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

PHẠM VĂN DƯƠNG

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN LỌC THƯ RÁC VÀ ỨNG DỤNG TRONG LỌC EMAIL NỘI BỘ CỦA VIỄN THÔNG TỈNH BẮC KẠN

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2017

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

PHẠM VĂN DƯƠNG

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN LỌC THƯ RÁC VÀ ỨNG DỤNG TRONG LỌC EMAIL NỘI BỘ CỦA VIỄN THÔNG TỈNH BẮC KẠN

Chuyên ngành: Khoa học máy tính
Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: TS. NGUYỄN HẢI MINH

Thái Nguyên - 2017

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn là công trình nghiên cứu của riêng cá nhân tôi, không sao chép của ai do tôi tự nghiên cứu, đọc, dịch tài liệu, tổng hợp và thực hiện. Nội dung lý thuyết trong trong luận văn tôi có sử dụng một số tài liệu tham khảo như đã trình bày trong phần tài liệu tham khảo. Các số liệu, chương trình phần mềm và những kết quả trong luận văn là trung thực và chưa được công bố trong bất kỳ một công trình nào khác.

Thái Nguyên, tháng 4 năm 2017

Học viên thực hiện

Phạm Văn Dương

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời biết ơn sâu sắc đến TS. Nguyễn Hải Minh người đã tận tình hướng dẫn, chỉ bảo, giúp đỡ em trong suốt quá trình làm luận văn.

Em cũng xin gửi lời cảm ơn đến các thầy cô giáo trường Đại học Công nghệ thông tin và Truyền thông, các thầy cô Viện Công nghệ thông tin đã truyền đạt những kiến thức và giúp đỡ em trong suốt quá trình học của mình.

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 4 năm 2017

Học viên

Phạm Văn Dương

MỤC LỤC

CÁC HÌNH VẼ, BẢNG BIỂU TRONG LUẬN VĂN

MỞ ĐẦU

Mạng Internet ra đời đã mang lại cho con người những tiện 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: KẾT QUẢ 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

Một trong những dịch vụ Internet mang lại đó là dịch vụ thư điện tử, đây là phương pháp giao tiếp rất đơn giản, tiện lợi, rẻ và hiệu quả giữa mọi người. Tuy nhiên, chính vì những lợi ích của dịch vụ thư điện tử mang lại mà số lượng thư trao đổitrên Internet ngày càng tăng và hầu hết trong số những thư đó là thư rác(Email spam). Thư rác thường được gửi với số lượng lớn, người dùng không mong đợi với nhiều mục đích khác nhau như: Quảng cáo, đính kèm virus, gây phiền toái khó chịu cho người dùng, làm giảm tốc độ Internet và tốc độ xử lý của server, gây thiệt hại lớn về kinh tế. Chương này khái quát các vấn đề về thư rác, ảnh hưởng của thư rác trong cuộc sống và các phương pháp ngăn chặn thư rác. Các khái niệm trong chương này được tham khảo trong [1], [2], [3], [4].

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]: