HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Đinh Quang Định

NGHIÊN CỨU XÂY DỰNG HỆ THỐNG LỌC NỘI DUNG THƯ ĐIỆN TỬ GỬI THEO GIAO THỨC SMTP

Chuyên ngành: Truyền dữ liệu và mạng máy tính

Mã số: 60.48.15

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2013

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. Nguyễn Thành Phúc

Phản biện 1: ……

Phản biện 2: ……..

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

HÀ NỘI - 2013

MỞ ĐẦU

Sự phát triển không ngừng của khoa học công nghệ, đặc biệt là tốc độ bùng phát mạnh mẽ của công nghệ thông tin đã khiến cho Internet trở thành một công cụ hữu ích không thể thiếu trong đời sống hiện nay. Không chỉ là kho tài nguyên thông tin khổng lồ, Internet còn cung cấp cho người sử dụng các công cụ khai thác thông tin tiện lợi, nhanh chóng và hiện đại. Một trong số các dịch vụ mạng được người dùng sử dụng rộng rãi nhất là dịch vụ thư điện tử (e-Mail). Dịch vụ này cho phép người sử dụng trao đổi thư tín một cách dễ dàng, đồng thời có thể phổ biến kiến thức, thông tin, thông báo về nội dung chính sách của một vấn đề trong một cơ quan, tổ chức… Với tính chất dễ sử dụng, không phụ thuộc vào không gian và thời gian, thư điện tử đã trở thành một phương tiện trao đổi thông tin quan trọng đối với nhiều người.

Cùng với sự phát triển tiện lợi của Internet, việc lấy cắp thông tin, thâm nhập bất hợp pháp, phá hoại thông qua Internet cũng gia tăng về số lượng, loại hình và kỹ thuật. Bên cạnh đó, các đối tượng có tư tưởng chính trị cực đoan, các thế lực phản động ở trong nước và ngoài nước tận dụng triệt để những khả năng của mạng Internet để phục vụ cho mục đích tuyên truyền, phát tán tài liệu phản động và thực hiện các hành vi phản động khác nhằm chống phá nhà nước cộng hòa xã hội chủ nghĩa Việt Nam. Do đó, việc đấu tranh chống các hoạt động sử dụng các dịch vụ trên mạng phục vụ cho mục đích xấu đã trở thành một yêu cầu cấp thiết của lực lượng Công an.

Trước nhu cầu thực tế đó, tôi đã nghiên cứu và mạnh dạn chọn đề tài: **"*Nghiên cứu xây dựng hệ thống lọc nội dung thư điện tử gửi theo giao thức SMTP*"** làm báo cáo tốt nghiệp nhằm đề xuất một giải pháp kiểm soát các nội dung của thư điện tử gửi đi trên mạng Internet.

Mục tiêu của đề tài

Xây dựng hệ thống lọc nhằm tìm ra những thư được gửi trên mạng Internet theo giao thức SMTP có nội dung liên quan đến công tác nghiệp vụ Công an cần quan tâm; để phục vụ yêu cầu nghiệp vụ trong lực lượng.

Phạm vi nghiên cứu

Những thư có nội dung liên quan đến các vấn đề mà công tác Công an quan tâm (nội dung của bức thư được gửi có thể hiển thị dưới dạng text .doc, .docx, .pdf, .txt…; thư có nội dung thuộc các địa chỉ cần theo dõi và những cụm từ mà công tác nghiệp vụ công an phải quan tâm…).

Giới hạn nội dung những thư được viết bằng ngôn ngữ tiếng Việt với các loại font chữ tiếng Việt hiện có (TCVN3, Unicode, VNI, UTF8…). Đồng thời, thư được gửi đi trên mạng theo giao thức SMTP.

Phương pháp nghiên cứu

- Nghiên cứu một thư được gửi theo giao thức SMTP để phân tách ra các phần trong nội dung bức thư như: địa chỉ gửi, nhận, chủ đề bức thư, nội dung bức thư, các file văn bản đính kèm… Trên cơ sở đó sử dụng phương pháp so sánh đa chuỗi (multi matching) để tìm kiếm các cụm từ cần quan tâm và đối sánh để xây dựng chính sách lọc theo yêu cầu nghiệp vụ.

- Nghiên cứu các font chữ tiếng Việt và phương pháp chuyển đổi các font chữ, từ đó chuyển đổi các font chữ sang một font tiêu chuẩn rồi áp dụng phương pháp tìm kiếm và so sánh đa chuỗi thực hiện việc lọc nội dung thư.

Ngoài phần lời mở đầu, kết luận, tài liệu tham khảo và phụ lục, luận văn được chia thành 3 chương:

Chương 1: Tổng quan về hệ thống thư điện tử

Chương này nói về cấu trúc của một thư điện tử, một hệ thống thư tín điện tử và các giao thức được sử dụng cho dịch vụ thư điện tử. Ngoài ra, chương này cũng giới thiệu giao thức SMTP, đi sâu vào nghiên cứu việc gửi một thư điện tử qua giao thức SMTP, nghiên cứu proxy cho SMTP.

Chương 2: Phương pháp và các kỹ thuật sử dụng cho hệ thống lọc nội dung thư điện tử

Chương này trình bày về những thông tin cần kiểm soát đối với một thư điện tử; một số biện pháp kiểm soát nội dung thư điện tử; các kỹ thuật sử dụng để phục vụ cho việc lọc thư: kỹ thuật tìm kiếm, so sánh chuỗi (multi matching), xử lý phông tiếng Việt (Unicode, TCVN3,VNI, UTF8…)

Chơơng 3: Lọc nội dung thơ điện tử gửi theo giao thức SMTP

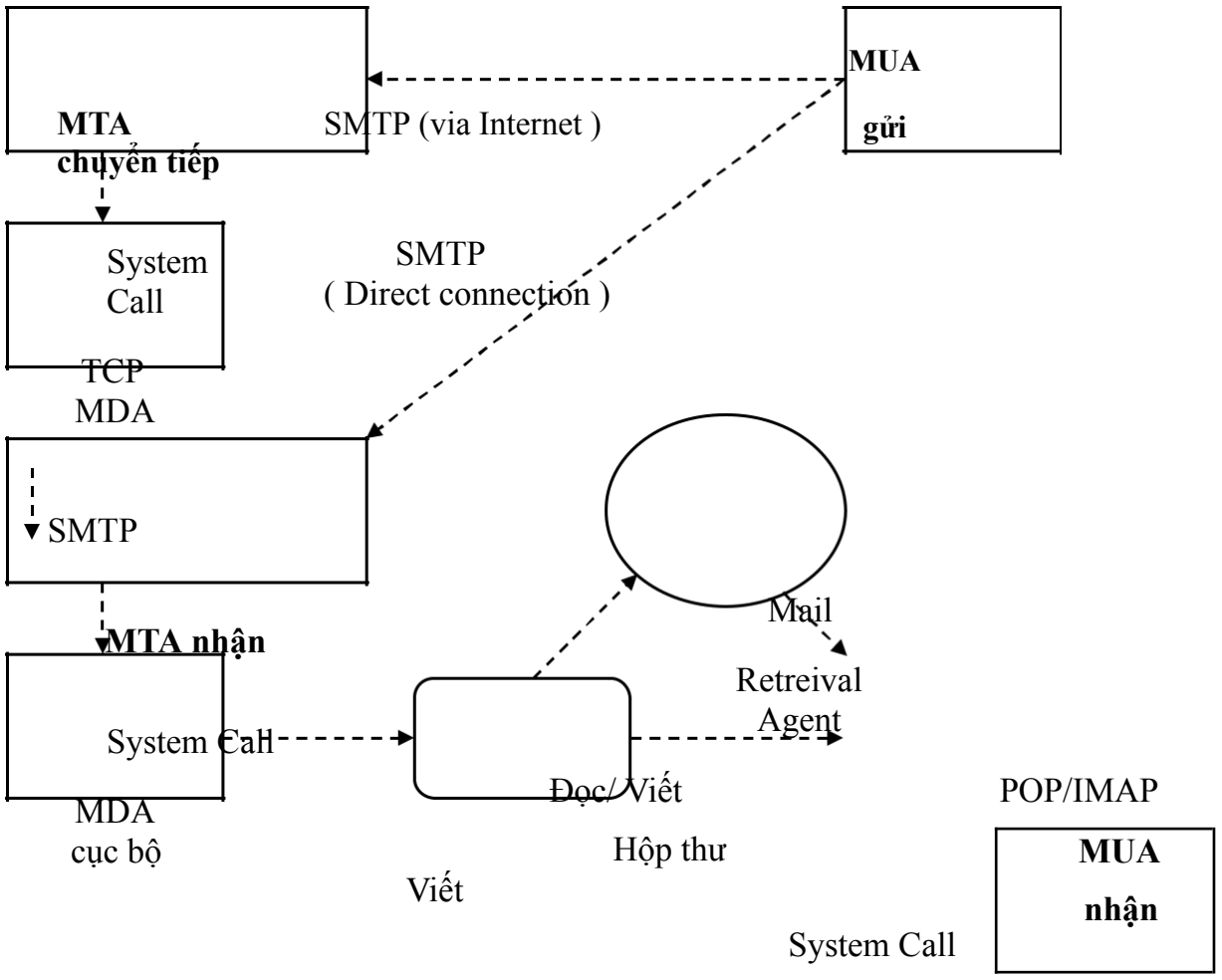
Chương này đề xuất phương pháp lọc về mô hình, xây dựng hệ thống lọc nội dung thư điện tử gửi theo giao thức SMTP.

Chương 1- TỔNG QUAN VỀ HỆ THỐNG THƯ ĐIỆN TỬ

1.1 Lý thuyết chung về thư điện tử

Hệ thống thư điện tử cho phép người dùng trao đổi thư điện tử với nhau. Hệ thống này bao gồm một hoặc nhiều máy chủ thư tín (mail server), trên đó có cài đặt một phần mềm mail server để quản lý tài khoản của người dùng, thực hiện việc trao đổi thư giữa những người dùng và trao đổi thư với các máy chủ thư tín khác.

1.1.1 Cấu trúc một hệ thống thư điện tử



Hình 1.1: Mô hình một hệ thống thoị tín điện tử

Để thực hiện việc trao đổi thư với người dùng, giữa máy chủ (mail server) và các máy khách (mail client) thống nhất sử dụng chung một bộ giao thức gửi và nhận thư, trong đó có quy định cụ thể về cổng làm việc, quy trình thao tác, các câu lệnh trao đổi, cấu trúc của thư điện tử... Hình 1.1 trình bày mô hình một hệ thống thư tín điện tử với giao thức gửi thư SMTP và giao thức nhận thư POP hoặc IMAP.

Hệ thống này bao gồm bốn phần tử chính: MUA (Mail User Agent), MTA (Mail Transfer Agent), MDA (Mail Delivery Agent), MRA (Mail Retrieval Agent).

Mail User Agent (MUA): là chương trình phần mềm của máy client được người dùng sử dụng để gửi, nhận, soạn thảo, xử lý thư điện tử.

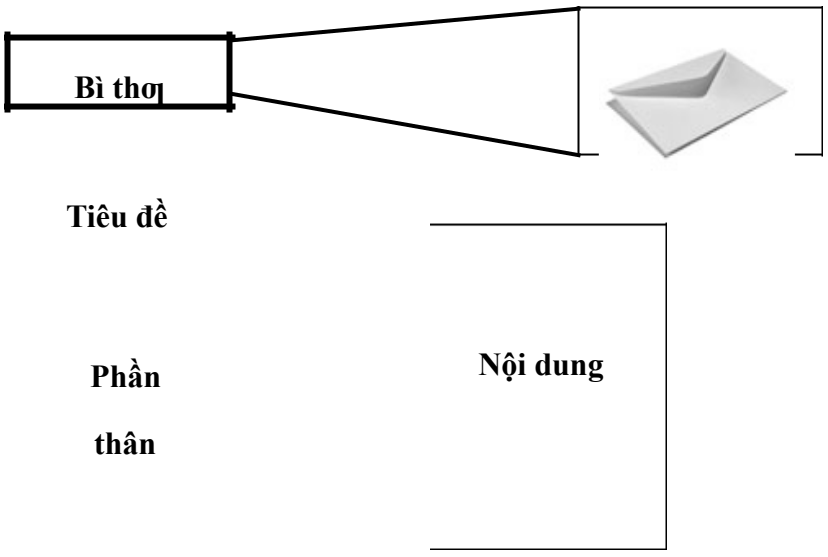
Mail Tranfer Agent (MTA): là một chương trình thư của máy chủ, cho phép truyền tải thư điện tử từ máy này sang máy khác.

Mail Delivery Agent (MDA): là chương trình mà MTA sử dụng để chuyển thư vào hộp thư của người dùng hoặc để truyền tải thư tới một MTA khác. Mỗi MTA sử dụng một hoặc nhiều MDA, mỗi MDA được sử dụng cho một loại yêu cầu phân phát thư riêng.

Mail Retrieval Agent (MRA): là một chương trình hoặc một dịch vụ có chức năng lấy thư điện tử về từ một hộp thư trên một máy chủ ở xa và đưa chúng tới một MUA. Các MRA truy vấn các thư và các phần header từ những hộp thư ở xa và phân phát chúng tới các MUA trên máy của người dùng.

1.1.2 Cấu trúc một thư điện tử

Thư điện tử thường có hai phần chính: phần đầu (Header) và phần thân (Body) là văn bản chứa nội dung của thư. Khi gửi đi, toàn bộ thư điện tử được gói trong nội dung (content). Ngoài ra, hệ thống thư còn tạo thêm một phần nữa được gọi là bì thư (envelope), phần này chứa các thông tin cần thiết cho việc chuyển thư đến nơi nhận. Hình 1.2 trình bày cấu trúc một thư điện tử.



Hình 1.2: Cấu trúc của một thọ điện tử

1.1.3 Một số giao thức sử dụng để gửi và nhận thư điện tử

Hệ thống Mail được xây dựng dựa trên một số giao thức: Simple Mail Transfer Protocol (SMTP), Post Office Protocol (POP), Multipurpose Internet Mail Extensions (MIME) và Interactive Mail Access Protocol (IMAP) được định dạng trong RFC 1176 là một giao thức quan trọng để thay thế POP, nó cung cấp nhiều cơ chế tìm kiếm văn bản, phân tích message từ xa mà ta không tìm thấy trong POP.

1.1.3.1 Một số giao thức sử dụng để gửi thư điện tử

1.1.3.1.1 Giao thức SMTP (Simple Mail Transfer Protocol)

SMTP là giao thức tin cậy, chịu trách nhiệm phân phát thư điện tử. Nó chuyển thư điện tử từ hệ thống mạng này sang hệ thống mạng khác, chuyển thư trong hệ thống mạng nội bộ. Giao thức này sẽ được trình bày cụ thể ở phần *Xây dựng proxy cho SMTP.*

1.1.3.1.2 Giao thức X.400

X.400 là giao thức được ITU-T và ISO định nghĩa và đã được ứng dụng rộng rãi ở Châu Âu, Canada. X.400 cung cấp tính năng điều khiển và phân phối e-Mail, sử dụng định dạng nhị phân, do đó không cần mã hóa nội dung khi phân phát thư trên mạng Internet.

1.1.3.2 Một số giao thức sử dụng để nhận thư điện tử

Có hai giao thức chính thường được dùng bởi các ứng dụng máy thư khách để truy cập thư tín từ các máy chủ: Post Office Protocol (POP) và Internet Message Access Protocol (IMAP).

1.1.3.2.1 Giao thức POP

POP là giao thức được thiết kết để hỗ trợ tiến trình thư "offline", trong tiến trình này thư được phân phát tới một máy chủ. Một máy tính cá nhân người dùng gọi định kỳ một chương trình thư khách được kết nối tới máy chủ và tải tất cả thư treo đó tới máy tính của người dùng. Cách truy cập offline là một loại dịch vụ *store-to-forward*, được sử dụng để chuyển thư (theo đơn đặt hàng) từ máy chủ thư (vị trí đưa về) tới máy của người đọc thư, thường là một PC hoặc Mac.

1.1.3.2.2 Giao thức IMAP (Internet Message Access Protocol)

IMAP là một giao thức chuẩn cho việc truy cập thư điện tử từ máy chủ thư cục bộ. Nó là một giao thức chủ/khách trong đó thư điện tử được nhận và duy trì bởi máy chủ thư.

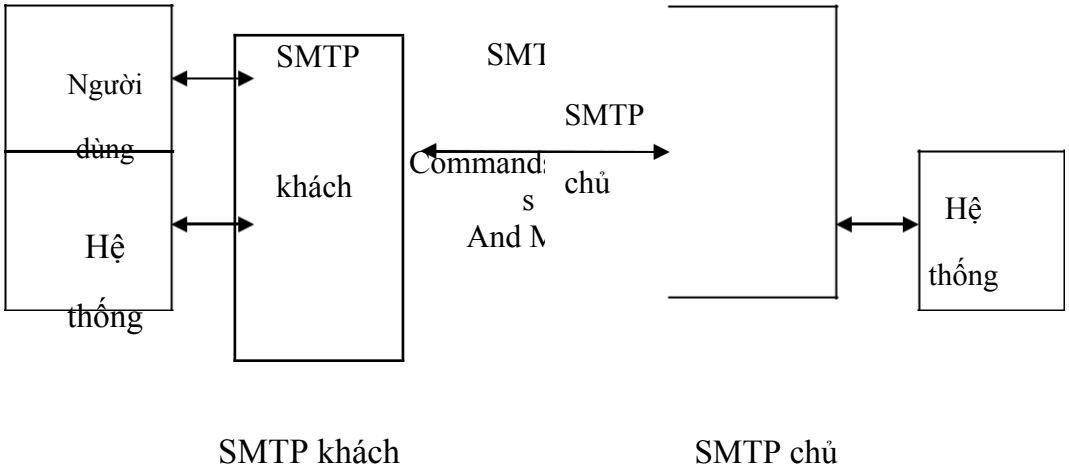
Với những yêu cầu này chỉ một trao đổi dữ liệu nhỏ làm việc tốt thậm chí qua một kết nối chậm như là một modem. Chỉ khi người dùng yêu cầu đọc một thư điện tử cụ thể thì nó sẽ

được tải về từ máy chủ thư đó. Người dùng có thể tạo và chế tác các thư mục hoặc các hộp thư trên máy chủ, xóa những tin nhắn…

1.2 Giao thức SMTP

1.2.1 Giới thiệu cơ bản về SMTP

SMTP (Simple Mail Transfer Protocol) là giao thức quy định việc truyền thư chủ yếu trên Internet, được sử dụng như một cơ chế chung cho việc chuyển tải thư điện tử giữa các máy tính với nhau trong giao thức TCP/IP. Khi một tiến trình SMTP thực hiện, SMTP client mở một kết nối TCP tới một tiến trình SMTP server nằm trên một máy chủ ở xa và cố gắng để gửi mail thông qua kết nối. SMTP server lắng nghe một kết nối TCP trên một cổng 25.

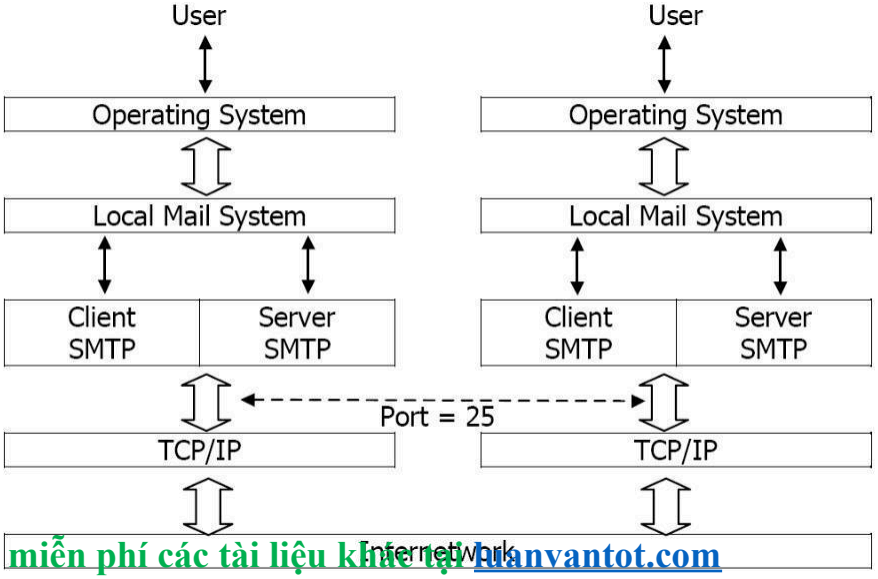


Hình 1.3: Mô hình truyền thọ sử dụng giao thức SMTP

Khi SMTP client có một thông điệp được truyền đi, nó thiết lập một kênh truyền hai chiều tới một SMTP server. Trách nhiệm của SMTP client là chuyển giao những thông điệp thư cho một hoặc nhiều SMTP server (hoặc báo những lỗi sai khi thực hiện).

Hoạt động của giao thức SMTP trong hệ thống thọ điện tử

Mối quan hệ giữa SMTP và hệ thống thư điện tử được trình bày trong hình 1.4:



Hình 1.4: Hoạt động của giao thức SMTP trong hệ thống thơị điện tử

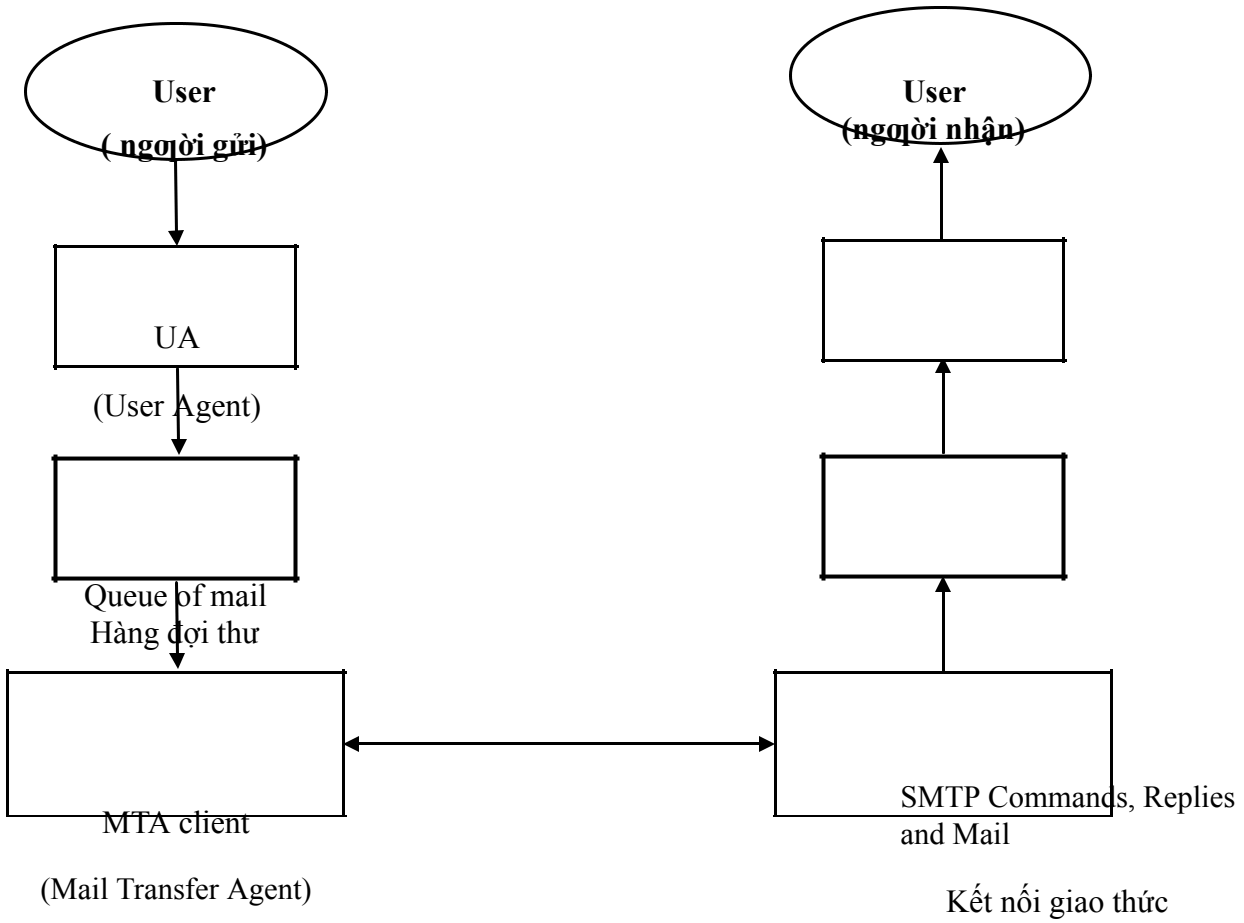
Client liên quan đến thư đi, Server liên quan đến nhận thư. Hệ thống thư cục bộ chứa hộp thư (mailbox) của mỗi user. Mailbox có 2 phần: phần cục bộ và phần toàn cục.

Sau khi tháo bức thư trong khuôn dạng chuẩn, hệ thống thư cục bộ xác định tên người nhận ở hộp thư cục bộ hay phải gửi ra ngoài. Để bức thư được gửi đi, Client SMTP phải biết địa chỉ IP của nơi nhận thông qua DNS và gửi qua cổng địa chỉ SMTP (25) để bắt đầu thiết lập kết nối server SMTP nơi nhận. Khi mối nối đã được thiết lập, Client bắt đầu chuyển thư đến Server bởi các lệnh của SMTP.

1.2.2 Mô hình giao thức SMTP

Mô hình SMTP hỗ trợ cả hai phương pháp truyền phát thư *end-to-end* (không có các MTA trung gian) và *store-and-forward.* Phương pháp *end-to-end* được sử dụng giữa các mạng nội bộ của các tổ chức và phương pháp *store-and-forward* được lựa chọn cho các hệ điều hành giữa các tổ chức có mạng sử dụng giao thức TCP/IP và SMTP cơ sở.

Một tiến trình SMTP cơ bản có thể truyền tải thư điện tử tới một tiến trình khác trên cùng một mạng hoặc tới một mạng khác thông qua một tiến trình truyền tiếp hoặc qua cổng nối có thể tới được cả hai mạng. Một mô hình đơn giản các thành phần của hệ thống SMTP được trình bày trong hình 1.5:



TCP, cổng 25

UA

(User Agent)

User
Mailboxes

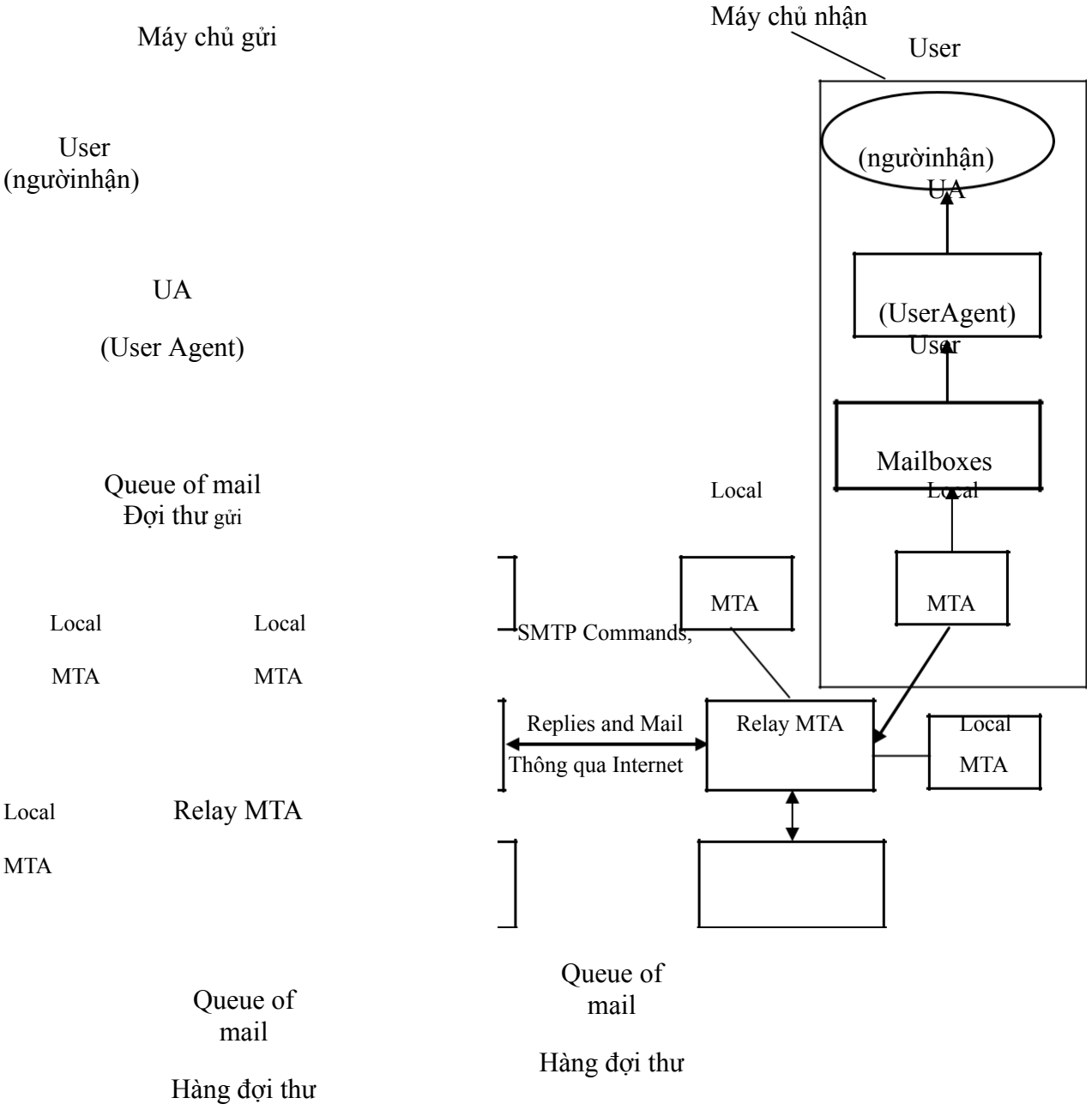
MTA Server

(Mail Transfer
Agent)

Hình 1.5: Mô hình giao thức SMTP

1.2.3 Hệ thống chuyển tiếp thư theo giao thức SMTP

Người dùng làm việc với UA (User Agent). Việc trao đổi thư sử dụng giao thức TCP được thực hiện nhờ một MTA. MTA gửi truyền thư qua mạng tới cổng 25 của giao thức TCP của MTA nhận. Việc truyền thông tin giữa máy chủ gửi và máy chủ nhận ở mạng ngoài thì việc chuyển tiếp có thể phức tạp (xem Hình 1.6). Việc thêm một MTA vào phía người gửi và một MTA vào phía người nhận, các MTA khác thực hiện như máy chủ và máy khách, có thể chuyển tiếp thư điện tử qua mạng.



Hình 1.6: Mô hình SMTP với các MTA chuyển tiếp

Hệ thống các MTA relay cho phép những nơi không sử dụng bộ giao thức TCP/IP để gửi thư điện tử tới những người dùng ở những nơi khác có thể hoặc không thể sử dụng bộ giao thức TCP/IP.

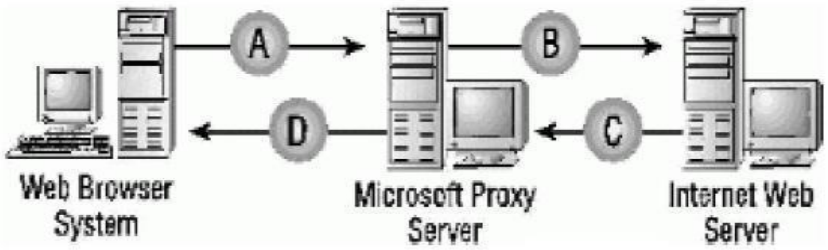
Sự giữ trễ quá tình phân phát thơ

Giao thức SMTP cho phép giữ trễ sự phân phát, và thư điện tử có thể được giữ trễ tại vị trí người gửi, chỗ người nhận, hoặc các máy chủ trung gian.

1.3 Xây dựng Proxy cho SMTP

1.3.1 Khái niệm Proxy

Proxy là một máy chủ trung gian cho phép kết nối từ máy cá nhân (client) tới các máy chủ chạy web trên Internet (theo nghĩa trực tiếp), không cho phép các packet (gói tin) đi trực tiếp giữa hệ thống sử dụng và Internet mà phải đi gián tiếp thông qua các dual home host hoặc qua sự kết hợp giữa basion host (pháo đài phòng ngự) và screening router.



Hình 1.7: Mô hình proxy

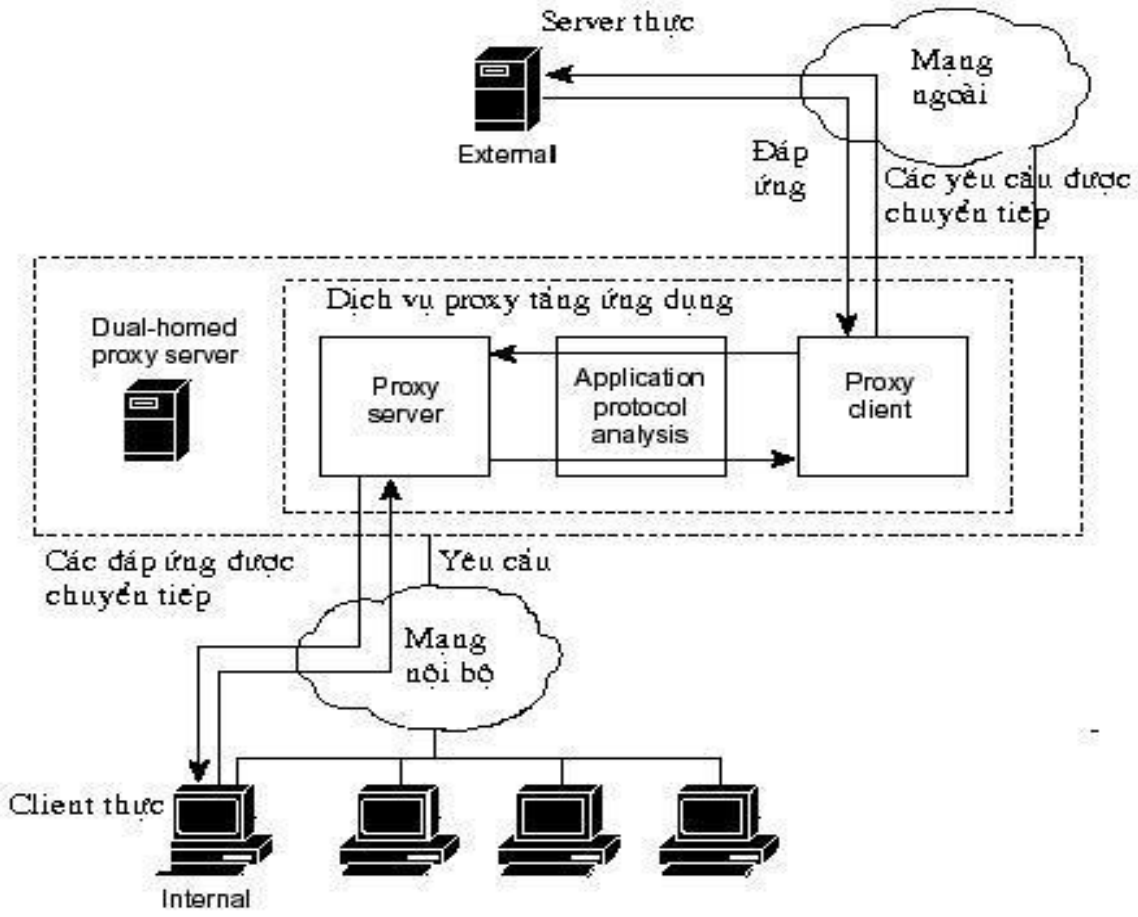
Các dịch vụ Proxy không cho phép kết nối trực tiếp, chúng buộc tất cả các gói tin trên mạng phải được kiểm tra và lọc theo quy tắc thích hợp. Thay vì trao đổi thông tin trực tiếp với dịch vụ thực sự, người dùng phải trao đổi thông tin với Proxy server.

1.3.2 Proxy tầng ứng dụng

Proxy tầng ứng dụng là một máy chủ trung gian cho phép kết nối từ máy trạm khai thác dịch vụ (client) tới các máy chủ cung cấp dịch vụ (server), không cho phép các packet (gói tin) đi trực tiếp từ client tới server mà phải đi gián tiếp thông qua các proxy. ProxyServer giống như cầu nối trung gian giữa Client và Server. Khi có yêu cầu kết nối Client đến Server thật, ProxyServer sẽ quyết định đáp ứng hay không đáp ứng.

Một dịch vụ proxy có ba bộ phận riêng biệt, đó là: proxy server, proxy client, và bộ phân tích giao thức. Trong đó, proxy server làm nhiệm vụ chuyển các yêu cầu được chấp nhận của client tới server thực và chuyển đáp ứng của yêu cầu này tới client thực.

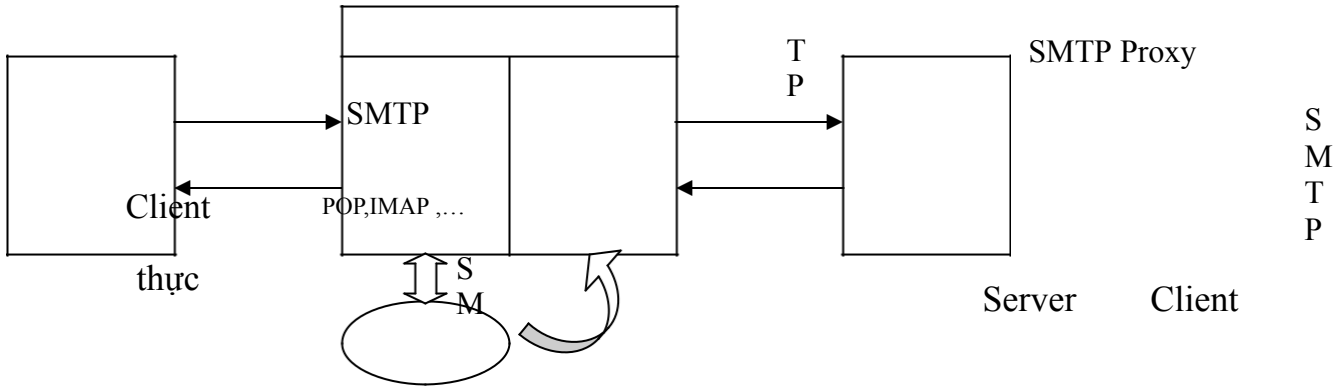
Hình 1.8 trình bày mô hình trao đổi thông tin giữa Client và Server qua Proxy.

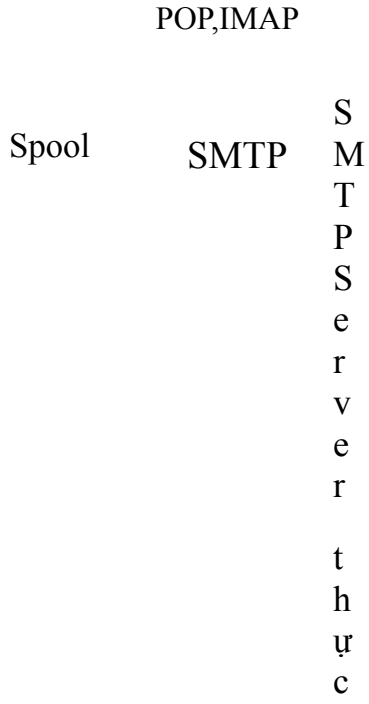


Hình 1.8: Mô hình trao đổi thông tin giữa Client và Server qua Proxy

1.3.3 Xây dựng proxy cho SMTP

SMTP proxy là proxy thuộc lớp ứng dụng sử dụng cho dịch vụ thư tín để kiểm soát thư điện tử được gửi đi qua giao thức SMTP. Nó gồm hai thành phần thực hiện những chức năng cách riêng biệt: Proxy Server và Proxy Client. Hình 1.9 trình bày một mô hình proxy cho giao thức SMTP.





Hình 1.9: Mô hình hoạt động của SMTP Proxy

Chương 2- PHƯƠNG PHÁP VÀ CÁC KỸ THUẬT SỬ DỤNG CHO HỆ THỐNG LỌC NỘI DUNG THƯ ĐIỆN TỬ

2.1 Những thông tin cần kiểm soát của một thư điện tử

Dựa trên mục tiêu của hệ thống lọc là kiểm soát các nội dung của một thư điện tử được gửi đi, những thông tin cần kiểm soát là:

Địa chỉ người gửi (trong trường "*from"*), người nhận (trong trường *"to*");

Chủ đề, tiêu đề (trong trường *"Subject"*);

Các địa chỉ trong trường "CC" và "Bcc": đây là trường gồm nhiều địa chỉ người nhận;

Nội dung của bức thư được viết trong phần *"body"* được người gửi soạn thảo;

- Các file đính kèm (có thể được nén, mã hoá, hoặc để định dạng .doc, docx,txt,pdf….).

2.2 Một số biện pháp kiểm soát nội dung thư điện tử

2.2.1 Lọc thư rác

Thư rác (spam) là thư điện tử, tin nhắn được gửi đến người nhận mà người nhận đó không mong muốn hoặc không có trách nhiệm phải tiếp nhận theo quy định của pháp luật.

Vấn đề thư rác là vấn đề gây nhức nhối trong xã hội trong những năm gần đây. Nhiều công trình nghiên cứu về phương pháp lọc thư rác đã được đầu tư và tiến hành từ khá lâu. Để đánh giá hiệu quả của một công cụ lọc thư rác người ta thường dựa trên hai độ đo sau:

False Positive – Tỷ lệ thư thường bị lọc nhầm thành thư rác.

False Negative – Tỷ lệ thư rác bị lọc nhầm thành thư thường.

Tất cả những công cụ lọc có giá trị ngày nay thường sử dụng một trong số những phương pháp hoặc kết hợp của các phương pháp sau:

Phương pháp lọc Bayesian

Phương pháp lọc SpamAssassin

Phương pháp dùng danh sách trắng/đen

Phương pháp lọc thư rác dùng chuỗi hỏi đáp (Challenge/Response filters)

Phương pháp lọc dựa vào vị trí của các bộ lọc (Filter Placement)

Phương pháp lọc dựa trên xác nhận danh tính của người gửi

2.2.2 Lọc theo từ khóa

Phương pháp lọc theo từ khóa là một phương pháp truyền thống trong việc lọc thư. Người ta dựa vào những từ hay cụm từ có trong đầu đề của thư (subject) và nội dung của thư để lọc. Phương pháp này có ưu điểm là:

- Tính thích nghi: Người dùng có thể dễ dàng biến đổi bộ lọc và thiết lập chính sách lọc theo mục đích của mình để nó có thể lọc ra những thư có nội dung cần quan tâm để phục vụ mục đích của người dùng.

- Tính mở rộng: phương pháp này cũng có thể sử dụng cho việc lọc thư rác. Khi một thư mới được gửi đi, các từ hay cụm từ này sẽ xác định đó là thư rác hay không.

Phương pháp này đặc biệt hiệu quả với mục đích kiểm soát nội dung thông tin, địa chỉ của thư nhằm phục vụ mục đích của người dùng. Ví dụ một số thư điện tử có những nội dung mang tính chất phản động như sau: "bản chất cướp của cộng đảng Việt Nam"… Những nội dung này cần được kiểm soát, ngăn chặn và tìm ra nguồn gốc người phát tán, cũng như được gửi đến người nhận nào.

2.3 Một số thuật toán sử dụng so sánh chuỗi (multi matching)

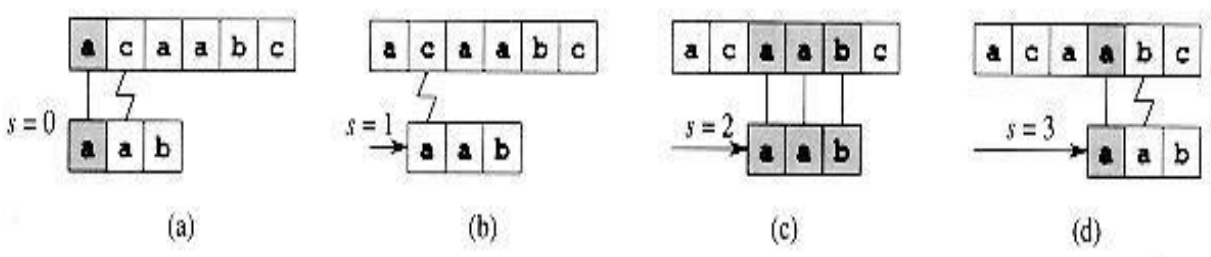
2.3.1 Khái niệm so sánh chuỗi

So sánh chuỗi là quá trình tìm kiếm một hoặc nhiều vị trí xuất hiện của mẫu (pattern), tìm kiếm từ khóa trong văn bản. Phần này sẽ tìm hiểu và so sánh ba giải thuật cơ bản trong so sánh chuỗi bao gồm Naïve, Knutt – Morris – Pratt và Boyer – Moore. Kỹ thuật này được sử dụng để lọc nội dung theo phương pháp lọc từ khóa.

2.3.2 Một số thuật toán so sánh chuỗi

2.3.2.1 Thuật toán Naïve

Đây là giải thuật cơ bản và đơn giản nhất, sử dụng nguyên lý vét cạn để kiểm tra tất cả các khả năng xuất hiện của một chuỗi ký tự có trong văn bản.

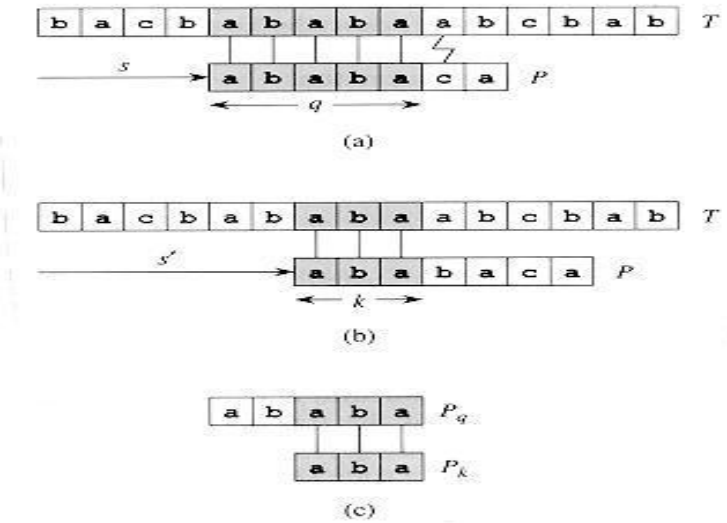


Hình 2.1: Minh họa thuật toán Naïve String Matcher

Thuật toán Knutt – Morris – Pratt được trình bày dưới đây được cải thiện, tốt ơn nhiều so với Naïve vì tận dụng được các thông tin hữu ích khi tìm kiếm.

2.3.2.2 Thuật toán Knuth – Morris – Pratt

Thuật toán Knuth-Morris-Pratt được xây dựng dựa trên thuật toán Naïve với ý tưởng lợi dụng lại những thông tin của lần thử trước cho lần sau.



Hình 2.2: Minh họa hàm tiền tố ∏

(a) *So sánh mẫu P = ababaca với văn bản T, cho ra kết quả 5 ký tự đầu tiên được khớp nối thành công. (b)Với những kết quả đã biết về văn bản và 5 ký tự được ghép nối, có thể thấy độ dịch chuyển s + 1 sẽ không hợp lệ, nhưng s' = s + 2 có khả năng hợp lệ.*

(c) *Thông tin hữu ích này sẽ giúp giảm số lần thử, và nó có thể được tính toán bằng cách so sánh chính các ký tự trong chuỗi mẫu.*

Mặc dù đã giảm được số lần so sánh trên mỗi ký tự, thuật toán KMP vẫn cố gắng kiểm tra mỗi ký tự một lần. Do đó, thời gian xử lý chưa được cải thiện nhiều. Điều quan trọng là phải giảm bớt số ký tự cần so khớp. Đây cũng chính là ý tưởng của thuật toán Boyer – Moore được trình bày tiếp theo.

2.3.2.3 Thuật toán Boyer – Moore

Thuật toán Boyer - Moore là thuật toán tìm kiếm chuỗi rất có hiệu quả trong thực tiễn, hoạt động theo quy tắc kiểm tra các ký tự của mẫu từ phải sang trái và khi phát hiện sự khác nhau đầu tiên, thuật toán sẽ tiến hành dịch chuyển.

Bằng cách loại bỏ tối đa số ký tự cần so sánh, tốc độ xử lý của thuật toán BM đã được tăng lên đáng kể. Tuy nhiên, khi độ dài của văn bản lớn và số lượng văn bản nhiều thì quá trình tìm kiếm sẽ rất tốn kém. Trên thực tế, rất nhiều nghiên cứu đề xuất thuật toán mới và cải tiến thuật toán cũ đã được tiến hành nhằm giải quyết bài toán so sánh đa mẫu. Bên cạnh các thuật toán so sánh đa mẫu điển hình như Aho – Corasick, Wu – Manber (WB) được cho là thuật toán tối ưu hơn cả trong quá trình tìm kiếm và so sánh chuỗi. Phần tiếp theo sẽ trình bày chi tiết hơn về thuật toán này.

2.3.2.4 Thuật toán Wu – Manber

a) *Giới thiệu thuật toán:*

Thuật toán WM được xây dựng với hai cơ chế lõi, đó là cơ chế lọc dựa trên công nghệ băm và cơ chế dịch chuyển khối ký tự dựa trên công nghệ dịch chuyển ký tự xấu của thuật toán Boyer – Moore.

Để phục vụ việc tính toán độ dịch chuyển, giai đoạn tiền xử lý sẽ xác định kích thước của cửa sổ đối sánh; thông thời thiết lập ba bảng dữ liệu cơ bản gồm bảng SHIFT lưu trữ khoảng cách dịch chuyển của các khối ký tự xuất hiện trong văn bản; bảng HASH chứa danh sách liên kết các mẫu có chuỗi hậu tố giống nhau bên trong cửa sổ đối sánh; bảng PREFIX chứa danh sách liên kết các mẫu có chuỗi tiền tố giống nhau bên trong cửa sổ đối sánh.

Quá trình so khớp mẫu được thực hiện thông qua tính toán giá trị băm, thiết lập danh sách những mẫu có chuỗi hậu tố giống nhau, sau đó so sánh với giá trị băm của khối ký tự bên trong cửa sổ đối sánh hiện thời. Khi một chuỗi mẫu được so khớp, cửa sổ tiếp tục được dịch chuyển sang phải. Công nghệ dịch chuyển ký tự xấu được ứng dụng để dịch chuyển cửa sổ đối sánh, tuy nhiên ở đây sẽ dịch chuyển theo khối ký tự, thay vì từng ký tự đơn lẻ như trong Boyer – Moore.

Nhằm khắc phục ảnh hưởng của các mẫu ngắn, thuật toán High Concurrence Wu – Manber (HCWM) đã được đề xuất với nhiều cải tiến đáng kể. Bằng cách tiến hành chia tất cả mẫu thành nhiều bộ mẫu khác nhau tùy theo độ dài của chúng; thuật toán HCWM sau đó lần lượt xử lý các bộ mẫu theo từng cách khác nhau. Bên cạnh đó, cấu trúc dữ liệu độc lập được sử dụng cho các bộ mẫu khác nhau đã tạo ra sự tương tranh cao, có thể đồng thời xử lý

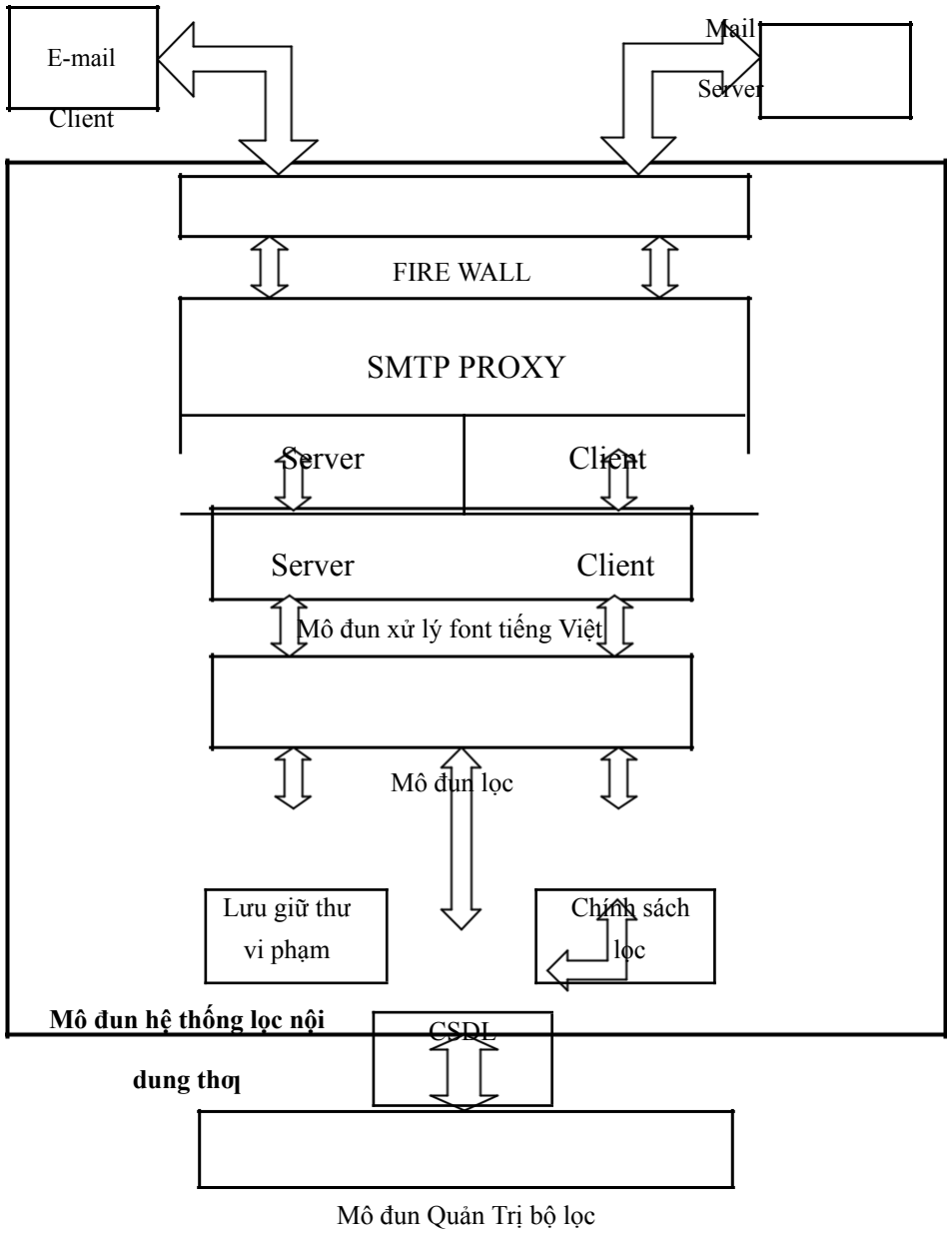
các bộ mẫu, nhờ đó tăng cường tốc độ so sánh mẫu của HCWM, làm cho thuật toán này đạt hiệu suất cao hơn nhiều so với WM.

Chương 3- LỌC NỘI DUNG CHO THƯ ĐIỆN TỬ GỬI THEO GIAO THỨC SMTP

Để có thể kiểm soát được việc gửi thư điện tử qua giao thức SMTP, cần phải có một hệ thống lọc nội dung thư nằm giữa đường đi của máy gửi thư và các mail server. Chính sách lọc thư dựa trên việc kiểm tra các phần *envelope, content* và các file đính kèm của thư điện tử để phát hiện những thư vi phạm chính sách bảo đảm an ninh an toàn thông tin.

3.1 Mô hình hệ thống lọc nội dung Thư điện tử qua giao thức SMTP

Mô hình một hệ thống lọc thư điện tử được trình bày trong Hình 3.2.



Hình 3.1: Mô hình hệ thống lọc nội dung thư điện tử sử dụng giao thức SMTP

Bộ lọc thư điện tử được gửi đi theo giao thức SMTP gồm các mô đun chính sau: mô đun Firewall, mô đun SMTP proxy, mô đun xử lý font tiếng Việt, mô đun lọc và mô đun quản trị.

3.1.1 Mô đun Firewall

Mô đun này thực hiện chức năng định hướng lại gói tin của thư điện tử được gửi từ máy Client tới một SMTP server theo cổng 25. Các gói tin có địa chỉ cổng đích TCP là 25 sẽ được định hướng lại để đi tới mô đun SMTP proxy server với một cổng đích đã được xác định.

3.1.2 Mô đun SMTP proxy

Khi hệ thống chưa có cơ chế lọc nội dung, mô đun SMTP proxy có chức năng như một Proxy bình thường nhận thư từ phía máy người gửi rồi chuyển tiếp thư tới máy mail server để truyền phát thư tới địa chỉ đích.

Khi hệ thống có cơ chế lọc nội dung, thì hoạt động của các thành phần của Proxy có sự thay đổi. Các thành phần này sẽ đóng vai trò như người trung gian tiếp nhận các yêu cầu từ máy Client cũng như tiếp nhận các yêu cầu từ Mail server để chuyển đến mô đun xử lý font tiếng Việt.

3.1.3 Mô đun xử lý font tiếng Việt

Khi nhận thư từ SMTP proxy server gửi đến, mô đun này sẽ phân tích phần *envelope* và *content* của bức thư. Một thư được người dùng tạo ra và được gửi đi nó được viết dưới nhiều dạng font chữ khác nhau. Hiện nay, có nhiều cách mã hóa các kí tự tiếng Việt khác nhau, dẫn tới có nhiều bảng mã khác nhau được sử dụng. Theo thống kê, có tới trên 40 bảng mã tiếng Việt khác nhau được sử dụng, do đó, việc khai thác tài liệu cũng như xử lý dữ liệu rất phức tạp.

Tuy nhiên, hiện nay việc sử dụng tiếng Việt trên máy tính vẫn chưa có sự thống nhất cao về chuẩn mã tiếng Việt, gây khó khăn lớn cho việc thu thập, khai thác và xử lý tiếng Việt. Đòi hỏi các hệ thống xử lý văn bản tiếng Việt cần phải có bước tiền xử lý để nhận dạng và quy chuẩn các ký tự về một bảng mã chung.

Mô đun xử lý font tiếng Việt có chức năng:

+ Chuyển các từ khóa được mã hóa dưới nhiều dạng font chữ khác nhau về một một dạng quy chuẩn các ký tự về một bảng mã chung được hệ thống thiết lập từ trước, sau đó chuyển tới mô đun lọc để thực hiện việc lọc thư.

+ Chuyển đổi mã của nội dung bức thư về định dạng ban đầu của người dùng soạn thảo để chuyển tiếp cho SMTP proxy client để gửi thư đi đến địa chỉ nhận.

3.1.4 Mô đun lọc

Đây là phần quan trọng nhất của hệ thống lọc nội dung thư điện tử. Khi nội dung bức thư được *mô đun xử lý font tiếng Việt* thực hiện, *mô đun lọc* sẽ sử dụng các kỹ thuật tìm kiếm và so sánh chuỗi (multi matching) để thực hiện đối sánh các phần địa chỉ gửi đi, địa chỉ nhận thư, và các từ khóa của phần tiêu đề bức thư và trong nội dung với các từ và cụm từ khóa cấm có trong cơ sở dữ liệu của bộ lọc (*từ điển lọc*) để kiểm tra những nội dung đó có trùng khớp hay không? Nếu một trong các thông tin đối sánh trùng nhau thì bộ lọc sẽ thực hiện việc ghi lại nhật ký và thực hiện chính sách lọc.

3.1.5 Mô đun Quản trị

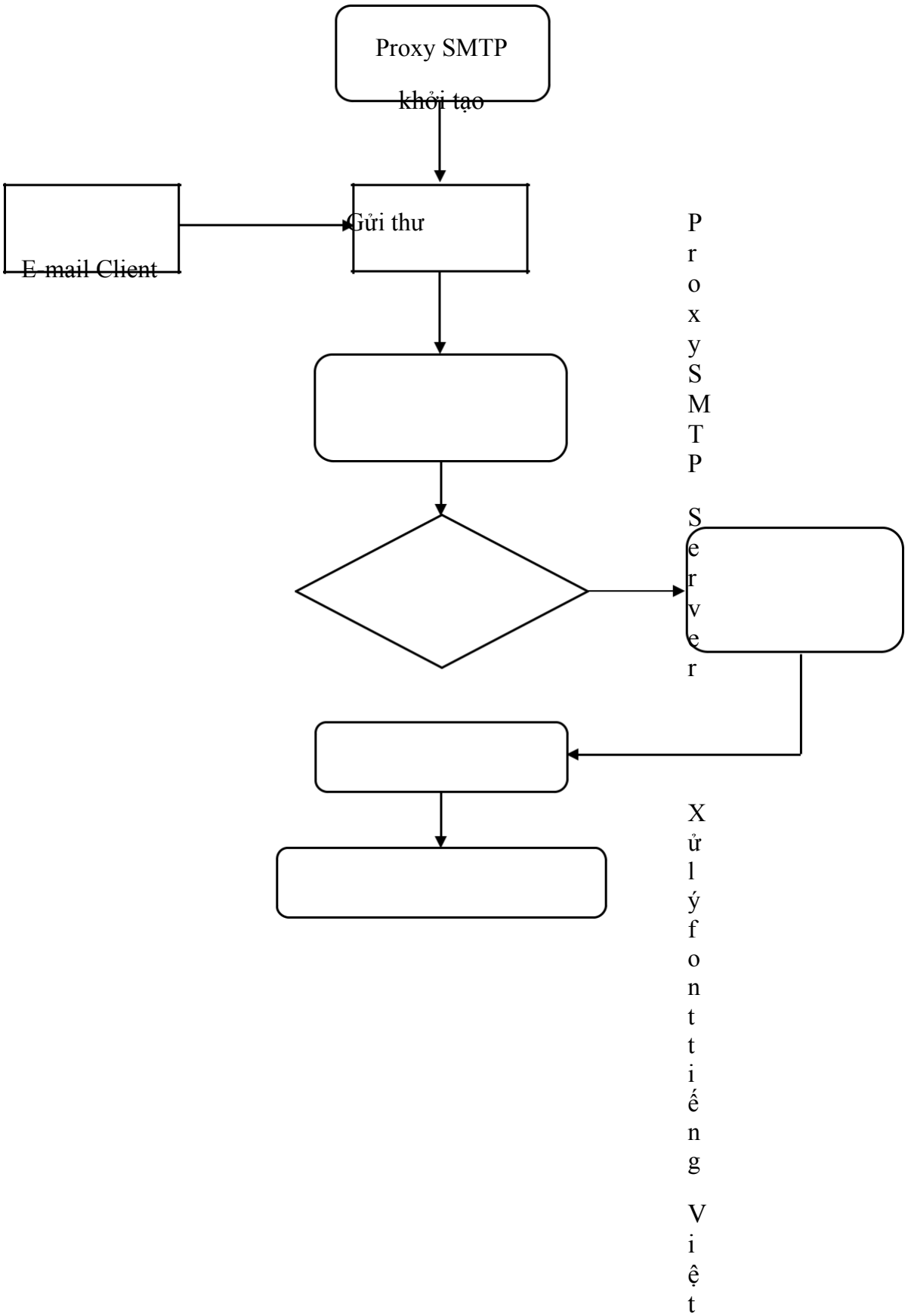
Cho phép người quản trị hệ thống điều khiển bộ lọc, thiết lập các chính sách lọc và các công cụ lọc như: thiết lập các địa chỉ cần lọc, các từ khóa cần lọc, kết xuất nhật ký vi phạm, lưu giữ thư vi phạm v.v...

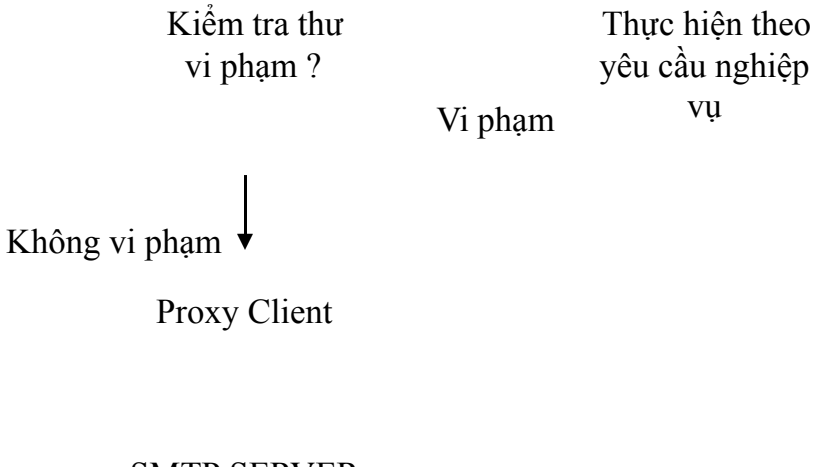
3.2 Lợu đồ hoạt động của hệ thống lọc thợ điện tử

3.2.1 Lưu đồ tiếp nhận và gửi một thư không vi phạm

Lưu đồ tiếp nhận và gửi một thư không vi phạm được biểu diễn ở Hình 3.2.

Khi SMTP proxy khởi tạo kết nối, máy mail client sẽ gửi yêu cầu kết nối tới SMTP proxy server. Sau khi SMTP proxy server xác nhận và cho phép mail client gửi thư, hệ thống sẽ tiến hành xử lý font tiếng Việt. Tiếp theo, hệ thống lọc thư sẽ kiểm tra thư gửi đi có vi phạm chính sách lọc hay không. Nếu thư được gửi không vi phạm chính sách lọc, thư đó sẽ được chuyển ngay tới Proxy SMTP client.



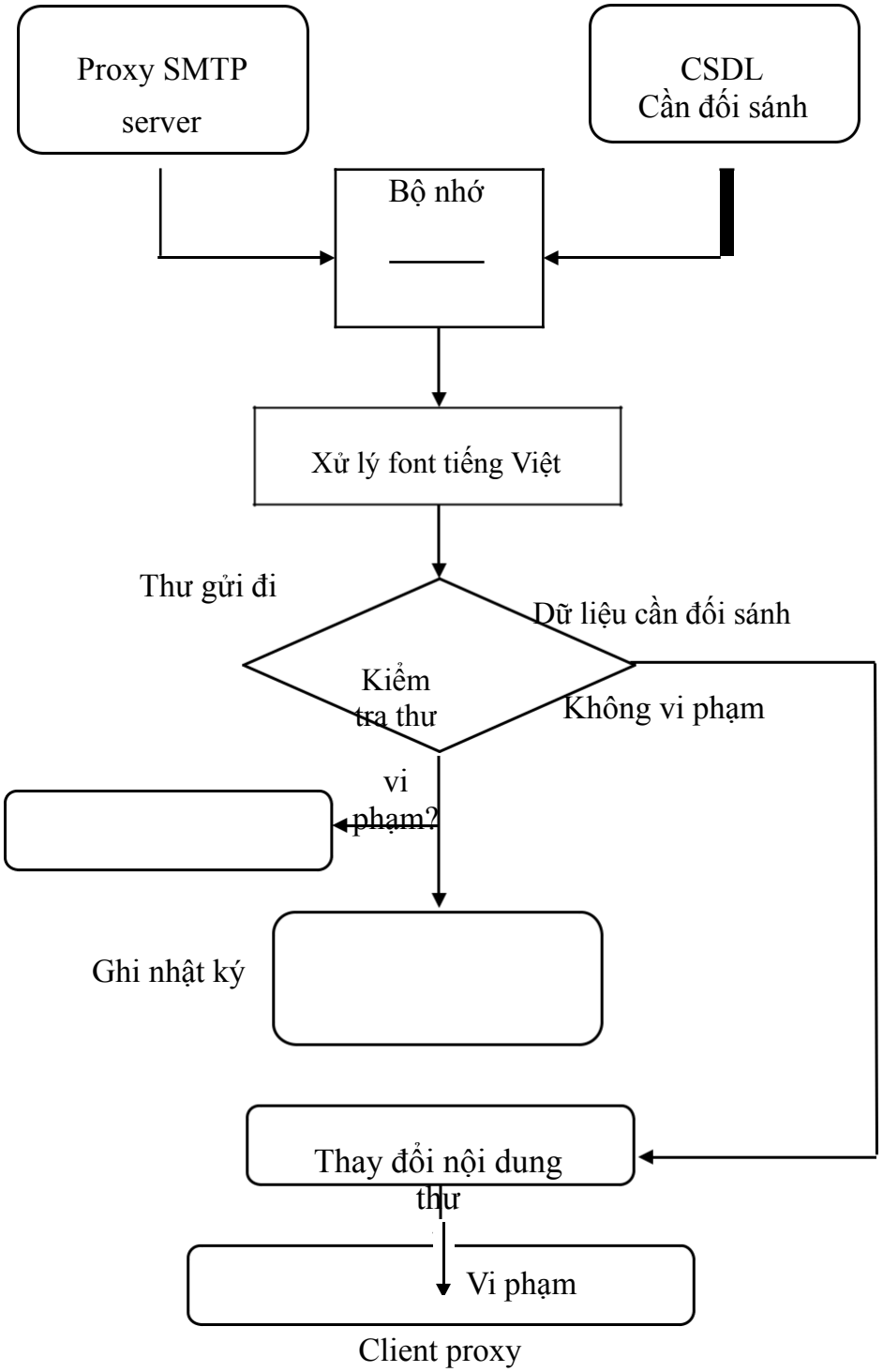


Hình 3.2: Lợu đồ tiếp nhận và gửi thợ điện tử không vi phạm

Ngược lại, nếu thư vi phạm chính sách lọc thì thư đó sẽ được xử lý theo các chính sách vi phạm được trình bày ở phần 3.2.2.

3.2.2 Lưu đồ xử lý thư vi phạm

Hình 3.3 thể hiện các bước xử lý đối với một thư vi phạm chính sách an ninh an toàn thông tin.



Hình 3.3: Lợu đồ xử lý thợ vi phạm

Thư được SMTP proxy tiếp nhận từ mail client sẽ được lưu vào một vùng nhớ. Tại đây hệ thống lọc sẽ xử lý font tiếng Việt trước khi lấy địa chỉ và từ khóa cần ngăn chặn có

trong CSDL của hệ thống lọc để đối sánh với địa chỉ và nội dung của thư (như đã nói ở phần trên). Khi có sự trùng khớp thì thư đó đã vi phạm chính sách lọc. Hệ thống sẽ ghi lại nhật ký vi phạm, đồng thời thay đổi nội dung thư rồi chuyển tiếp thư đó tới Proxy client để chuyển thư tới SMTP server.

3.2.3 Thiết lập chính sách lọc

Chính sách lọc thư điện tử được thiết lập theo các tiêu chí: lọc theo địa chỉ và lọc theo từ khóa.

Lọc theo địa chỉ bao gồm: lọc địa chỉ người gửi và địa chỉ người nhận.

Lọc theo địa chỉ người gửi.

Lọc theo địa chỉ nhận gồm có các trường: *To:* ; *Cc:* ; *Bcc:* . Nếu địa chỉ gửi đi trùng với địa chỉ có trong CSDL của hệ thống lọc cần chặn thì bộ lọc sẽ thay đổi trường địa chỉ người nhận bằng trường địa chỉ người gửi đồng thời thay nội dung bức thư bằng thông báo "Thư đã vi phạm an ninh an toàn thông tin".

Lọc theo từ khóa được thực hiện lọc từ tiêu đề bức thư (subject) và phần nội dung của bức thư. Những từ cần kiểm soát sẽ được thiết lập trong từ điển lọc, hệ thống sẽ tiến hành so sánh với nội dung thư, nếu vi phạm sẽ bị hệ thống ngăn chặn hoặc gửi thông báo để người quản trị có biện pháp xử lý kịp thời.

3.3 Cài đặt hệ thống và hoạt động của phần mềm lọc thọ điện tử

3.3.1 Cài đặt Firewall

3.3.2 Cài đặt hệ thống lọc thư điện tử

Để cài đặt hệ thống lọc thư điện tử cần thao tác theo thứ tự các bước như sau:

Bước 1: Chép bộ cài đặt *emailrelay.tar* vào trong thư mục */opt* dùng lệnh */tar –xvzf*

emailrelay.tar để giải nén bộ cài đặt vào trong thư mục được lựa chọn để cài đặt cho bộ lọc mail. Sửa file cấu hình của bộ lọc theo chính sách lọc và cổng lọc đã được ấn định. Trong đó:

Thư mục */usr/local/emailrelay/etc* chứa file cấu hình của chương trình.

Thư mục */usr/local/emailrelay/libexec* chứa các file thư viện và file chạy của chương trình.

Thư mục */usr/local/emailrelay/tmp* chứa các file nhật ký.

Thư mục */usr/local/emailrelay/var/spool* lưu giữ thư tạm thời, chờ hệ thống chuyển đi.

Bước 2: Kích hoạt hệ thống lọc thư bằng việc đánh lệnh:

/usr/local/emailrelay/libexec/emailrelay/init/emailrelay start

Bước 3: Kiểm tra hoạt động của hệ thống: *ps –ef | grep emailrelay*

KẾT LUẬN

Lọc nội dung thư điện tử theo giao thức SMTP nhằm mục đích kiểm soát nội dung thư điện tử được gửi trên Internet là một hướng mới trong việc đảm bảo an ninh an toàn thông tin. Hướng tiếp cận đề tài đặc biệt hiệu quả trong việc thu thập thông tin, theo dõi đối tượng phục vụ cho các hoạt động nghiệp vụ của lực lượng Công an. cũng như góp phần không nhỏ trong việc hạn chế việc phát tán tài liệu, tuyên truyền, kích động và nói xấu Đảng, lãnh đạo, và Nhà nước. Đề tài cũng đáp ứng được một phần yêu cầu thực tiễn đặt ra về vấn đề kiểm soát nội dung thông tin trên Internet.

Khóa luận đã hệ thống hóa được một số vấn đề lý thuyết về hệ thống thư điện tử, giới thiệu một số biện pháp kiểm soát nội dung thư điện tử, một số phương pháp được ứng dụng phổ biến. Tuy nhiên với mục tiêu đề tài đặt ra thì việc sử dụng phương pháp lọc từ khóa sử dụng kỹ thuật multi matching đã thu được các kết quả thực sự khả quan và đáp ứng được nhu cầu rất cấp thiết hiện nay đối với một số vấn đề cần quan tâm.

Tuy nhiên, do thời gian nghiên cứu có hạn chương trình lọc mới chỉ thực hiện được các chức năng lọc cơ bản dựa theo cấu trúc của một bức thư điện tử như lọc theo các trường địa chỉ người nhận, người gửi, tiêu đề thư và nội dung thư theo từ khóa tiếng việt.

Với các kết quả đã đạt được, đề tài đang tiến hành thử nghiệm tại phòng thí nghiệm An ninh an toàn thông tin, Cục Tin học nghiệp vụ, Bộ Công an. Nếu có điều kiện, thời gian và được sự cho phép, tác giả mong muốn được nghiên cứu phát triển hoàn thiện các chức năng lọc của chương trình như lọc nội dung các tệp đính kèm với các định dạng khác nhau Trong thời gian tới sẽ tiến hành thử nghiệm tích hợp với các mô đun lọc khác như: lọc thư điện tử sử dụng giao thức khác như: giao thức POP, IMAP; lọc nội dung trang web, kiểm tra lỗ hổng an ninh trang web… để đưa hệ thống vào hoạt động trong thực tế.