ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

Khamchan PHOMTHAVONG

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN LỌC THƯ RÁC VÀ ỨNG DỤNG TRONG LỌC EMAIL NỘI BỘ

Chuyên ngành: Khoa học máy tính
Mã số: 8 48 0101

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: TS. NGUYỄN HẢI MINH

THÁI NGUYÊN – 2019

LỜI CẢM ƠN

Để hoàn thành chương trình cao học và viết luận văn, tôi đã nhận được sự hướng dẫn, giúp đỡ góp ý nhiệt tình của quý thầy cô trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên.

Trong quá trình học tập và rèn luyện tại trường Đại học Công nghệ Thông tin và Truyền thông – Đai học Thái Nguyên, đến nay em đã kết thúc khóa học 2 năm và hoàn thành luận văn tốt nghiệp. Để có được kết quả này em xin chân thành cảm ơn:

Ban Giám hiệu trường Đại học Công nghệ Thông tin và Truyền thông cùng các thầy, cô giáo trong trường đã giảng dạy, quan tâm và điều kiện thuận lợi để chúng em học tập và rèn luyện trong suốt thời gian theo học tại trường.

TS. Nguyễn Hải Minh người đã tận tình hướng dẫn, chỉ bảo, giúp đỡ em trong suốt quá trình làm luận văn.

Và cuối cùng tôi xin gửi lời cảm ơn tới các đồng nghiệp, gia đình và bạn bè những người đã ủng hộ, động viên tạo mọi điều kiện giúp đỡ để tôi có được kết quả như ngày hôm nay.

Thái Nguyên, tháng .... năm 2019

Học viên

Khamchan PHOMTHAVONG

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

MỞ ĐẦU

Mạng Internet ra đời đã mang lại cho con người những tiên ích hết sức to lớn và quan trọng, một trong những tiện ích đó là dịch vụ thư điện tử. Vì, đó là phương tiện giao tiếp đơn giản, tiện lợi, rẻ và hiệu quả giúp mọi người gắn kết và liên lạc với nhau thường xuyên hơn. Tuy nhiên, lợi dụng tính mở của công nghệ và cơ chế trao đổi thư mà hàng ngày người dùng nhận được một số thư ngoài mong đợi đó là thư rác (Spam). Thư rác thường được gửi với số lượng rất lớn thường vì mục đích quảng cáo, thậm trí là đính kèm mã độc dưới dạng Virus gây phiền toài cho người dùng, làm giảm tốc độ xử lý của máy chủ mail server.

Thư rác (spam) là thư điện tử được gửi hàng loạt với nội dung mà người nhận không mong đợi, không muốn xem, hay chứa những nội dung không liên quan đến người nhận và thường được sử dụng để gửi thông tin quảng cáo. Do có giá thành tương đối thấp so với các phương pháp quảng cáo khác, thư rác hiện chiếm một tỷ lệ lớn và ngày càng tăng trong tổng số thư điện tử được gửi qua Internet. Sự xuất hiện và gia tăng thư rác không những gây khó chịu và làm mất thời gian của người nhận mà còn ảnh hưởng tới đường truyền Internet và làm chậm tốc độ xử lý của máy chủ thư điện tử, gây thiệt hại lớn về kinh tế.

Xuất phát từ lý do đó, đề tài đặt vấn đề nghiên cứu một số thuật toán LỌC THƯ RÁC, một trong những thuật toán mới được công bố gần đây để đề xuất một mô hình thực nghiệm trên một dịch vụ email thực tế. Qua đó hướng tới xây dựng ứng dụng bằng cách tích hợp thêm một số Module trong hỗ trợ sử dụng dịch vụ sử dụng email.

Nội dung luận văn gồm có 3 chương:

Dự kiến nội dung báo cáo của luận văn gồm: Phần mở đầu, 3 chương chính, phần kết luận, tài liệu tham khảo, phụ lục. Bố cục được trình bày như sau:

Phần mở đầu: Nêu lý do chọn đề tài và hướng nghiên cứu chính

Chương 1: THƯ RÁC VÀ CÁC PHƯƠNG PHÁP LỌC THƯ RÁC

Chương 2: TỔNG QUAN CÁC THUẬT TOÁN NSA, PSA, PNSA TRONG LỌC THƯ RÁC.

Chương 3: CÀI ĐẶT CÁC THUẬT TOÁN.

Phần kết luận: Tóm tắt các kết quả đã đạt được và hướng phát triển tiếp theo của đề tài.

Chương 1.
THƯ RÁC VÀ CÁC PHƯƠNG PHÁP LỌC THƯ RÁC

1.1. Một số khái niệm cơ bản

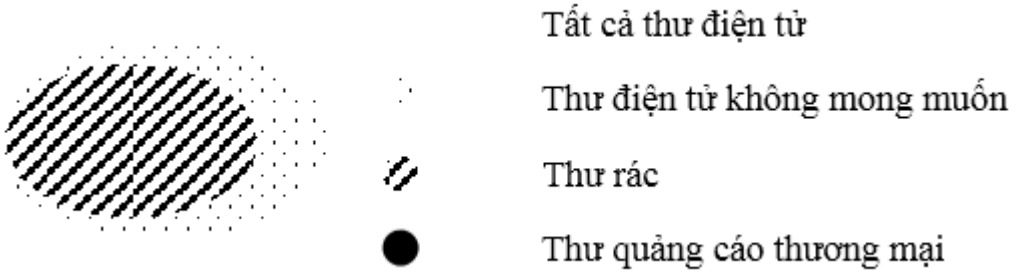
1.1.1. Định nghĩa thư rác

Có nhiều tranh cãi về việc đâu là định nghĩa chính xác của thư rác (spam email), bởi vì thư rác mang tính cá nhân hóa nên khó mà nói lên được hết ý nghĩa của thư rác. Nhiều ý kiến cho rằng thư rác là những "thư điện tử (email) không mong muốn". Định nghĩa này cũng không thực sự chính xác, như một nhân viên nhận những thư điện tử về công việc từ sếp của họ, đây là những thư điện tử người nhân viên không mong muốn nhưng chúng không phải là thư rác. Lại có ý kiến khác cho rằng thư rác là những "thư điện tử thương mại không được yêu cầu từ phía người nhận" những thư này bao gồm các thư điện tử quảng cáo về các sản phẩm và thư điện tử lừa gạt. Nhưng định nghĩa này cũng không thực sự chính xác, nó làm mọi người nghĩ rằng thư rác giống như là thư đáng bỏ đi (junk mail). Sau đây sẽ đưa ra một định nghĩa thông dụng nhất về thư rác và giải thích các đặc điểm của nó để phân biệt thư rác với thư thông thường [1,2]:

Thư rác (spam mail) là những bức thư điện tử không yêu cầu, không mong muốn và được gửi hàng loạt tới người nhận.

Một bức thư nếu gửi không theo yêu cầu có thể đó là thư làm quen hoặc thư được gửi lần đầu tiên, còn nếu thư được gửi hàng loạt thì nó có thể là thư gửi cho khách hàng của các công ty, các nhà cung cấp dịch vụ. Vì thế một bức thư bị coi là rác khi nó không được yêu cầu, và được gửi hàng loạt.

Hình vẽ sau sẽ thể hiện rõ định nghĩa của thư rác:



Hình 1.1: Tất cả các thư điện tử

1.1.2. Phân loại thư rác

Có rất nhiều cách phân loại thư rác[1] .

- Dựa trên kiểu phát tán thư rác: Tính tới thời điểm hiện tại, thư rác có thể bị gửi thông qua thư điện tử, nhóm thảo luận (newsgroups), điện thoại di động (Short Message Service - SMS) và các dịch vụ gửi tin nhắn trên mạng (như Yahoo Messenger, Windows Messenger...)

- Dựa vào quan hệ với người gửi thư rác: bao gồm người lạ mặt, bạn bè, người quen và các dịch vụ quyên góp giúp đỡ…

- Dựa vào nội dung của thư rác: các kiểu nội dung phổ biến như thư về thương mại, thư về chính trị, thư về công nghệ, chuỗi thư (chain e-mail) và các loại khác (như thư phát tán virus...).

- Dựa trên động lực của người gửi: Thông thường, thư rác được gửi đi cho những mục đích quảng bá thông tin. Ngoài ra, còn có một số loại thư rác được gửi tới một người nhận xác định nào đó nhằm mục đích phá vỡ và gây cản trở công việc của người nhận hay mạng của nhà cung cấp dịch vụ thư điện tử (ESP) được gọi là “bom thư”. Thư rác còn được cố ý gửi đi nhằm thông báo tin sai lệch, làm xáo trộn công việc và cuộc sống của người nhận.

Sự phân loại thư rác rất quan trọng không chỉ trong lĩnh vực tạo những bộ lọc thư rác có hiệu quả cao mà còn giúp cho việc ban hành các bộ luật chống thư rác phù hợp.

1.2. Các phương pháp lọc thư rác

1.2.1. Lọc thư rác thông qua việc đưa ra luật lệ nhằm hạn chế, ngăn chặn việc gửi thư rác.

Khi tình trạng thư rác ngày càng tăng trên đường truyền internet gây ra nhiều phiền toái và thiệt hại lớn trên thế giới rất nhiều các quốc gia đã đưa ra các luật để ngăn chặn thư rác. Dưới đây là một số nội dung cơ bản liên quan tới giải pháp ngăn chặn thông qua luật lệ pháp lý được đưa ra trên báo điện tử của bộ viễn thông .

Mỹ là một những nước đầu tiên trên thế giới cố gắng ban hành các văn bản pháp luật để giải quyết vấn đề thư điện tử rác tràn ngập. Từ tháng 7 năm 1997, bang Nevada đã dẫn đầu trong việc ban hành các quy phạm pháp luật quy

định về hành vi phục vụ và sử dụng thư tín điện tử. Tính đến tháng 3 năm 2003, đã có 26 bang ban hành quy phạm pháp luật quy định về dịch vụ và hành vi sử dụng thư tín điện tử. Đến tháng 11 năm 2003, con số này lên đến 36. Về phía chính quyền liên bang, từ những năm 1990, cả Thượng nghị viện và Hạ nghị viện đều quan tâm đến sự lan rộng của thư tín điện tử quấy rối và thư rác, và đã đưa ra nhiều dự án luật như "Luật bảo vệ hộp thư không bị quấy rối" (1999), "Luật Bảo vệ người sử dụng thư điện tử", "Luật Khống chế thư điện tử không được phép" (2000), "Luật Khống chế thư rác truyền qua đường điện thoại vô tuyến" (2000) , "Luật Chống thư rác" (2001).

Mười năm gần đây, Liên minh Châu Âu cũng đã ban hành một số chỉ lệnh, đưa ra các quy phạm và chỉ dẫn đối với các vấn đề thương mại điện tử, thông tin điện tử, bảo hộ dữ liệu.

Trong các chỉ lệnh nói trên, có không ít các qui định có liên quan mật thiết, thậm chí là trực tiếp với phục vụ và sử dụng thư điện tử như "Chỉ lệnh Bảo vệ dữ liệu cá nhân ở Châu Âu", "Chỉ lệnh về thông tin điện tử và bảo mật dữ liệu" ... Ngày 12 tháng 7 năm

2002, Nghị Viện Liên minh Châu Âu đã thông qua "Chỉ lệnh Bảo mật riêng tư và Thông tin điện tử trong Liên minh Châu Âu". Chỉ lệnh quy định: Từ 31 tháng 10 năm 2003, trong phạm vi Liên minh Châu Âu, nếu chưa được người nhận đồng ý trước, không được gửi thư điện tử thương mại hay nhằm mục đích tuyên truyền cho cá nhân. Tiếp theo sau khi Liên minh Châu Âu đưa ra các qui định về phục vụ và sử dụng thư điện tử, các nước thành viên Liên minh Châu Âu, như Italia, Anh, Đan Mạch, Tây Ban Nha ... đều đã ban hành quy phạm pháp luật trong nước quy định hành vi cung cấp và sử dụng thư điện tử, ngăn chặn sự tràn ngập của thư rác.

Tại Lào vấn đề thư rác bắt đầu nhận được sự quan tâm từ phía các cơ quan có trách nhiệm. Bộ Thương mại đang soạn thảo Thông tư quản lý hoạt động quảng cáo thương mại trên các phương tiện điện tử. Trên trang báo điện tử của bộ viễn thông, Bà Lại Việt Anh, Trưởng Phòng chính sách, Vụ Thương mại điện tử, Bộ Thương mại, nhận xét: mục tiêu của Thông tư này trước mắt tập trung