ERVER-BASED E-MAIL FILTERING

SITI ZUBAIDAH BINTI ZAKARIA

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Computer Network)

FACULTY OF INFORMATION TECHNOLOGY AND COMMUNICATION
KOLEJ UNIVERSITI TEKNIKAL KEBANGSAAN MALAYSIA
2006

BORANG PENGESAHAN STATUS TESIS^

JUDUL : SERVER-BASED EMAIL FILTERING

SESI PENGAJIAN : 2006

Saya SITI ZUBAIDAH BINTI ZAKARIA

(HURUF BESAR)

mengaku membenarkan tesis (PSM/~~Sarjana/Doktor Falsafah~~) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut :

1. Tesis adalah hakmilik Kolej Universiti Teknikal Kebangsaan Malaysia.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

_____	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
_____	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
/	TIDAK TERHAD	

(TANDATANGAN PENULIS)

Alamat tetap : 550, JLN 2/9, TAMAN BUKIT KEPAYANG, 70200 SEREMBAN, N.S.

Tarikh : 09 JUN 2006

(TANDATANGAN PENYELIA)

ZAKIAH BINTI AYOP

Nama Penyelia

Tarikh : 9 JUNE 2006

CATATAN : ** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

^ Tesis dimaksudkan sebagai Laporan Projek Sarjana Muda (PSM)

DECLARATION

I hereby declare that this project report entitled

SERVER BASED E-MAIL FILTERING

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT : _____ Date : 9 JUN 2006

(SITI ZUBAIDAH BINTI ZAKARIA)

SUPERVISOR : _____ Date : 9 JUNE 2006

(ZAKIAH BINTI AYOP)

DEDICATION

To my much-loved Emak and Abah...

ACKNOWLEDGEMENTS

Special thanks and appreciation dedicated to my supervisor, Cik Zakiah binti Ayop for her willingness and patience in giving me her guidance, advices and critics upon developing this project. Her excellent supervision is one of the main reasons for the success of this Server-based E-mail Filtering project. This appreciation also goes to all my lecturer and panels for this project.

My deeply thanks I would like to tender to my beloved parents, family and loved one, who continuously give their support, encouragement, and love throughout developing this project. My thanks also go to all my friends, classmates, and personal person for their support and help along the way.

All of your encouragements, understanding, guidance, helps, and love are my strengths that lead me to the completion of this project. Thank you so much.

ABSTRACT

This *Projek Sarjana Muda* (PSM) Thesis Report contains a report of Server-Based E-mail Filtering Project. The report is divided into seven chapters, which are Introduction, Literature Review and Methodology, Analysis, Design, Implementation, Testing, and Project Conclusion. In Introduction Chapter, it reviews on the description of the project in whole but briefly, including the objectives, scopes, problem analysis and project significance. Chapter II which is Literature Review and Methodology chapter, it discuss about the research that had been made for the project, fact and findings, project methodology, project requirements and project milestones. As for Chapter III, this is where problem analysis and requirement analysis reviewed. Chapter IV is the project design which includes high-level design, network architecture, logical design, physical design and security requirements. Chapter V is the Implementation Phase that will describe about the activities that will involve in configuration and setup environment and what the expected output after completing the phase, and as for Testing chapter, it will review about test plan, test strategy, test design, test results and test analysis of the project. The last chapter, which is the Project Conclusion, will observe about the weakness and strengths of the project, prepositions for improvement, contribution and the conclusion. This project report also reviews about the e-mail filtering method to be used, the rules to be applied, the methodology, the e-mail network environment, e-mail flow, e-mail architecture, e-mail filtering issues, the methods involve in e-mail filtering, current issues, the problems occur and the best ways that hopefully to solve the problems. Reviews in full details can be found in the later chapters of this report.

ABSTRAK

Laporan Tesis Projek Sarjana Muda (PSM) ini mengandungi laporan tentang Projek Penapisan Emel di Peringkat Server *(Server-based E-mail Filtering)*. Laporan ini dibahagikan kepada tujuh bab, yang merangkumi Pendahuluan *(Introduction)*, Kajian Literatur dan Metodologi *(Literature Review and Mthodology)*, Analisis *(Analisis)*, Lakaran *(Design)*, Pelaksanaan *(Implementation)*, Percubaan *(Testing)*, dan Kesimpulan Projek *(Project Conclusion)*. Bab Pertama menghalkan tentang gambaran projek secara ringkas dan menyeluruh, termasuklah objektif, skop, analisa masalah, dan kepentingan projek. Bab Kedua pula membincangkan tentang kajian yang telah dijalankan terhadap projek, fakta-fakta dan hasil kajian, metodologi projek yang dijalankan, keperluan projek dan jadual perjalanan projek. Bab Ketiga merangkumi analisa masalah, dan keperluan terhadap projek. Bab Empat meliputi lakaran projek, termasuk lakaran rangkaian, lakaran logikal, lakaran fizikal, dan keperluan sekuriti projek. Bab Kelima, iaitu Pelaksanaan pula menerangkan tentang aktiviti-aktiviti yang dijalankan semasa konfigurasi dan pelaksanaan projek, serta hasil yang dijangka selepas projek berjaya dijalankan. Bab Keenam iaitu Percubaan meliputi pelan percubaan, strategi, lakaran percubaan, hasil percubaan, dan analisa percubaan projek. Bab yang terakhir sekali mengkaji tentang kekurangan dan kelebihan projek, cadangan pembangunan, dan kesimpulan. Laporan projek ini juga mengkaji tentang metod yang terbaik untuk menapis emel di peringkat server, persekitaran rangkaian emel, perjalanan emel mesej, rekabentuk emel, isu-isu tentang penapisan emel, metod-metod lain yang terlibat dalam penapisan emel, masalah-masalah yang dihadapi disebabkan oleh emel spam dan gangguan-gangguan lain serta langkah terbaik yang dikenalpasti boleh membantu mengurangkan masalah yang dihadapi. Laporan penuh tentang projek ini boleh dilihat dalam bab-bab seterusnya.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS / ACRONYMS

AI	-	Artificial Intelligence
TCP	-	Transfer Control Protocol
UDP	-	User Datagram Protocol
TCP/IP	-	Transmission Control Protocol / Internet Protocol
ISP	-	Internet Service Provider
PC	-	Personal Computer
RAM	-	Random Access Memory
ICMP	-	Internet Controls Message Protocol
DNS	-	Domain Name System
IP	-	Internet Protocol
FQDN	-	Fully Qualified Domain Name
FTMK	-	Fakulti Teknologi Maklumat dan Komunikasi
MTA	-	Mail Transfer Agent
MUA	-	Mail User Agent
MDA	-	Mail Delivery Agent
RFC	-	Request for Comments
BCC	-	Blind Carbon Copy
CC	-	Carbon Copy
MIME	-	Multipurpose Internet Mail Extensions
RBLs	-	Realtime Blackhole Lists
DNSBLs	-	DNS Blacklists
NIC	-	Network Interface Card
LAN	-	Local Area Network
POP	-	Pop Office Protocol

LIST OF ATTACHMENTS

CHAPTER I

INTRODUCTION

1.0 Project Background

E-mail filtering is the processing of organizing electronic e-mails according to specified criterion. This refers to the automatic processing of incoming e-mail and outgoing e-mail messages, as well as those being received which relies on the server software that uses the basic form of Artificial Intelligence (AI) involvement in order to separate spam from legitimate (valid) e-mail messages.

E-mail filtering has several benefits and criticisms. For incoming e-mail, filtering is performed to:

- Eliminate the sending and receipt of unsolicited e-mails and computer viruses. Viruses are often attached to unsolicited e-mails.
- Move e-mails for efficient access
- Automatically reply to e-mails, notifying the sender that their message has been received successfully. Common auto responders are the out of office reply and vacation message, advising that the intended recipient is away from the keyboard, when they are scheduled to return, and alternative contact details.

Content filtering is a method that can be used to isolate and identify keywords that signal the presence of malicious types of e-mails. Content filtering deals with what information is allowed into a network. Organizations must work not to only protect against outside hackers breaking into secure networks (access control), but they must work to protect the information that comes into the network via e-mail (content control). This is done through content filtering.

Content filtering or e-mail filtering will protect an organization's network from infection from e-mail-borne-viruses, network congestion from system misuse, as well as loss of network service from spam and spoof attacks.

When using the content filtering tool, all e-mail is filtered at the server before it reaches the intended recipient. E-mail can be filtered based on sender, subject, excessive file size, prohibited content, profanities, corrupted data, pornography, or racist or hate e-mails.

1.2 Problem Statement

Some organizations that have their own e-mail server face the same problem : spam e-mails. This lead to many other problems to occur, such as :

1. Spam makes up 30% to 60% of mail traffic and is on the rise
2. Storage of unwanted e-mail that can cause shutdown of mailboxes
3. Managing and deleting unwanted messages, negative effect on productivity
4. Difficult or impossible to unsubscribe brings to the same problem to be occurred again.

Nor does virus-scanning software protect against all e-mail viruses and attacks. Anti-virus vendors cannot always update their signatures in time against the deadly viruses that are distributed worldwide via e-mail in a short time (such as the

LoveLetter virus and its variants). This means that organizations that are using a single virus-scanning engine alone are not necessarily safeguarded when a new virus is released. Yet, once an e-mail virus has entered the system, it takes one quick click for an unwitting user to activate it.

Hopefully, the project can filter or will develop ways to filter out anything at the server side that an organization or a user doesn't want to receive. Usually, these messages fall into three general categories: viruses (malicious code), spam (bulk, automated, or commercial unsolicited e-mail), and abuse (targeted offensive or abusive messages).

1.3 Objectives

The objectives of this project are :

1. To study the best e-mail filtering method to be used and applied in the e-mail server

2. To find and apply selected filtering rules that is suitable to be applied in the e-mail server

3. To build an e-mail filtering server that will filter e-mail at the server side according to the rules applied before it reaches the intended recipients

1.4 Scopes

The main purpose of this project is to study and apply selected filtering rules that are most suitable to be applied in the e-mail server, and then will filter all e-mails at the server before they reach the intended recipient. E-mail can be filtered

based on sender, subject, excessive file size, prohibited content, profanities, corrupted data, pornography, or racist or hate e-mails. Thus when the e-mails are filtered at the server, the recipient will only receive legitimate e-mails and somehow can protect the network from malicious attacks and so the user can save their e-mail account spaces taken by those spam e-mails.

The system administrator will control the handling of e-mail filtering tool or software. System administrators are able to assign numerous quarantine areas, plus also in some cases, they are responsibilities determine further actions to the e-mails that are suspected as spam or malicious.

1.4.1 System Administrators

As the system administrators are the 'admin' of the system, they have to know 'how' the network is protected and from what it is protected from. Therefore, the tool allows System Administrators to assign numerous quarantine areas, how the quarantines e-mails are dealt with, in most cases, in predefined manner. Sometimes in some cases, the System Administrator is responsibility to determine further actions to the e-mails that are suspected as spam or unwanted. The software can be a helpful tool for the System Administrators.

1.4.2 Clients

Clients basically will receive e-mails that are already filtered at the server. This is the main reason why the project is developed: to filter e-mails at server side so that the clients/recipients will only receive legitimate e-mails that are clean and unharmed. Their e-mail account spaces taken by spam e-mails also can be saved for another uses.

1.5 Project Significance

This system is very useful tool for e-mail server of any network or organizations, who their always want to receive only 'clean' e-mails that are free from spam and viruses, thus can protect the network itself from attacks and viruses. The system also can help the System Administrator to protect the network from malicious attacks that come with e-mails.

1.6 Expected Output

Some assumptions had been made to make sure the effectiveness and the smoothness of the project upon developing. The expected output for the project are :

- The e-mail server will be able to apply and execute the filtering rules that had been applied the server
- The e-mail server will be able to filter e-mails at the server side before they reach the recipients

1.7 Conclusion

Server based e-mail filtering is important for every network which it can protect the network in a way of securing organization network from outside attacks that come with e-mails, filters all incoming e-mails stored in the e-mail server and delete all harmful and unnecessary e-mails to make sure the clients will receive only legitimate e-mails and therefore will protect the network itself from deadly attacks.

The next chapter will discuss more deeply about the development, methodology and requirements of this project.

CHAPTER II

LITERATURE REVIEW AND PROJECT METHODOLOGY

2.1 Introduction

A literature review is a "critical analysis of a segment of a published body of knowledge through summary, classification, and comparison of prior research studies, reviews of literature, and theoretical articles." (Wisconsin)

Generally, the purpose of a literature review is to analyze critically a relevant part of a published body of knowledge through summary, classification, and comparison of prior research studies, white papers, reviews of literature, and theoretical articles.

Project methodology describes the methodology or approaches that will be used in developing the project. This including the activities and actions that will be taken for every stage stated in the methodology.

2.2 Fact and Finding

2.2.1 Components of E-mail Messages

There are three components to an e-mail message (www.bath.ac.uk/bucs/e-mail/anatomy.shtml) :

- The envelope
- The headers
- The message body

The Envelope

Users never see the Envelope as it is used internally by the Message Transfer Agent (MTA) to route the message. The MTA is the server software used to transfer e-mail over the network. Examples of MTAs include Sendmail, Exim, Qmail and Postfix.

Thus, the user do not really need to be concerned with the Envelope except to understand it is the Envelope that gets the user e-mail from sender to recipient not what the user see in the To: and From: headers when the user is reading the message.

The Headers

The most interesting part of an e-mail in the opinion of the systems administrators and engineers anyway.

- KEY: VALUE pairs that conform to RFC 822.
- Each header transmitted as a single line of text.
- Some are Mandatory: Date From, To, (or BCC).
- Others are optional but widely used: Subject, Cc, Reply-To, Received, Message-Id.
- Any others are ignored by the mail system but all headers are

propagated, recognized or not.

- Headers starting with 'X-' are for personal application or institution use.

a. Components of E-mail Headers

It was mentioned above that e-mail headers are the most interesting art of an e-mail message for system administrators and engineers. This is arguable but what is not is the fact that headers are the most important part of an e-mail for tracing and diagnosing problems. Headers may look garbled and incomprehensible but that can be said of anything upon first contact.

The ability to read and decipher e-mail headers is a useful skill to learn for tracing messages to their original source and diagnosing many other problems. Headers may contain a lot of information but the most important information will always be contained in every e-mail header.

- Pragmatically the 'Received' header lines are the most important.
- Each intervening MTA adds a Received header.

Here is an example header extract from an e-mail:

> ***Received:*** *from kelly.bath.ac.uk ([138.38.32.20]) by serena.bath.ac.uk with esmtp (Exim 3.36 #2) id 19NkR7-0003g5-00 for ccsxxx@imaps.bath.ac.uk; Thu, 05 Jun 2003 03:25:05 +0100*
> ***Received:*** *from mta1.silver.ddc.dartmail.net ([146.82.220.228]) by kelly.bath.ac.uk with smtp id 19NkQa-00008C-4E for ccsxxx@bath.ac.uk; Thu, 05 Jun 2003 03:24:32 +0100*
> ***Date:*** *Wed, 04 Jun 2003 22:25:43 -0400 (EDT)*

Figure 2.1 : Example of E-mail Header

i. Reading e-mail header from top to bottom:

- The message was received FROM the machine 'kelly.bath.ac.uk' - the IP address is also given

- It was received BY the machine 'serena.bath.ac.uk' - the local MTA's version (in this case Exim) and a local ID is given
- The intended recipient and the receipt-date are given
- New 'Received' lines are added from the top so...in turn
- 'kelly.bath.ac.uk' received the message from 'mta1.silver.ddc.dartmail.net'
- Note the time stamps both MTAs inserted and in the Date field. Date and Time are given in UTC (GMT) time with a (local) offset.

At this point it is probably worth noting that spammers will often add fake 'Received' lines.

ii. Other Header Lines

- 'Message-Id' - this is added by the originating MTA; it should be a globally unique identifier
- 'Return-Path' - this is supposed to contain the ENVELOPE sender's address; BOUNCED mail gets sent to this address.
- 'Date' - this is added by the originating User Agent. It may not be truthful.
- 'CC' - the recipients of a cc'd e-mail
- 'Reply-To' - the reply-to address if the sender set it.
- 'MIME-Version' - used for encoding binary content as attachments.
- 'X-headers' - For personal application or institution use. Either added by the users client program or an MTA.
- And many others

Figure 2.2 shows an example of a slightly more widely e-mail header: