that 90% of spam today is sent by these bots which are in millions all over the internet.

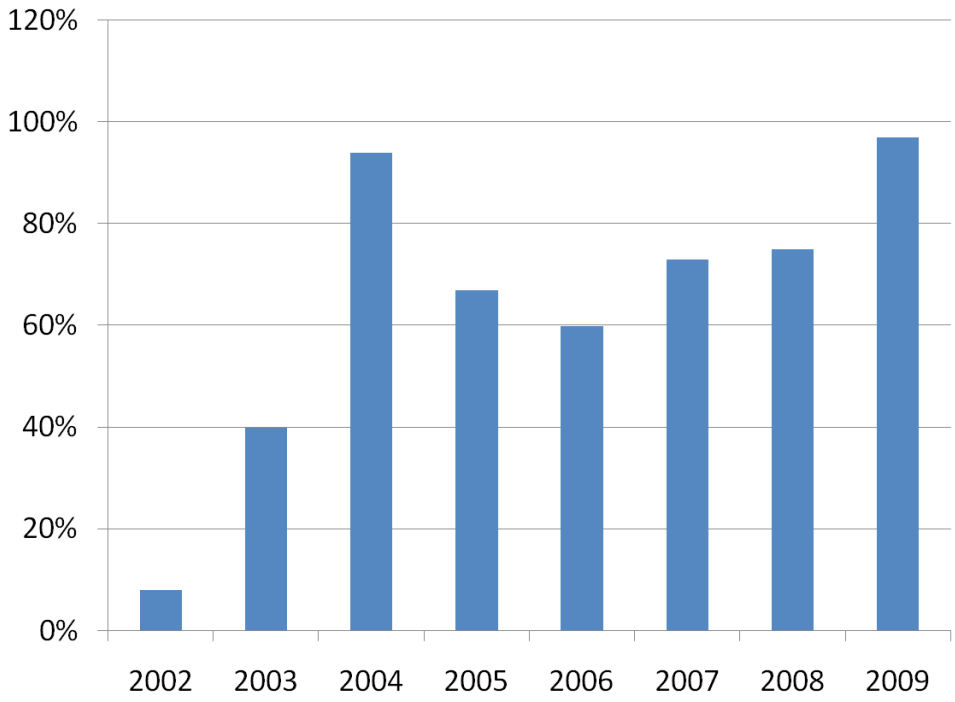

Figure 2.2: Spam Level from 2002 to 2009

2.4 Spamming

The process of spamming involves many sophisticated steps. Each and every step is really important from the spammer's point of view in order to deliver the spam in user's mail box, eventually, because at some extent spam filters are also becoming smart to distinguish between a spam and a ham. In the following sub-sections describe the process of spamming in detail.

2.4.1 Process of Spamming

Spamming activity has basically three phases. These are as follows:

- E-mail Harvesting
- Creation of Spam
- Sending of Spam

Each phase in itself is very challenging for spammers because internet security experts are working hard to fail each and every attempt of spamming. It has become a kind of war between spammers and internet security experts. Every time security experts come up with new barrier to stop spam, spammers come up with new ideas and strategies to bypass that barrier.

2.4.1.1 Obtaining e-mail ID's

In order to send spam to millions of users spammers needs millions of e-mail addresses. Spammers can get e-mail addresses by *renting*,*buying* and by *harvesting* them. The reason behind getting e-mail id's by these companies and spammers is the common intentional or unintentional mistakes made by internet users. Some of these mistakes are - posting on Usenet with e-mail id, posting on public forums (discussion groups), subscribing to a website that goes out of business and selling out the e-mail list of its members, responding to an opt-out link or e-mail and having an easy guessable e-mail id.

2.4.1.1.1 Renting: The list of the e-mail id's can be rented from the company managing it. In the process of renting, e-mail ids are not actually provided to the spammer instead the company, at a small charge per e-mail, sends spams on behalf of the spammer. Renting an e-mail list is typically cheaper than buying but if e-mails are needed repeatedly then renting may be prove to be expensive. some of the sites which rent e-mail lists are.

- http://www.postmasterdirect.com
- http://www.horizon-place.com
- http://www.meesels.com
- http://direc-tel.com
- http://www.optininc.com

2.4.1.1.2 Buying: Purchasing e-mail list(s) is better than renting if there is repeated use of it. The prices of lists vary from company to company depending on the quality of lists. For Ex. if the list if quite old then it will have low delivery rate and hence, the cost of the list will be low. Those companies which sell classified e-mail lists such as based on business type and geographical location, charge more because list may offer more delivery rates.

2.4.1.1.3 Harvesting: E-mail harvesting is a process of scanning e-mail ids over the internet using an application. These e-mail harvesters are automated tools which analyzes the internet data to find certain patterns which match the pattern of an e-mail id. To find an e-mail id the application may scan `HTML` source for different tags like `mal From:` and `mail To:`. Search engines can also be used by e-mail harvesters to return specified pages which harvesters can scan for e-mail ids.

2.4.1.2 Creation of Spam

The spam is composed in a way to catch the attention of the user and compel him to respond to the e-mail (spam). Since, these days users have become cautious of spam and accustomed to delete the spam as soon as they see it. Therefore, it is challenge for spammers to compose a spam message which could lure the user to open it or to visit specified site.

Before reaching to the user mailbox spam has to deceive the filter. Spam should be composed in such a way so that filter should classify it as a ham. Some of the techniques of creation of spam message which spammers use are described as follows:

2.4.1.2.1 Blank HTML: Blank `HTML` e-mail messages are the messages which do not contain any plain text. The message contains an image which is very hard for a spam filter to parse because significant amount of artificial intelligence would be required to parse such an image. Example of blank HTML has been shown in Fig.(2.3)

```
<html><img src="http://www.Myhi-five-value.com/dre/favrt.gif"><div>
<ahref="http://www.Myhi-five-value.com/dre/qd.php?x=52c">
<img src="http://www.Myhi-five-value.com/dre/high.gif"></a></html>
```

Figure 2.3: Ex. of Invisibility- using blank HTML

2.4.1.2.2 Invisible Text: Invisible text is a textual content which a user can not see but the spam filter can easily read. Techniques involved in making the text invisible attempt to hide valid text inside a message to make it appear a valid message. The reason behind hiding the text by the spammers is that many the spam filters calculate percentage and make the decision, whether an incoming e-mail is spam or ham, on the basis of number of spam words rather than ham words. Therefore inclusion of such texts, having random words, would offset the percentage to the level where the spam filter considers any incoming e-mail to be of an acceptable type for the delivery.

The techniques to hide the text, include the inclusion of real random numbers or text or both before `HTML` begins as shown in the Fig.(2.4).

```
Real words suspension carelessly obfuscation maintanance<html>
```

Figure 2.4: Ex. of Invisibility- using data before HTML

Words can also be secretly included which makes the spam look like ham by writing white text on a white background, as shown in the Fig.(2.5).

```
<font color=.white.>Real words suspension carelessly
obfuscation maintanance</font>
```

Figure 2.5: Ex. of Invisibility- using white text on white background

Another method to hide the text is by using the header fields as shown in Fig.(2.6).

```
X-Mime-Key: search words: repositoryobfuscation
mercahandisoryformation nonsenseaction
```

Figure 2.6: Ex. of Invisibility- using header

2.4.1.2.3 Splitting Words: Many spam filters use corpus of words which help in classifying a message as spam or ham. Empty `HTML` tags can be used with split words so that the spam could not detect it as single word, which is actually a single word but it can be detected by a human eye. Therefore, in order to be effective spam filters should be knowledgable enough to understand `HTML` very well. An example of split words with HTML tags is shown in Fig.(2.7).

```
<HTML><BODY>explan<! xe64>ation</BODY></HTML>
```

Figure 2.7: Split words with HTML

2.4.1.2.4 Bogus HTML Tags: Insertion of bogus `HTML` tags is quite effective way to accomplish the purpose of spammers because some spam filters may not be able to parse the message due to large amount of text that is not properly formatted. Fig.(2.8) describes an example of inserting invalid `HTML` tags with large amount of data. The main objective behind such insertion as described in the paragraph of *invisible text*, is to hamper the filter's ability to distinguish between spam and ham.

```
<Unsolicited bulk email, commonly known as spam, represents a
significant problem on the Internet. The seriousness of the
situationis reflected by the fact that approximately 97\% of
the total e-mail traffic currently (2009) isspam. For example,
one missed call means NO, two missed calls means YES etc.
Missed calls consume approximately the same amount resources in
the signalling channel as the normal voice calls. The traffic
generated in the signalling channel by the missed calls is
huge, and during rush hours it is very hard to make a voice
call because of network congestion. To fight this problem,
various anti-spam methods have been proposed and are
implemented to filter out spam before it gets delivered to
recipients,but none of these methods are entirely satisfactory.
.>
```

Figure 2.8: Invalid HTML tags with large amount of text

2.4.1.2.5 Vertical Hiding: A Spammer can hide the words by using `HTML` table. According to this technique words a printed vertically in the table instead of horizontally, as shown in the Fig.(2.9). For the user the output will meaningful but for the spam filter it will only be fragments of words. In the figure the output (bottom part) shows how the message would be displayed to the user but each strip(`HFT, EIH,..`) shown in the output is placed in the table as shown in the upper part of the figure.

2.4.1.2.6 MIME Partition: A `MIME` document is separated in two parts, one `HTML` part and the other plain text. Spammers exploit this functionality by placing an invalid text in the section of the plain text, which is generally never displayed and placing a spam message in the `HTML` section. The spam

```
<table cellpadding=0 cellspacing=0 border=0><tr> <td>
<table cellspacing=0 cellpadding=0 border=0><tr><td><font
face="Courier New, Courier, mono" size=2> H<br>F<br>T
</font></td></tr></table></td>

<td><table cellspacing=0 cellpadding=0 border=0><tr><td><font
face="Courier New, Courier, mono" size=2> E<br>I<br>H
</font></td></tr></table></td>
...
...
```

H E L L O
F I N E
T H A N K Y O U

Figure 2.9: Vertical hiding of the text

filters generally parse the message as a single message and therefore, if the invalid message succeeds in having more likelihood value of being a ham than the message in the `HTML` section have of being a spam then the whole message would deceive the filter and pass through it. Fig.(2.10) shows the exploitation of partition of `MIME` document.

```
------=_NextPart_001_1A3FC_05N30G82.4576190
Content-Type: text/plain;
<Unsolicited bulk email, commonly known as spam, represents a
significant problem on the Internet. The seriousness of the
situationis reflected by the fact that approximately 97\% of
the total email traffic currently (2009) isspam. To fight this
problem, various anti-spam methods have been proposed and are
implemented to filter out spam before it gets delivered to
recipients,but none of these methods are entirely satisfactory
.>
------=_NextPart_001_1A3FC_05N30G82.4576190
Content-Type: text/html;
<p><b><font face=Arial>Buy Viagra a low price and get back your
sexy life in just one week.</font></b></td>
```

Figure 2.10: MIME document partition exploitation

2.4.1.2.7 Character and Space Tricks: By placing spaces in between the characters spammer can fool spam filters. For example in Fig.(2.11) the spam filter would read the word `M O N E Y` as `M<space>O<space>N<space>E<space>Y`. Even if the spaces are replaced by any other characters, as shown

in the second line of same figure, the spam filter would still be not able to parse it and would allow the message to pass through it like a ham.

M O N E Y

M*O*N*E*Y

Figure 2.11: Example of character and space tricks

2.4.1.2.8 URL Hiding: Spammers use several techniques in order to hide the URL as shown in the Fig.(2.12). First line in the figure shows 32 bit encoding of URL, second line shows hexa decimal encoding of URL, third line shows octal encoding of URL and the fourth line shows the URL as a combination of password and IP address before and after @ and HTML page name infinite.htm. The advantage of hiding of the URL is to avoid the matching with the URL's that are present in the database of the spam filter.

```
http://7634629437/infinite.htm
http://0xC70F333D/infinite.htm
http://0707.0036.0314.075/infinite.htm
http://3334750091@3334750091/o%51a%45ae%32%5b%76f%9c
```

Figure 2.12: Example of hidden URL's

2.4.1.2.9 JavaScript: Since many spam filters do not have the functionality of JavaScript parser therefore, filter ignores the JavaScript and allows the message to pass through it. This loophole is exploited the spammers by placing the entire spam message inside the JavaScript. Therefore, in order to avoid such spamming proper decoding of JavaScript is needed.

2.4.1.3 Sending of Spam

After getting the list of e-mail addresses and having the spam message composed, the spammer sends the message to the collected addresses, using one of the many mass mailer tools. In the process of spam sending the spammer avoids getting tracked back because spam sending violates the terms of service of internet service providers (ISP's) and therefore complaints of spam

sending generally results in the termination of the account of the sender. The point of origin of the spam message is concealed using *Open Relays* or *Open Proxies*.

2.4.1.3.1 Open Relays: Open relays are SMTP servers over the internet which are designed in such a way that they transfer an e-mail to and from anyone and not just an e-mail destined to and from the users in the server database. For example, normally server for A.com would accept an e-mail only from addresses at A.com but Open relays would also accept an e-mail for B.com and then would contact the B.com server to deliver the mail. Open relays exist for many reasons, some users use it because of the firewalls.

Open relays had been abused a lot by spammers in the past but now it has become less common now. Many ISP's use DNS based blocking lists to not allow the mails from Open relays. If any mail server is detected of allowing e-mails to pass through them on the behalf of some third party then that mail server would be added to the blocking list and in future would get rejected, by the severs using that blocking list, for sending any e-mails. Open relay technique for spamming has come to an extinction therefore spammers have adopted other techniques of spamming like botnets. Botnet is the collection of infected computers which work autonomously and automatically which the spammers use to send spam.

2.4.1.3.2 Open Proxies: Proxy servers are the servers which are designed in such a way that they bypass firewalls. Proxy servers are designed for those users who are behind the firewalls. The misconfigured proxy servers can be abused by spammers with the help of the command, `HTTP CONNECT`. Unlike Open relays in proxy server it is quite impossible to find out the correct origin of the e-mail. Therefore, proxy servers are preferred by the spammers. Open proxies are also created using viruses which then spammers abuse by sending spam. The open proxies created by the viruses are very hard to detect.

2.4.2 Measures against spamming

There are two different ways to stop spam.

- Non Filtering Techniques: These techniques try to stop spam by preventing bulk e-mailers. For example, by charging for every e-mail which is sent or by restricting access to e-mail servers for spammers.

- Filtering Techniques: These techniques are used after the spam messages are sent by the spammer. Filtering techniques detect and separates spam from ham before e-mail gets delivered to the user.

2.4.2.1 Non filtering techniques

Non filtering techniques can be categorized into three parts:(1) Prevention System (2) Time based System and (3) Money based System.

2.4.2.1.1 Prevention System: Spam prevention is the direct way to stop spam. This can be done by closing all open relays over the internet and by strengthening `SMTP` protocol that would force the sender to go through the authentication process to track the origin of the spam. This forces bulk e-mailers to send spam through their own ISP's, but relies on these ISP's to block their accounts. This approach of stopping spam goes against the principles of the internet. Moreover this approach is not sufficient as spammers have now started using open proxies which hides the place of origin of the message. In addition to this hacked computers are also used for spamming.

2.4.2.1.2 Time based System: This is one of the economic solutions. According to the this solution the sender of the message is forced to spend some time for Ex. in solving some problem before he can send the message. This problem is moderately expensive function, called a *pricing function*[13]. The idea behind this solution is to waste the computer time in order to discourage the spammer from spamming. For a legitimate internet user it is not very expensive in terms of computer time to send an e-mail but for spammers who send millions of e-mails it would take significant amount of time to send spam. Therefore, it makes it tough for the spammer to send large amount of messages in an acceptable time.

This technique has not been yet incorporated in the internet. Even if it would be there, it is hard to tell how much will it succeed in practice. Issues related with this technique are:

- This feature has to be incorporated into the Internet which is not easy.
- There is the problem of hardware backward compatibility. A user using an old computer must be able to send an email in a reasonable amount of time. This rules out the use of too costly pricing functions. But then for a spammer using modern hardware, the cost in time to send a

message may become almost equal to zero. It seems impossible to find a pricing function that suits both needs.

2.4.2.1.3 Money based System: The significant cheapness of sending large amount of spam is the main motivation factor for spammers. Money based solution was proposed in order to discourage spammers.

Money based System this is also an economic solution. The main idea behind this solution is to charge the sender some amount of money for each e-mail. Basically this is based on channelised e-mail system where the sender has to pay to the recipient, before the recipient reads the e-mail arriving on specific channel[13]. The payment can be in the form of electronic cash to automate the process. Since, spammers send large amount of spam therefore this technique may make it unpleasant for them to send spam.

There are some issues with money based system like presently there is no global electronic cash system and the other major concern is the adoption of the system by the user (assuming the system is present and working).

2.4.2.2 Filtering Techniques

Filter based techniques against spam can be divided into two categories:

1. Cooperative Filtering: This kind of filtering would require cooperation between spammers and the recipient of spam. Cooperative Filtering would also require implementation of set of standards all over the network and adhering to those standards in order to identify spam. This kind of filtering is less likely to work because spammers try to hide the place of origin of spam.

2. Heuristic Filtering (Rule based filtering): Heuristic Filtering on the other hand works without any cooperation with spam originators and assumes that it is possible to detect and classify spam from ham.

Since the cooperative filtering is less likely to work therefore following part of this subsubsection will discuss about only heuristic filtering. Heuristic based filtering can be classified in to three categories: List based filtering, Traffic analysis based filtering and Content based filtering.

2.4.2.2.1 List Based Filtering: List based filters work on the idea of categorizing the sender of the e-mail as a spammer or a non-spammer (trusted

user) and then stop spam by blocking or allowing e-mails accordingly. List Based Filtering is also called Origin Based Filtering as e-mails according to this technique are filtered before getting to the user's computer. Following are the lists used for filtering e-mails.

Blacklist:

Blacklist is the most prominent and popular method to stop spam. It contains the list of e-mail address and IP (Internet Protocol) addresses which previously have been involved in spamming. When any e-mail arrives, the spam filter checks to find our if the IP address or the e-mail address of the incoming e-mail is in the balcklist. If the spam filter finds out match the e-mail is classified as spam and rejected.

Blacklist can also sometime misidentify a legitimate sender as spammer because blacklists can be bypassed by relaying mail through the SMTP servers of the legitimate users that are not on the blacklist. Another disadvantage is that spammers routinely switch IP addresses and e-mail addresses to hide their tracks therefore, a blacklist may not catch newest sapmming cases.

Whitelist:

Whitelist makes an attempt to stop spam using method which is just opposite to that of a blacklist. Unlike blacklist, whitelist contains the IP addresses and e-mail addresses of the users who are allowed to send the e-email and others are rejected by default. These addresses are placed on a trusted user list. In order to enable the legitimate sender to reach the recipient, the whitelist based system will send a request for confirmation to the sender and the sender is supposed to reply in specific short period of time.

The whitelist is generally used along with another filtering technique in order to reduce the number of ham that accidentally get classified as spam. If just whitelist is used by the spam filter then each and every ham sent by unknown legitimate users (not on the whitelist) will be classified as spam.

There is also an automatic way of creating a whitelist. According to this method, sender addresses is checked against the blacklist; if the sender has no history of spamming then his addresses added to the whitelist after dropping the e-mail to the intended mailbox.

Greylist:

Greylist spam filtering technique in comparison with blacklist and whitelist is newer. It takes the advantage of the fact that spammers generally attempt

to send a batch of spam only once. Greylist based system initially rejects the message from an unknown sender and sends a failure notice to the sender server. If the sender server attempts to send the message again (which is done by most legitimate servers) then the Greylist based system assumes that the message is not a spam and hence delivers the message to the recipient's inbox. In addition to this, the system will add the e-mail address or the IP address of the sender to the to the Greylist.

One of the disadvantage of Greylist filters is that they may delay the delivery of the e-mail which can be sometimes inconvenient when any particular e-mail is expected urgently.

Real-Time Blackhole List:

The Real-Time Blackhole List technique works in quite similar manner as blacklist but requires less hands-on maintenance. The reason behind this is the maintenance of most of the real-time blackhole lists third parties. These third parties build blacklists on the behalf of their subscribers. According to this technique each time the spam filter receives an e-mail it connects to the third party system and then compares the sender's address against the Real-Time Blackhole List.

Blackhole lists are large and updated regularly therefore, there is no need to spend time manually including new IP addresses in the list, to increase the probability of the spamfilter to catch the newest spam scam. The disadvantage of real-time blackhole lists is that like blacklist it may also classify ham as spam if spammers happen to use a legitimate IP address as a similar passage for spam.

2.4.2.2.2 Content Based Filtering: Content based filtering technique is used after the full reception of the message (including the body of the message). Some of the Content based filtering techniques are mentioned below.

Key word based filtering:

Key word based spam filters are the simplest type of content based filters. These filters reject e-mails that contain certain words. The idea behind this technique is that most spammers do not use words that are used in personal or business communication. Hence, it can be used to fight spam, inspite of being the simplest.

But the disadvantage with this technique is that if the spam filter is configured to detect e-mails with more common words then this may classify

ham as spam. Key word based filters should be updated regularly because spammers quite often misspell the key words in order to fool the spam filter and pass through it.

Score based filtering:

Score based filters are more advance than Keyword based filters because instead of blocking e-mails that have suspicious words score based filters take into account multiple words in an e-mail. Score based filters scans an incoming e-mail and assigns a specific score (points) to words and phrases. Words that are found quite frequently in spam messages like 'Viagra' 'free credit' would receive higher scores than those words which are found in ham messages. The total score is calculated by adding up all the points. If the e-mail receives certain score or higher (determined by the anti-spam application's administrator), the e-mail is classified as spam and e-mails that receive low score than the target score are delivered to the users inboxes.

Score based filters are quite effective and also minimize delay but may also result in classifying spam as ham if filter finds certain combination of words in an e-mail sent by legitimate user. In addition to this, spammers may also learn to avoid certain words thereby deceiving the spam spam filters.

Naïve Bayesian filtering:

Bayesian filtering technique is the most advanced form of content-based filtering. It uses the laws of mathematical probability to classify spam from ham. Before the Bayesian filter starts functioning, they are trained with a set of spam and a set of ham, by manually flagging each message as either spam or ham. The filter makes two list one for ham and another for spam. When e-mails are received by the filter it scans e-mails (ham+spam) for words and phrases and adds them to the respective lists.

In order to check whether an e-mail is spam, the Bayesian filter scans the e-mail and looks for certain words and phrases and then compares them against the list for spam and the list for ham to find out the probability that the message is spam. For example, if the e-mail contains the word "Viagra" and it appears 50 times in spam list but it only appears 5 times in ham list, then there is 91% chance that the incoming e-mail is a spam.

Bayesian filter regularly builds its lists on the basis of e-mails received by the user therefore, filter becomes more effective the longer it's used.

2.5 Signal Detection Theory

This section presents a model for analyzing spam filters based on SDT (Signal Detection Theory)[12, 4, 11, 23]. SDT is based on probability theory and is an effective means to analyze ambiguous data. In the SDT framework each event is assumed to be either:

- signal (from a known process) or
- noise (from an unknown process)

SDT provides a formal framework for setting optimal thresholds for distinguishing between signal and noise. For example, in radar system the operator tries to determine from the display on the radar screen whether it is a signal (aircraft) or a noise (bird or something else), and setting the optimal decisioin threshold is importance for the success of military operations.

SDT assumes that signal and noise distributions overlap each other and that an observed stimulus may come from any side of the distribution. In addition to this SDT also assumes that the signal is added to the noise and that the decision maker behaves rationally and tries to find out the optimal performance.

Fig.(2.13) shows the SDT model with the two distributions (signal and noise) assuming that both distributions are normal with equal standard deviations. The X-axis / horizontal axis represents the strength of the internal response (also called hidden variable, decision variable or internal variable) which is a function of the external observed stimulus. The internal response gives the information about the event. The Y-axis / vertical axis represents the probability of the internal response. These distributions are used in the process of making the decision whether the stimulus represents signal or noise. The vertical line between the two distributions is the decision criterion for the internal response that is used to make a decision. The decision criterion is fixed and is defined on the basis of the hidden variables.

In the process of decision making any internal response with a value less than the value of the decision criterion is determined to come from the noise distribution while an internal response with a value greater than the value of the decision criterion is determined to come from the signal distribution.

The overlap between noise and signal distributions results in four possible decisions as shown in Fig.(2.14).

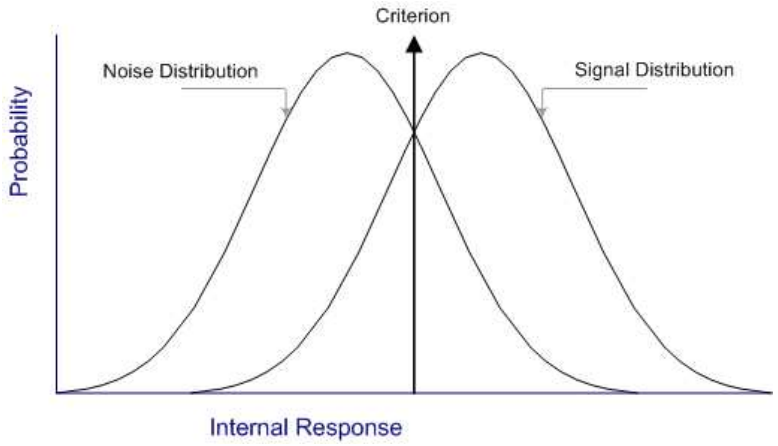

Figure 2.13: SDT model showing overlap between signal and noise distribution

- False Negative (FN): Stimulus coming from the signal distribution incorrectly detected as noise[1].
- True Positive (TP): Stimulus coming from the signal distribution correctly detected as signal[2].
- False Positive (FP): Stimulus coming from the noise distribution incorrectly detected as signal[3].
- True Negative (TN): Stimulus coming from the noise distribution correctly detected as noise [4].

FP and FN are also known as Type I error and Type II errors respectively in statistics. The SDT decision making method is based on the concepts of TP Rate and FP Rate. The TP Rate is the total number of times a genuine signal is detected as signal divided by the total number of genuine signals. Hence, it can be calculated as follows:

$$\text{TP Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.1}$$

[1]Called *"Miss"* in SDT terminology.
[2]Called *"Hit"* in SDT terminology.
[3]Called *"False Alarm"* or *"FA"* in SDT terminology.
[4]Called *"Correct Identification"* or *"CI"* or *"Correct Rejection"* or *"CR"*in SDT terminology.

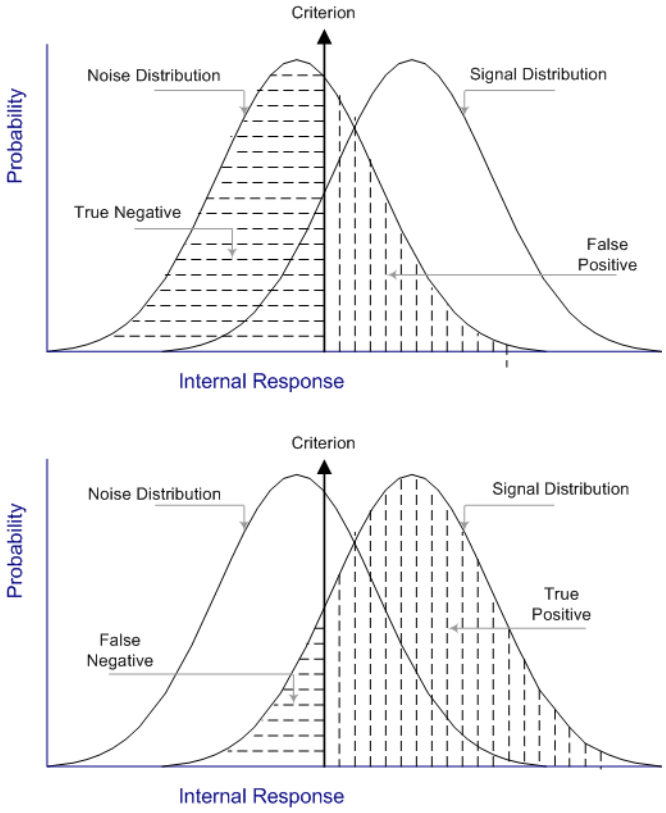

Figure 2.14: The model of SDT showing TP,FN,FP and TN

The FP Rate is the total number of times genuine noise is detected as signal, divided by the total number of genuine noise instances. Hence the FP Rate can be calculated using the following formula:

$$\text{FP Rate} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2.2}$$

It can be noted that the sum of the TP and FN Rates, as well as the sum of the FP and TN Rates both are equal to 1. This can be expressed as:

$$\begin{cases} \text{FN Rate} = 1 - \text{TP Rate} \\ \\ \text{TN Rate} = 1 - \text{FP Rate} \end{cases} \tag{2.3}$$

Fig.(2.15) illustrates the analysis of TP and FP rates. The lower half of figure sets the decision criterion at the left-most edge of the signal distribution.

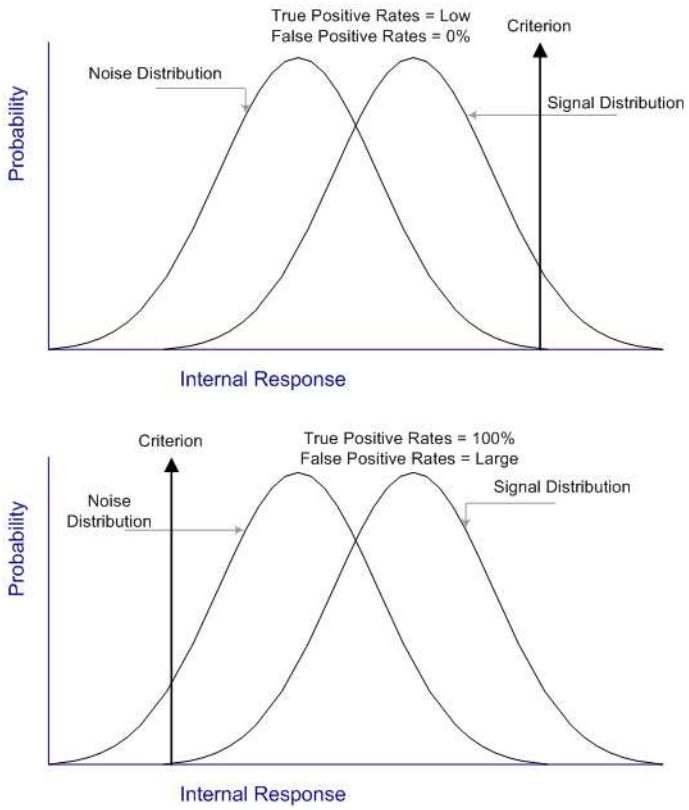

Figure 2.15: SDT model showing showing criterion at two different places: FP Rates=0% and TP Rates=100%

Statistically, it means that the TP Rate is 100%.

Let us assume the example of a doctor who makes the decision whether there is a tumor in the brain based on the internal response of a brain scan. If the value of the decision criterion is lowered such that the TP Rate is 100% then the FP Rate also increases as shown in the lower half of Fig.(2.15). The doctor will therefore never miss a real tumor, but a negative side-effect of increasing TP Rate is a corresponding increase in the FP rate. In case value of the decision criterion is increased to the rightmost edge of the noise distribution as shown in the upper half of Fig.(2.15) then the FP Rate becomes 0%, but at the same time the TP Rate also gets very low. This means that the doctor gets no false alarms, but will miss many real tumors.

SDT assumes that it is practically impossible to simultaneously have a 100% TP Rate and 0% FP Rate because of the overlap between the signal and the noise distributions. STD offers a method for defining the decision criterion

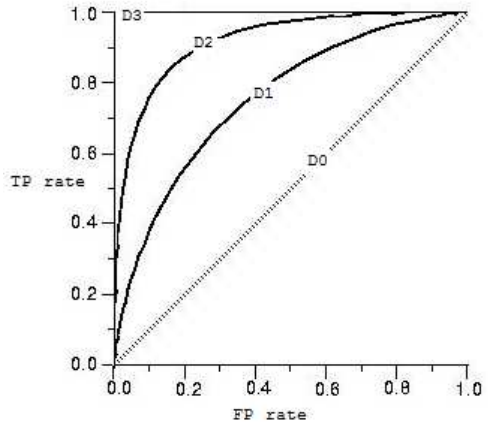

Figure 2.16: Showing ROC curves

value which will result in optimal decision making. In this paper we use STD and Bayesian methods for analyzing spam filters.

2.5.1 ROC: Receiver Operating Characteristics

After the decision have been made by the decision maker, four types of results are obtained as described earlier in this chapter. Receiving Operating Characteristics or just ROC curve [21] can be used to all the four types of results.

ROC is a graphical plot of TP rate Vs. FP rate as shown in the Fig(2.16). ROC curve changes as the value of the decision criterion is varied. It shows the comparison of two operating characteristics: TP rate and FP rate.

In the Fig(2.16), D0, D1, D2 and D3 shows the distance that is the amount of overlap between the two distributions (signal and noise). For each of the distance it shows that as the value of the decision criterion decreases or increases the rate of FP and TP changes accordingly. It can also be noticed that for reasonable value of decision criterion the TP rate is always higher than the FP rate.

The shape of the ROC curve depends on the the noise and signal distributions. The more overlap between the distributions, the more the shape of the ROC curve will be a straight line at 45 deegrees angle. The more distinct the distributions, the more the ROC curve will change angle. A specific point on the curve called the likelihood ratio (LR) depends on a certain decision

criterion. The LR has been explained in the following subsection.

2.5.2 Likelihood Ratio

The likelihood ratio (LR) is the ratio of TP rate and the FP rate. LR is calculated using the following formula:

$$\text{LR} = \frac{\text{TP Rate}}{\text{FP Rate}} \tag{2.4}$$

LR in the ROC curve represents one of the points on the curve. Likelihood ratio is very important in signal detection theory as it has many things to offer for Ex. it gives a general and principled basis for the process of decision making. It suggests what the observer may be doing in making a judgment and the most important characteristic is that the LR makes the optimal use of the information.

Signal Detection Theory says that in order to find the optimal value of the decision criterion for a particular user i.e. in order to maximize the utility for the particular user, the following equation should be satisfied:

$$\text{LR} = \frac{P(\text{noise})}{P(\text{signal})} \cdot \frac{\text{Benefit of TN} + \text{Cost of FP}}{\text{Benefit of TP} + \text{Cost of FN}} \tag{2.5}$$

The left side of the Eq.(2.5) dependent on the base rate probabilities of the stimulus being signal or noise, and also on the costs of incorrect and the benefits of correct detection and it is calculated by multiplying the ratio of the base rate probability of noise $P(\text{noise})$ and the base rate probability of signal $P(\text{signal})$ with the ratio of the cost of error and benefit of correct identification. Note that for every stimulus, the equation $P(\text{noise}) + P(\text{signal}) = 1$ holds.

Chapter 3

Related Work

In the context of spam filtering, genuine (non-spam) email messages are commonly called *"ham"*. Since spam filters are trying to identify spam, a message identified as spam is called a *"positive"*. A ham message incorrectly classified as spam therefore represents an instance of false positive (FP), and a spam message identified as ham represents a false negative (FN).

Various analyzes of the performance of spam filters have been done in previous studies. The effectiveness of a spam filter is affected by the domain in which it is used. For example the cost of a lost genuine email message incorrectly detected as spam will depend on the recipient's (and sender's) business area, as well as on the recipient's (and sender's) perception, attitude and level of frustration.

Some of the methods of analyzing spam filters which have been proposed are described in the following sections.

3.1 Error Based Function

A method for analyzing spam filters was proposed by Garcia *et al.* in 2004 [9]. Garcia's analysis was restricted to open source filters, and only considered content based filters, i.e. not for example black/white lists. According to [9] both FN rate and FP rate can not be 0 at the same time therefore, intention was to rank the performance of spam filters on the basis of FN and FP rates because a good spam filter will have low FN and FP rate.

This method of analysis took into consideration FP as an error and FN as an indicator of effectiveness of the spam filter.Garcia *et al.* a proposed function

'W' for calculating a single measure of a filter's error rate as a function of its false positive and false negative rates.

$$W(FN\_rate, FP\_rate) = (FP\_rate + \epsilon)^2 (FN\_rate + \epsilon)$$

where ϵ=.01 (constant).

The main idea is to tolerate small amount of FPs for significant amount of decrease in FNs.

3.2 Precision (P) and Recall (R)

Another approach to analyzing spam filter performance is through the Precision and Recall metrics. This method was extensively used for spam filter classification in [19].

Precision is the ratio of spam messages classified as spam relative to the total number of messages classified as spam.

Recall is the ratio of spam messages classified as spam relative to the total number of spam messages. For example, if 5 out of 10 spam messages are correctly identified as spam then the Recall rate is 0.5. As long as no ham messages are classified as spam the Precision will be 1, but as soon as some ham messages are incorrectly classified as spam the Precision will fall below 1. Therefore, formally, if:

- N1=Number of spam classified as spam
- N2=Number of spam classified as ham
- N3=Number of ham classified as ham
- N4=Number of ham classified as spam

then the formula for Precision and Recall can be written as follows:

$$\mathrm{P} = \frac{\mathrm{N1}}{\mathrm{N1} + \mathrm{N4}}$$

$$\mathrm{R} = \frac{\mathrm{N1}}{\mathrm{N1} + \mathrm{N2}}$$

For spam filters, an instance of FP is normally considered more problematic than an instance of FN. Precision which reflects a filter's FP property is therefore considered to be a more important measure than Recall which reflects the filter's FN property. The Precision value therefore needs to be higher than the Recall value, but at the same time there should be a proper balance between the two values. Therefore, spam filters with higher precision value are considered good.

3.3 Weighted Accuracy

Another proposed method for measuring the effectiveness of spam filters is Weighted Accuracy which uses the accuracy and error rate as measures. Weighted accuracy 'W' of a spam filer can be calculated as:

$$W = \frac{\lambda \cdot N3 + N1}{\lambda \cdot Nh + Ns}$$

where Nh and Ns are the total number of ham and spam messages respectively.

Equal relative weight (λ) is assigned to the error types FP (False Positive) and FN (False Negative), as well as to the correct classification types. An instance of FP counts λ times an instance of FN. An instance of TN (True Negative), i.e. a correct classification of a genuine email message, counts λ times an instance of TP (true positive), i.e. a correct classification of spam. This method reflects that an instance of FP is λ times more costly than an instance of FN [6].

3.4 10-fold cross validation

Cross validation technique is a straightforward way of finding out the effectiveness of a spam filter [15].

According to this technique data set 'm' is splitted into 10 mutually exclusive parts 'm1, m2,...m10' of approximately equal size. The inducer is trained and tested on m/m_i and against m_i, 10 times respectively, with different i's (i=1, 2,..10).

At last the performance of the spam filter is calculated by taking the average of total number of tests. For 10-fold cross validation the precision 'P' and recall 'R' 3.2 can be calculated as follows:

$$\mathrm{P} = \frac{1}{n} \cdot \sum_{i=1}^{10} P_i$$

$$\mathrm{R} = \frac{1}{n} \cdot \sum_{i=1}^{10} R_i$$

where P_i is a precision for each of the 10 tests and R_i is a recall for each of the 10 tests.

Chapter 4

Investigating Spam Filters

In this chapter it has been described that how Signal detection theory can be applied to investigate the spam filters. Characteristics of a spam filters have been analyzed in detail using SDT. In addition to this, this chapter analyzes spam filters of some of the most popular webmail services like Gmail, Yahoo Mail, Hotmail and Microsoft Outlook (Exchange Server).

4.1 Spam Filter Analysis Using SDT

Spam filters are used to separate spam from ham. A spam filter carries out this separation using different techniques. For example, content based filtering [9] is done by analyzing the body of the message. Origin based filtering[9] is done by judging the source of the message. SDT can be used to analyze the spam filters based on a single technique as well as filters based on multiple technique like those used by email service providers like: Gmail, Yahoo mail and Hotmail. First single technique spam filters after that multiple technique spam filters are discussed.

4.1.1 Spam Filters Based on Single Technique

When applying SDT to spam filter analysis, we will use the terminology convention that:

- an instance of spam is considered as a signal
- an instance of ham is considered as noise

Within the SDT framework, the difficulty of distinguishing between spam and ham increases with the degree of overlap between the two distributions, as would be expected. The overlap between spam and ham distributions results in two types of incorrect and two types of correct decisions, defined as:

1. Ham classified as ham (TN)
2. Spam classified as ham (FN)
3. Spam classified as spam (TP)
4. Ham classified as spam (FP)

The 3rd and 4th outcomes are important from the SDT point of view as they are used in the mathematical expressions. In the following S denotes a genuine spam message, and S' denotes an assumed spam message. Similarly, H denotes a genuine ham message, and H' denotes an assumed ham message.

The four possible outcomes of the spam filter are shown in Fig. 4.1. $P(S'|S)$, $P(H'|S)$, $P(S'|H)$ and $P(H'|H)$ in the Fig. 4.1 represents the four conditional probabilities.

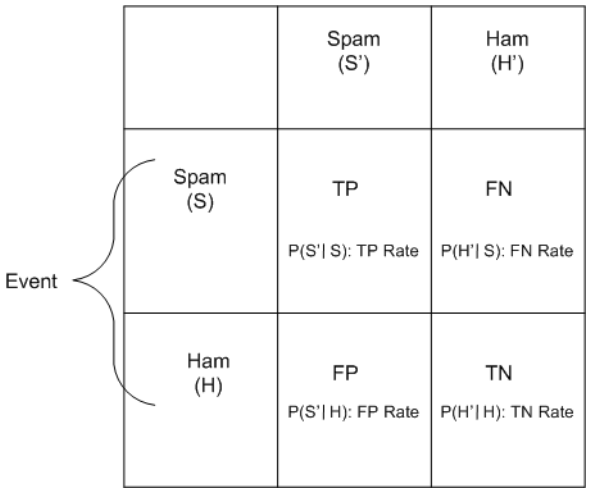

Figure 4.1: Decision Matrix for a spam filter showing four possible cases

All the four possible cases are dependent on each other. For example, when the message really is spam (1st row) the proportion of TP and FN add up to 1 because the filter can only respond in one of the two ways- either Yes

or No. Likewise when the message really is ham (2nd row), the proportion of FP and TN add up to 1. Thus all the information in the decision matrix can be obtained from TP and FP. Therefore we have

$$P(H'|S) = 1 - P(S'|S) \tag{4.1}$$

$$P(H'|H) = 1 - P(S'|H) \tag{4.2}$$

The conditional probabilities $P(S'|S)$ and $P(S'|H)$ represent the TP and FP rates respectively. The TP rate indicates the successful filtering of spam messages, and can therefore be used to analyze the effectiveness of the spam filter. The FP rate on the other hand shows errors which can be used to determine the efficiency of spam filters. Efficiency can be increased by reducing the FP rate. The effectiveness of the spam filter increases as the TP rate gets closer to 1 and the efficiency increases as the FP rate gets closer to 0.

It can be easily concluded that spam filters will behave in the best way when the TP rate is maximum and the FP rate is minimum. Practically no automated spam filter can be both 100% effective and 100% efficient at the same time. The reason for this is of course that clever composition of spam messages give them similar characteristics to ham messages. For automated filters that do not have the same cognitive and semantic capabilities as humans, separation between ham and spam is not always possible.

4.1.1.1 Actual LR and Optimal LR

After the receiving the four types of results in the inbox and spam folder it can be calculated that the output produced by the specific filter provides negative or positive utility to the particular user.

Spam filters makes use of the TP rate and the FP rate to calculate the LR (Likelihood Ratio). The formula to calculate the LR is as follows:

$$\begin{aligned} \text{LR} &= \frac{\text{TPrate}}{\text{FPrate}} \\ &= \frac{P(S'|S)}{P(S'|H)} \end{aligned} \tag{4.3}$$

We can call the LR in the Eq.(4.3) as the *Actual LR* as it has been calculated from the actual data after the filtering of the e-mails.

In order to find the utility for specific user the actual LR is compared with the value in the Eq.(4.4). We have named the value in the equation 4.4 as the *Optimal LR = LR'* because it is used to find out whether the spam

filter provides the positive utility to the user or not. If the spam filter is not optimal for the user then it is tuned for optimality.

$$\mathrm{LR}' = \frac{P(H)}{P(S)} \cdot \frac{(B_{H'|H} + C_{S'H})}{(B_{S'|S} + C_{H'|S})} \tag{4.4}$$

where $P(H)$ and $P(S)$ represent the base rate probabilities of ham and spam in the message set.

The additivity $P(H) + P(S) = 1$ always holds.

In the above equation $B_{H'|H}$ denotes the benefit associated with TN, and $B_{S'|S}$ denotes the benefit associated with TP. Similarly $C_{S'|H}$ denotes the cost associated with FP, and $C_{H'|S}$ denotes the cost associated with FN.

In the Eq.(4.4) LR'[1] has been calculated using the base rate probabilities of occurrence of spam messages in a representative set of messages and the cost associated with incorrect decisions and the benefits associated with correct decisions. The LR' varies from one user to another because the costs and benefits involved in receiving an e-mail is different for different users.

In Eq.(4.4) if the cost of errors is the same as the benefits of correct responses as shown in the Eq.(4.5)

$$(B_{H'|H} + C_{S'|H}) = (B_{S'|S} + C_{H'|S}) \tag{4.5}$$

then the LR' becomes equal to the fraction of base rate probabilities of spam and ham. This can be written mathematically as follows:

$$LR' = \frac{P(H)}{P(S)}$$

From empirical researches [19, 6, 5] it has been found that the base rate probability of spam affects the detection of spam. The base rate probability will therefore influence the decision criterion value of the filter.

The cost of FP is normally significantly higher than the cost of FN. People are normally more concerned about the loss of a ham that about receiving a spam. With the help of Eq.(4.6) different aspects of the spam filter can be evaluated and analyzed.

While comparing LR and LR' the rule for assessing the value of the spam filter is as follows:

[1]The formula has been derived taking into account + and - signs but wherever else cost and benefits will be used they will be used with appropriate signs

$$\text{LR} = \text{LR'} \tag{4.6}$$

As described in subsection 2.5.1, in the ROC curve a particular point on the curve is determine by the decision criterion, which is the actual LR.

Like actual LR, the optimal LR can also be placed on the same curve and the optimal decision criterion is said to be obtained when both the points are same as shown below:

$$\begin{aligned} \text{LR} &= \text{LR'} \\ (P(S'|S), P(S'|H)) &= (P(H) \cdot (B_{H'|H} + C_{S'|H}), P(S) \cdot (B_{S'|S} + C_{H'|S})) \end{aligned} \tag{4.7}$$

In this situation the spam filter would behave optimally for the specific user.

If the spam filer is does not works in an optimal way for the user then it should be tuned taking in to consideration certain parameters. Therefore, it can be concluded that Actual LR is a function of tuning parameters. It can be represented mathematically as follows:

$$\begin{aligned} \text{LR} &= \text{f(x)} \\ \text{where, x} &= \text{Tuning Parameters} \end{aligned} \tag{4.8}$$

The value of x in the Eq.(4.8) will change each time the spam filer is tuned with new parameters.

4.1.2 Subjective Tuning Index

Based on the concepts developed in the previous sections we will here define the *Subjective Tuning Index*, or STI for short. This index expresses the degree of optimality of the tuning of a particular spam filter when seen from a specific user's point of view. This means that the utility of having a spam filter is maximized as a function of cost and benefit of incorrect and correct filtering.

From here onwards $B_{H'|H} = U_{H'|H}, C_{S'|H} = U_{S'|H}, B_{S'|S} = U_{S'|S}, C_{H'|S} = U_{H'|S}$ because we will talk in terms of utility.

The optimal likelihood ratio and the actual likelihood ratio are determined by their respective points on the ROC curve of Fig.(2.16). A spam filter is

tuned optimally when the two points are in the same position. The closer the points, the more optimal the tuning, and the further apart, the worse the tuning. Below we specify the STI as the distance in the plane of the ROC curve. Let σ represent the STI:

Definition 1 (Subjective Spam Filter Utility Index)

$$\sigma = \frac{\sqrt{(P(S'|S) - P(H)(U_{H'|H} + U_{S'|H}))^2 + (P(S'|H) - P(S)(U_{S'|S} + U_{H'|S}))^2}}{\sqrt{2}} \tag{4.9}$$

The maximum distance between two points in Fig.(2.16) would be $\sqrt{2}$. In order to let σ take a value in the range $[0, 1]$ the normalization factor $1/\sqrt{2}$ is used in Eq.(4.9).

We can use the value of σ to analyze the tuning of a spam filter. The smaller the value, the better the spam filter is tuned. In case $\sigma = 0$, the spam filter for a given user is perfectly tuned. When $\sigma \neq 0$ it means that the spam filter is not tuned according to the needs of the user.

Whether the spam filter actually provides positive or negative utility, and how much utility is providers to the user is not directly indicated by the STI σ. The utility U is given by the expression below.

$$U = P(S) \cdot [P(S'|S) \cdot U_{S'|S} + P(H'|S) \cdot U_{H'|S}] + P(H) \cdot [P(H'H) \cdot U_{H'|H} + P(S'|H) \cdot U_{S'|H}] \tag{4.10}$$

The overall utility U will depend on the probabilities of the various outcomes and their respective utilities.

4.1.3 Spam Filters Based on Multiple Techniques

When a spam filter has more than one filtering techniques, which is generally the case, then additional considerations must be taken.

All the filtering techniques are assumed to be in sequence. In addition to this, the inherent characteristics of each filtering technique are statistically independent of each other. If the filtering techniques are not statistically independent then the sequential set of filters is assumed to consist of just one filtering technique, and this filter would be relatively less effective. A filtering technique at one point in the chain will change the base rate probabilities

for the next filtering technique in the chain. If the base rate probabilities are changed by the stimulus emanating from the 1st filtering technique, it should result in actual LR equal to that of Eq.(4.3). This new value will be denoted as LR_1.

$$LR_1 = \frac{P(S_1'|S)}{P(S_1'|H)} \tag{4.11}$$

Therefore Eq.4.6 would look like:

$$\frac{P(S_1'|S)}{P(S_1'|H)} = \frac{P(H)}{P(S)} \cdot \frac{(U_{H'|H} + U_{S'|H})}{(U_{S'|S} + U_{H'|S})} \tag{4.12}$$

The base rate probability and the actual LR changes every time an e-mail passes through the new filtering technique. LR_1 indicates the actual LR after the 1st filtering technique.

If the filter incorporates n filtering techniques then the internal structure of the spam filter would more look like as one shown in the Fig.4.2. In addition to this, with 'n' filtering techniques the Eq.4.12 would change to:

$$\prod_{i=1}^{i=n} \frac{P(S_i'|S)}{P(S_i'|H)} = \frac{P(H)}{P(S)} \cdot \frac{(U_{H'|H} + U_{S'|H})}{(U_{S'|S} + U_{H'|S})} \tag{4.13}$$

where $P(S_i'|S)$ and $P(S_i'|H)$ represent the TP and the FP rates for the i^{th} filtering technique.

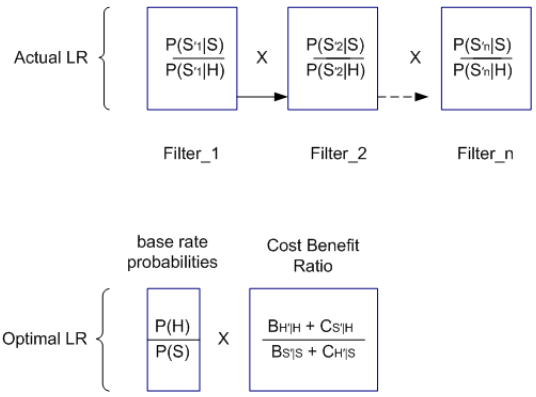

Figure 4.2: Sequential use of spam filters

4.2 Method of Analysis

This section describes about the method used to collect the data for analysis. The main objective is to analyze the different spam filters on realistic e-mails. These e-mails should reflect the fact that different types users have different priorities of receiving an e-mail.

To collect the data a survey was conducted by sending an email, shown below, to 224 people which included students and students who are also employees.

```
Hi,
This is a short survey about how many spam emails you receive.I
do this as part of my Masters research project at UNIK / Unive-
-rsity of Oslo under the supervision of Prof. Audun Josangs. By
participating you'll support research and thereby help fighting
the problem of spam. It won't take more than a minute or two to
fill in the survey.The information you provide does not have to
be totally exact. Simply answer the questions as precisely as
you can. The survey is anonymous.

The information you provide must relate to one specific e-mail
service, such as Gmail, Hotmail,Yahoomail and Microsoft Outlook
(Exchange Server). In case you use multiple e-mail services you
can fill in a survey for each one of them separately.

Please follow this link:
https://www.surveymonkey.com/s.aspx?sm=BEyJ8UZq51XyIqJntF2dmQ_3
d_3d

Your contribution really matters.
Any feedback on the survey design is also welcome.
Thank You.
```

To create a survey the service of surveymonkey.com were used. A snapshot of the survey, just the Gmail page, has been shown in the Fig.(4.3)

The e-mail was sent to 224 people in order to get the real values for Gmail, Yahoo Mail, Hotmail and MS Outlook (Exchange Server) spam filters. Fig.4.4 shows the classification of the people who replied to respective e-mail services. Each person was asked 6 questions related to spam messages as shown in the Fig.(4.3). Tables 7.1 and 7.2 in Appendix show the same questions with the respective options given to the surveyees.

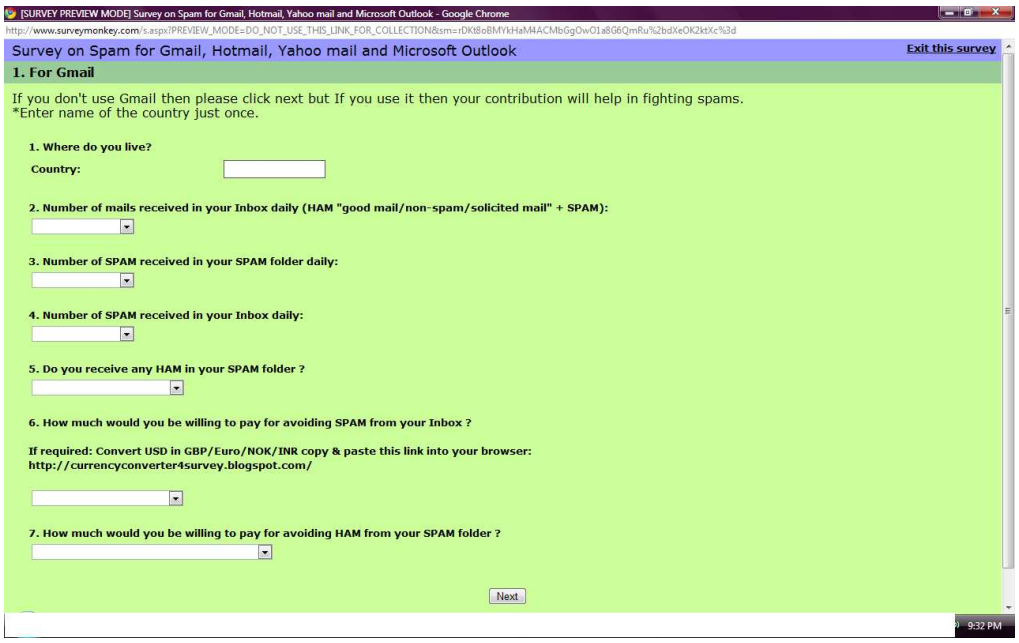

Figure 4.3: A snapshot of the survey (Gmail page)

Since Gmail, Yahoo Mail, Hotmail and MS Outlook (Exchange Server) are privately owned so, it was difficult to know if the spam filters used by these e-mail service providers are composed of single or multiple techniques. Therefore, initially I assumed them to be a single technique spam filters but later on conclusion has been made about the number of filtering techniques each of the analyzed filter may be composed of.

Since it was difficult to give the exact number as an option in the questionnaire so, a probable range was given for all the options. Therefore, for best results calculation has been done after averaging the respective data.

According to the latest data in [1] we have assumed the base rate probability of spam to be 97%.

In addition to this, the cost of a FP is assumed to be equal to the benefit of a TN and the cost of a FN is assumed to be equal to the benefit of a TP. Though these four values can also be different.

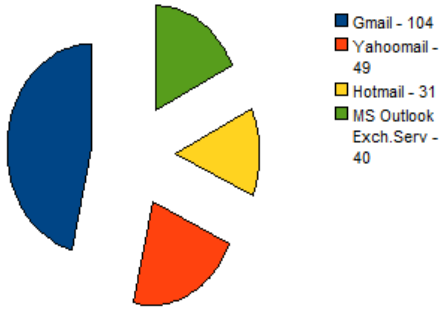

Figure 4.4: Shows numbers of people who replied to the survey

Table 4.1: Shows survey statistics obtained for Gmail. Statistics correspond to the total number of e-mails altogether received by 104 people in inboxes and spam folders in 1 day and money they are ready to pay for avoiding a ham ending up in the spam folder and a spam ending up in the inbox.

Number of mails received in your Inbox daily (HAM "good mail/non-spam/solicited mail" + SPAM).	1211
Number of SPAM received in your SPAM folder daily.	1017
Number of SPAM received in your Inbox daily.	238
Do you receive any HAM in your SPAM folder?	23
How much would you be willing to pay for avoiding that SPAM ever ends up in your Inbox ?	221 $ cents
How much would you be willing to pay for avoiding that HAM ever ends up in your SPAM folder ?	228 $ cents

4.2.1 Analysis of Gmail Filter

This section will first investigate the Gmail spam filter based on statistics obtained from the survey.

Out of 224 people 104 were Gmail users. Table4.1 shows the average number of mails collectively received in inboxes and spam folders by 104 people in 1 day and the amount of money they are ready to pay for avoiding a ham ending up in the spam folder and a spam ending up in the inbox.

As shown in the Table4.1:

$$\text{TP} = 1017 \text{ and } \text{FN} = 238$$

$$\text{FP} = 23 \text{ and } \text{TN} = 1211 - 238 = 973$$

In order to find out whether the spam filter is perfectly tuned according to the needs of the users we need actual LR (LR_{Gmail}) and optimal LR' (LR'_{Gmail}). Therefore:

$$\begin{aligned} LR_{Gmail} &= \frac{TP\_rate}{FP\_rate} = \frac{\frac{TP}{TP+FN}}{\frac{FP}{FP+TN}} \\ &= \frac{\frac{1017}{1017+238}}{\frac{23}{23+973}} \\ &= \frac{0.8103}{0.0230} \end{aligned} \tag{4.14}$$

And the Optimal LR is calculated according to the Eq.4.4 (we have assume that $B_{H'|H} = C_{S'|H}$ and $(B_{S'|S} = C_{H'|S})$:

$$\begin{aligned} LR'_{Gmail} &= \frac{P(H)}{P(S)} \cdot \frac{(U_{H'|H}+U_{S'H})}{(U_{S'|S}+U_{H'|S})} \\ &= \frac{0.00684}{0.21437} \end{aligned} \tag{4.15}$$

Both LR_{Gmail} and LR'_{Gmail} can be represented on the ROC curve as points (0.0230, 0.8103) and (0.21437, 0.00684) respectively. Distance between the two points will show the tuning of the spam filter. Therefore:

$$\begin{aligned} \sigma &= \frac{\sqrt{(0.8103-0.00684)^2+(0.0230-0.21437)^2}}{\sqrt{2}} \\ &= 0.584 \end{aligned} \tag{4.16}$$

$\sigma \neq 0$, which implies that the Gmail spam filter is not tuned according to the needs of this group of 104 students.

Utility provided by the Gmail spam filter is shown in the following calculation:

$$\begin{aligned} U &= 97 \cdot [0.8103 \cdot 221 + 0.1896 \cdot (-221)] + 3 \cdot [0.9769 \cdot 228 + 0.023 \cdot (-228)] \\ &= 13958.4135 \end{aligned} \tag{4.17}$$

Value of U=13958.4135 shows that the utility provided by the Gmail spam filter to the given users is positive and very high. Therefore this filter is good for the given users who are students in this case.

Gmail-ROC curve:

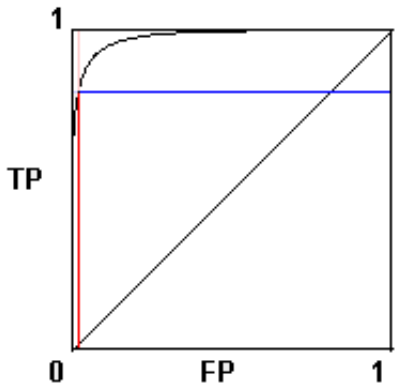

Figure 4.5: ROC curve for Gmail spam filter

Intersection of two line on the curve (Fig.4.5) shows the point LR_{Gmail} for the specific decision criterion such that TP and FP rates are 0.8103 and 0.0230 respectively.

4.2.2 Analysis of HotMail Filter

Similar to subsection (4.2.1) this section will also, with all the same assumptions, investigate the Hotmail spam filter based on statistics from the survey.

Out of 224 people 31 were Hotmail users. Table4.2 shows the average number of mails collectively received in inboxes and spam folders by 31 people in 1 day and the amount of money they are ready to pay for avoiding a ham ending up in the spam folder and a spam ending up in the inbox.

As shown in the Table4.2:

$$\text{TP} = 250 \text{ and } \text{FN} = 155$$

$$\text{FP} = 9 \text{ and } \text{TN} = 600 - 155 = 445$$

In order to analyze the spam filter we need $LR_{hotmail}$ and $LR'_{hotmail}$:

$$\begin{aligned} LR_{hotmail} &= \frac{TP\_rate}{FP\_rate} \\ &= \frac{0.6172}{0.0198} \end{aligned} \tag{4.18}$$

Table 4.2: Shows survey statistics obtained from people using Hotmail. Statistics correspond to the total number of e-mails altogether received by 31 people in inboxes and spam folders in 1 day and money they are ready to pay in 1 day for avoiding a ham ending up in the spam folder and a spam ending up in the inbox.

Number of mails received in your Inbox daily (HAM "good mail/non-spam/solicited mail" + SPAM).	600
Number of SPAM received in your SPAM folder daily.	250
Number of SPAM received in your Inbox daily.	155
Do you receive any HAM in your SPAM folder?	9
How much would you be willing to pay for avoiding that SPAM ever ends up in your Inbox ?	78 $ cents
How much would you be willing to pay for avoiding that HAM ever ends up in your SPAM folder ?	318 $ cents

And the Optimal LR is calculated according to the Eq.4.4:

$$LR'_{hotmail} = \frac{0.0954}{0.7566} \tag{4.19}$$

Both $LR_{hotmail}$ and $LR'_{hotmail}$ can be represented on the ROC curve as points $(0.0198, 0.6172)$ and $(0.7566, 0.0954)$ respectively. Distance between the two points will show the tuning of the spam filter. Therefore:

$$\begin{aligned} \sigma &= \frac{\sqrt{(0.0198-0.7566)^2+(0.6172-0.0954)^2}}{\sqrt{2}} \\ &= 0.6384 \end{aligned} \tag{4.20}$$

$\sigma \neq 0$, which implies that the Hotmail spam filter is not tuned according to the needs of this group of 31 students.

Utility provided by the Hotmail spam filter is shown in the following calculation:

$$\begin{aligned} U &= 97 \cdot [0.6172 \cdot 78 + 0.3828 \cdot (-78)] + 3 \cdot [0.9802 \cdot 318 + 0.0198 \cdot (-318)] \\ &= 2395.632 \end{aligned} \tag{4.21}$$

Value of U=2395.632 shows that the utility provided by the Hotmail spam filter to the given users is positive and high.

Hotmail-ROC curve:

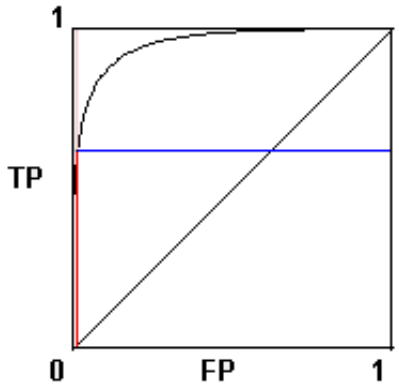

Figure 4.6: ROC curve for Hotmail spam filter

Intersection of two line on the curve (Fig.4.6) shows the point $LR_{hotmail}$ for the specific decision criterion such that TP and FP rates are 0.6172 and 0.0198 respectively.

4.2.3 Analysis of Yahoo Mail Filter

Similar to subsection (4.2.1) this section will also, with all the same assumptions, investigate the Yahoo Mail spam filter based on statistics from the survey.

Out of 224 people 49 were Yahoo Mail users. Table4.3 shows the average number of mails collectively received in inboxes and spam folders by 49 people in 1 day and the amount of money they are ready to pay for avoiding a ham ending up in the spam folder and a spam ending up in the inbox.

As shown in the Table4.3:

$$\text{TP} = 285 \text{ and } \text{FN} = 106$$

$$\text{FP} = 17 \text{ and } \text{TN} = 219 - 106 = 113$$

In order to analyze the spam filter we need LR_{YMail} and LR'_{YMail}':

$$\begin{aligned} LR_{YMail} &= \frac{TP\_rate}{FP\_rate} \\ &= \frac{0.7289}{0.1307} \end{aligned} \tag{4.22}$$

Table 4.3: Shows survey statistics obtained from people using Yahoo Mail. Statistics correspond to the total number of e-mails altogether received by 49 people in inboxes and spam folders in 1 day and money they are ready to pay for avoiding a ham ending up in the spam folder and a spam ending up in the inbox.

Number of mails received in your Inbox daily (HAM "good mail/non-spam/solicited mail" + SPAM).	219
Number of SPAM received in your SPAM folder daily.	285
Number of SPAM received in your Inbox daily.	106
Do you receive any HAM in your SPAM folder?	17
How much would you be willing to pay for avoiding that SPAM ever ends up in your Inbox ?	153 $ cents
How much would you be willing to pay for avoiding that HAM ever ends up in your SPAM folder ?	14 $ cents

And the Optimal LR is calculated according to the Eq.4.4:

$$LR'_{YMail} = \frac{0.00042}{0.14841} \tag{4.23}$$

Both LR_{YMail} and LR'_{YMail} can be represented on the ROC curve as points (0.1307, 0.7289) and (0.14841, 0.00042) respectively. Distance between the two points will show the tuning of the spam filter. Therefore:

$$\begin{aligned} \sigma &= \frac{\sqrt{(0.1307-0.14841)^2+(0.7289-0.00042)^2}}{\sqrt{2}} \\ &= 0.5153 \end{aligned} \tag{4.24}$$

$\sigma \neq 0$, which implies that the Yahoomail spam filter is not tuned according to the needs of this group of 49 students.

Utility provided by the Yahoomail spam filter is shown in the following calculation:

$$\begin{aligned} U &= 97 \cdot [0.7289 \cdot 153 + 0.2711 \cdot (-153)] + 3 \cdot [0.8693 \cdot 14 + 0.1307 \cdot (-14)] \\ &= 6825.231 \end{aligned} \tag{4.25}$$

Value of U=6825.231 shows that the utility provided by the Yahoomail spam filter to the given users is positive and high.

Yahoomail-ROC curve:

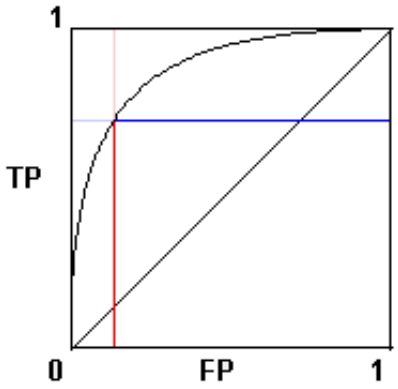

Figure 4.7: ROC curve for Yahoomail spam filter

Intersection of two line on the curve (Fig.4.7) shows the point LR_{YMail} for the specifics decision criterion such that TP and FP rates are 0.7289 and 0.1307 respectively

4.2.4 Analysis of MS Outlook (Exchange Server) Filter

Similar to subsection (4.2.1) this section will also, with all the same assumptions, investigate the MS Outlook (Exchange Server) spam filter based on statistics from the survey.

Out of 224 people 40 were MS Outlook (Exchange Server) users. Table4.4 shows the average number of mails collectively received in inboxes and spam folders by 40 people in 1 day and the amount of money they are ready to pay for avoiding a ham ending up in the spam folder and a spam ending up in the inbox.

As shown in the Table4.4:

$$\text{TP} = 281 \text{ and } \text{FN} = 201$$

$$\text{FP} = 13 \text{ and } \text{TN} = 640 - 201 = 439$$

$LR_{MS\_exch\_server}$ and $LR'_{MS\_exch\_server}$ are as follows:

Table 4.4: Shows survey statistics obtained from people using MS Outlook (Exchange Server). Statistics correspond to the total number of e-mails altogether received by 40 people in inboxes and spam folders in 1 day and money they are ready to pay for avoiding a ham ending up in the spam folder and a spam ending up in the inbox.

Number of mails received in your Inbox daily (HAM "good mail/non-spam/solicited mail" + SPAM).	640
Number of SPAM received in your SPAM folder daily.	281
Number of SPAM received in your Inbox daily.	201
Do you receive any HAM in your SPAM folder?	13
How much would you be willing to pay for avoiding that SPAM ever ends up in your Inbox ?	69 $ cents
How much would you be willing to pay for avoiding that HAM ever ends up in your SPAM folder ?	27900 $ cents

$$\begin{aligned} LR_{MS\_exch\_server} &= \frac{TP\_rate}{FP\_rate} \\ &= \frac{0.5829}{0.0287} \end{aligned} \tag{4.26}$$

$$LR'_{MS\_exch\_server} = \frac{0.8370}{0.0669} \tag{4.27}$$

Both $LR_{MS\_exch\_server}$ and $LR'_{MS\_exch\_server}$ can be represented on the ROC curve as points $\overline{(0.0287, 0.5829)}$ and $\overline{(0.0669, 0.8370)}$ respectively. Distance between the two points will show the tuning of the spam filter. Therefore:

$$\begin{aligned} \sigma &= \frac{\sqrt{(0.0287-0.0669)^2+(0.5829-0.8370)^2}}{\sqrt{2}} \\ &= 0.1817 \end{aligned} \tag{4.28}$$

$\sigma \neq 0$, which implies that the Yahoomail spam filter is not tuned according to the needs of this group of 40 students.

Utility provided by the Yahoomail spam filter is shown in the following calculation:

$$\begin{aligned} U &= 97 \cdot [0.5829 \cdot 69 + 0.4171 \cdot (-69)] + 3 \cdot [0.9713 \cdot 27900 + 0.0287 \cdot (-27900)] \\ &= 80005.3144 \end{aligned} \tag{4.29}$$

Value of U=80005.3144 shows that the utility provided by the Yahoomail spam filter to the given users is positive and very high.

MS exchange server spam filter-ROC curve:

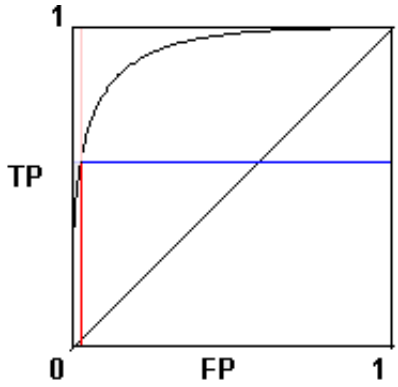

Figure 4.8: ROC curve for MS exchange server spam filter

Intersection of two line on the curve (Fig.4.8) shows the point $LR_{MS\_exch\_server}$ for the particular decision criterion such that TP and FP rates are 0.5829 and 0.0287 respectively

Chapter 5

Spam Filter Comparison and Discussion

Generally the comparison of spam filters are done on the basis of the TP, FN, FP and TN rates. The less the FP rate or the more the TP rate is the better the spam filter. This is the conventional rule to evaluate any filter by now.

Very often the main concern is on FPs because generally a FP carries more weight than other alternatives because in case of e-mails one would normally prefer receiving a spam message over losing a ham message but it may also depend on the user priorities.

Unlike said above, here we are not going to compare the spam filters on the basis of any rates but on the basis of the needs of the users and on the basis of the utility provided to them by the spam filters. We will compare on the basis of STI (σ) and utility (U), defined in the section (4.1). The less the value of σ the more the spam filter works according to user needs and the more the value of U the better is the filter for the user.

The survey was conducted among students. At first the comparison is made by analyzing the tuning of the spam filter for students (Table 5.1) and later Table (5.2) shows the comparison of the spam filters on the basis of the utility provided by spam filters to the students.

Though not the same students are surveyed for each spam filter but since just the students are surveyed for all 4 spam filters so we can assume that the students have same kind of priorities when it comes to loosing or accepting e-mails.

Therefore from the Table (5.1) we can say that MS Outlook Exchange Server

Table 5.1: Comparison of spam filters on the basis of Subjective Tuning Index (σ)

	σ
(1) MS Outlook Exchange Server Spam Filter	0.1817
(2) Yahoo Mail Spam Filter	0.5153
(3) Gmail Spam Filter	0.584
(4) Hotmail Spam Filter	0.6384

Table 5.2: Comparison of spam filters on the basis of Utility (U)

	U
(1) MS Outlook Exchange Server Spam Filter	0.80005.134
(2) Yahoo Mail Spam Filter	0.5153
(3) Gmail Spam Filter	0.584
(4) Hotmail Spam Filter	0.6384

Spam Filter is quite close in working according to the needs of the users as it has the minimum value of σ and Hotmail Spam Filter needs to be tuned quite a lot to work according to the users need because it has the maximum value of σ but both of them are not tuned according the needs of the users. They are not optimal.

Since MS Outlook Exchange Server Spam Filter provides more utility to the users than other spam filters in question. Results obtained from the calculations of the utility of the respective spam filters in section (4.2) which are also shown in Table (5.2) below shows that MS Outlook Exchange Server Spam Filter provides the maximum utility and Hotmail spam filter provides the minimum utility to the intended users. Therefore it can be concluded that MS Outlook Exchange Server Spam Filter is good, than other filters in experiment, for students

With the use of Signal Detection Theory for analyzing spam filters we can easily get to know if the spam filter which the user is using is tuned according to his needs or not. In addition to this we could also know the utility provided to the user by the spam filter.

Using this method it can be easily found out which filter is suitable for which user/group/oraganization or how much a filter needs tuning to satisfy the needs of the user.

Interesting results can be concluded from the Eq.(4.12). We can see that cost of FP is inversely proportional to the FP rate i.e. as the cost of the FP will increase the FP rate will decrease. It is important to talk about the FP because generally the cost of FP is higher than the other costs.

Talking about the decision criterion, while tuning the spam filter it should be noted that one can not simultaneously decrease the FP rate and increase the TP rate. One of them increases as the other decreases, therefore, it is very important to set the optimal decision criterion in general but for specific user(s) it can be set according to their needs.

Chapter 6

Conclusion and Future work

This thesis describes the analysis of spam filters within the framework of signal detection theory.

The criterion value plays an important part in decision making. It represents the environment in which the spam filter operates with the user's subjective view of the cost and benefits of false and correct filtering.

This thesis talks about the optimality of the spam filters. It sheds light on how to know whether the spam filter is tuned according to the needs of the particular user or not and what utility (positive or negative) does it provides to the user. Thus the user could easily choose which filter to use.

This could also be useful for the companies which make spam filters as with the application of SDT they can easily know the needs of the users of the organization and could build a spam filter which matches the needs of the organization on the whole. Therefore spam filter in future could be easily customized.

Future work could be based on analyzing social aspects of using a spam filter. It would be really interesting to study how a spam filter could effect social behavior of the user. Studies could be done on what type of people prefer which kind of filter, what changes are seen on user's social behavior after using particular spam filter which is tuned to certain level and how it could effect the social life of the user. Considering the social aspects, after knowing which level of tuning is best for what type of people, the social satisfaction level of the of the users would increase.

Bibliography

[1] Security intelligence. Tech. rep., Microsoft, December 2008.

[2] The carbon footprint of email spam report. Tech. rep., McAfee Inc. and ICF International, April 2009.

[3] Jonathan B. Postel. Simple Mail Transfer Protocol, 1982. RFC821.

[4] Abdi, H. Signal detection theory (sdt) overview.

[5] Agustin Orfila, Javier Carbo, A. R. Decision model analysis for spam. *Information and Security: An International Journal 15*, 2 (2004), 151–161.

[6] Androutsopoulos, I., Koutsias, J., Ch, K. V., Paliouras, G., and Spyropoulos, C. D. An evaluation of naive bayesian anti-spam filtering. In *Proceedings of the workshop on Machine Learning in the New Information Age, G. Potamias, V. Moustakis and M. van Someren (eds.), 11th European Conference on Machine Learning* (2000), pp. 9–17.

[7] Cournane, A., and Hunt, R. An analysis of the tools used for the generation and prevention of spam. *Computers and Security 23*, 2 (March 2004), 154–166.

[8] Dyrud, M. A. "i brought you a good news": An analysis of nigerian 419 letters. In *Proc. of the 2005 Association for Business Communication Annual Convention* (2005).

[9] Garcia, F. D., henk Hoepman, J., and Nieuwenhuizen, J. V. Spam filter analysis. In *in 'Proceedings of 19th IFIP International Information Security Conference, WCC2004-SEC* (2004), Kluwer Academic Publishers, pp. 395–410.

[10] Graham, C. *Monty Python's Flying Circus-Just the Words*. Methuen Publishing Ltd., 1989.

[11] Green, D. M., and Swets, J. A. *Signal Detection Theory and Psychophysics*. Peninsula Publishing, 1966.

[12] Heeger, D. Signal detection theory. Tech. rep., November 1997. Available at:http://www.cns.nyu.edu/∼david/handouts/sdt-advanced.pdf.

[13] Hird, S. Technical solutions for controlling spam. In *In proceedings of AUUG2002* (September 2002), pp. 4–6.

[14] Jøsang, A., and Pope, S. User centric identity management. In *in Asia Pacific Information Technology Security Conference, AusCERT2005, Austrailia* (2005), pp. 77–89.

[15] Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI* (1995), pp. 1137–1145.

[16] Lee, Y. The can-spam act: a silver bullet solution? *Commun. ACM 48*, 6 (2005), 131–132. Available at:http://portal.acm.org/citation.cfm?doid=1064830.1064863.

[17] McWilliams., B. *Spam Kings*. O'Reilly Media, Inc, 2004.

[18] Mikko, S., and Carl, S. Effective anti-spam strategies in companies: An international study. In *HICSS '06: Proceedings of the 39th Annual Hawaii International Conference on System Sciences* (Washington, DC, USA, 2006), IEEE Computer Society, p. 127.3.

[19] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. A bayesian approach to filtering junk e-mail. In *AAAI Workshop on Learning for Text Categorization* (July 1998).

[20] Shirali-Shahreza, S., and Movaghar, A. A new anti-spam protocol using captcha. *Networking, Sensing and Control, 2007 IEEE International Conference on* (April 2007), 234–238.

[21] Swets, J. A. *Signal Detection Theory and Roc Analysis in Psychology and Diagnostics: Collected Papers*. Lawrence Erlbaum Associates, NJ, 1996.

[22] Thuerp, G. Father of spam-interview, May 2008. Available at:http://www.npr.org/templates/player/mediaPlayer.html?action=1&t=1&islist=false&id=90160617&m=90160591.

[23] Wickens, T. D. *Elementary Signal Detection Theory.* Oxford University Press (OUP), 2001.

Chapter 7

Appendices

Table 7.1: Questionnaire for Spam Survey for Gmail & Yahoo Mail Users

	Gmail	Yahoo Mail
Where do you live?	Country	11C
Number of mails received in your Inbox daily (HAM "good mail/non-spam/solicited mail" + SPAM)	(1) less than 5, (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0	(1) less than 5, (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0
Number of SPAM received in your SPAM folder daily	(1) less than 5 (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0	(1) less than 5, (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0
Number of SPAM received in your Inbox daily	(1) less than 5, (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0	(1) less than 5, (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0
Do you receive any HAM in your SPAM folder	(1) No, I don't receive any, (2) Yes 1 daily, (3) Yes 2 daily, (4) Yes 3 or more daily, (5) Yes 1 in a month, (6) Yes 2 in a month, (7) Yes 3 or more in a month, (8) Yes 1 in a year, (9) Yes 2 in a year, (10) Yes 3 or more in a year,	(1) No, I don't receive any, (2) Yes 1 daily, (3) Yes 2 daily, (4) Yes 3 or more daily, (5) Yes 1 in a month, (6) Yes 2 in a month, (7) Yes 3 or more in a month, (8) Yes 1 in a year, (9) Yes 2 in a year, (10) Yes 3 or more in a year,
How much would you be willing to pay for avoiding that SPAM ever ends up in your Inbox ?	(1) Nothing, I'm OK with spams, (2)1cent of $ daily, (3)2cent of $ daily, (4)5cent of $ daily, (5)10cent of $ daily, (6)25cent of $ daily, (7)50cent of $ daily, (8)1$ daily, (9)2$ daily, (10)5$ daily, (11)10$ daily, (12)25$ daily, (13)50$ daily, (14)100$ daily, (15) more than 100$ daily	(1) Nothing, I'm OK with spams, (2)1cent of $ daily, (3)2cent of $ daily, (4)5cent of $ daily, (5)10cent of $ daily, (6)25cent of $ daily, (7)50cent of $ daily, (8)1$ daily, (9)2$ daily, (10)5$ daily, (11)10$ daily, (12)25$ daily, (13)50$ daily, (14)100$ daily, (15) more than 100$ daily
How much would you be willing to pay for avoiding that HAM ever ends up in your SPAM folder ?	(1) Nothing, I'm OK with ham in spam folder, (2)1cent of $ daily, (3)2cent of $ daily, (4)5cent of $ daily, (5)10cent of $ daily, (6)25cent of $ daily, (7)50cent of $ daily, (8)1$ daily, (9)2$ daily, (10)5$ daily, (11)10$ daily, (12)25$ daily, (13)50$ daily, (14)100$ daily, (15) more than 100$ daily	(1) Nothing, I'm OK with ham in spam folder, (2)1cent of $ daily, (3)2cent of $ daily, (4)5cent of $ daily, (5)10cent of $ daily, (6)25cent of $ daily, (7)50cent of $ daily, (8)1$ daily, (9)2$ daily, (10)5$ daily, (11)10$ daily, (12)25$ daily, (13)50$ daily, (14)100$ daily, (15) more than 100$ daily

Table 7.2: Questionnaire for Spam Survey for Hotmail & MS Outlook (Exchange Server) Users

	Hotmail	MS Outlook (Exchange Server)
Where do you live?	Country	11C
Number of mails received in your Inbox daily (HAM "good mail/non-spam/solicited mail" + SPAM)	(1) less than 5, (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0	(1) less than 5, (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0
Number of SPAM received in your SPAM folder daily	(1) less than 5 (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0	(1) less than 5, (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0
Number of SPAM received in your Inbox daily	(1) less than 5, (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0	(1) less than 5, (2) between 6-10, (3) between11-15, (4) between 16-20, (5) more than 20, (6) 0
Do you receive any HAM in your SPAM folder	(1) No, I don't receive any, (2) Yes 1 daily, (3) Yes 2 daily, (4) Yes 3 or more daily, (5) Yes 1 in a month, (6) Yes 2 in a month, (7) Yes 3 or more in a month, (8) Yes 1 in a year, (9) Yes 2 in a year, (10) Yes 3 or more in a year,	(1) No, I don't receive any, (2) Yes 1 daily, (3) Yes 2 daily, (4) Yes 3 or more daily, (5) Yes 1 in a month, (6) Yes 2 in a month, (7) Yes 3 or more in a month, (8) Yes 1 in a year, (9) Yes 2 in a year, (10) Yes 3 or more in a year,
How much would you be willing to pay for avoiding that SPAM ever ends up in your Inbox ?	(1) Nothing, I'm OK with spams, (2)1cent of $ daily, (3)2cent of $ daily, (4)5cent of $ daily, (5)10cent of $ daily, (6)25cent of $ daily, (7)50cent of $ daily, (8)1$ daily, (9)2$ daily, (10)5$ daily, (11)10$ daily, (12)25$ daily, (13)50$ daily, (14)100$ daily, (15) more than 100$ daily	(1) Nothing, I'm OK with spams, (2)1cent of $ daily, (3)2cent of $ daily, (4)5cent of $ daily, (5)10cent of $ daily, (6)25cent of $ daily, (7)50cent of $ daily, (8)1$ daily, (9)2$ daily, (10)5$ daily, (11)10$ daily, (12)25$ daily, (13)50$ daily, (14)100$ daily, (15) more than 100$ daily
How much would you be willing to pay for avoiding that HAM ever ends up in your SPAM folder ?	(1) Nothing, I'm OK with ham in spam folder, (2)1cent of $ daily, (3)2cent of $ daily, (4)5cent of $ daily, (5)10cent of $ daily, (6)25cent of $ daily, (7)50cent of $ daily, (8)1$ daily, (9)2$ daily, (10)5$ daily, (11)10$ daily, (12)25$ daily, (13)50$ daily, (14)100$ daily, (15) more than 100$ daily	(1) Nothing, I'm OK with ham in spam folder, (2)1cent of $ daily, (3)2cent of $ daily, (4)5cent of $ daily, (5)10cent of $ daily, (6)25cent of $ daily, (7)50cent of $ daily, (8)1$ daily, (9)2$ daily, (10)5$ daily, (11)10$ daily, (12)25$ daily, (13)50$ daily, (14)100$ daily, (15) more than 100$ daily