ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

-------o0o-------

PHẠM THỊ KIM DUNG

PHÂN LOẠI THƯ RÁC BẰNG PHƯƠNG PHÁP HỌC MÁY

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái nguyên, 2015

ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

-------o0o-------

PHẠM THỊ KIM DUNG

PHÂN LOẠI THƯ RÁC BẰNG PHƯƠNG PHÁP HỌC MÁY

Chuyên ngành: Khoa học máy tính

Mã số: **60 48 01**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC

PGS. TS ĐỖ TRUNG TUẤN

Thái nguyên, 2015

MỤC LỤC

LỜI CAM KẾT

Dưới sự giúp đỡ nhiệt tình và chỉ bảo chi tiết của giáo viên hướng dẫn, tôi đã hoàn thành luận văn của mình. Tôi xin cam kết luận văn này là của bản thân tôi làm và nghiên cứu, không hề trùng hay sao chép của bất kỳ ai.

Tài liệu được sử dụng trong luận văn được thu thập từ các nguồn kiến thức hợp pháp.

Tác giả luận văn

Phạm Thị Kim Dung

LỜI CẢM ƠN

Để hoàn thành chương trình cao học và viết luận văn này, em đã nhận được sự giúp đỡ và đóng góp nhiệt tình của các thầy cô trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên.

Trước hết, em xin chân thành cảm ơn các thầy cô trong khoa Đào tạo sau đại học, đã tận tình giảng dạy, trang bị cho em những kiến thức quý báu trong suốt những năm học qua.

Xin chân thành cảm ơn gia đình, bạn bè đã nhiệt tình ủng hộ, giúp đỡ, động viên cả về vật chất lẫn tinh thần trong thời gian học tập và nghiên cứu.

Trong quá trình thực hiện luận văn, mặc dù đã rất cố gắng nhưng cũng không tránh khỏi những thiếu sót. Kính mong nhận được sự cảm thông và tận tình chỉ bảo của các thầy cô và các bạn.

DANH MỤC CÁC TỪ VIẾT TẮT

AI	Trí tuệ nhân tạo
Clustering	Phân cụm
Computer Vision	Nhìn máy
ESP	Email Service Provider
HAM	Thư điện tử không là thứ rác
ISP	Internet Service Provider, nhà cung cấp dịch vụ Internet
KNN	K người láng giềng gần nhất
MI	Mutual information, thông tin tương hỗ
NB	Phương pháp Naïve Bayes
Regression	Hồi qui
Search Engine	Máy tìm kiếm
Server	Máy chủ, phía máy chủ
SMO	Sequential Minimal Optimization
SMS	Short Message Service
Spam Email	Thư rác
SQL	Structured Query Language
Stemming	Gốc (của từ)
SVM	Support Vector Machine, máy vec tơ tựa
TTNT	Trí tuệ nhân tạo
UBE	Unsolicited Bulk Email, thư không lành mạnh
UCE	Unsolicited Commercial Email, thư không yêu cầu đến
VC	Kích thước Vapnik- Chervonenkis
XML	eXtensible Markup Language

DANH MỤC HÌNH VẼ VÀ BẢNG BIỂU

HÌNH

BẢNG

MỞ ĐẦU

Ngày nay, Internet mở ra nhiều kênh liên lạc, nhiều dịch vụ mới cho người sử dụng, một trong những dịch vụ mà Internet mang lại là dịch vụ thư điện tử (Email), đó là phương tiện giao tiếp rất đơn giản, tiện lợi và hiệu quả đối với cộng đồng người sử dụng dịch vụ này. Chính vì những lợi ích do thư mang lại nên số lượng thư trao đổi trên Internet ngày càng tăng, và một số không nhỏ trong đó là thư rác (Spam).

Trong những năm gần đây, spam hay các thư không mong muốn đã trở thành một vấn nạn và đe dọa khả năng giao tiếp của con người trên kênh liên lạc này, đó là một trong những thách thức lớn mà khách hàng và các nhà cung cấp dịch vụ phải đối phó. Spam đã trở thành một hình thức quảng cáo chuyên nghiệp, phát tán virus, ăn cắp thông tin với nhiều thủ đoạn và mánh khóe cực kỳ tinh vi. Người dùng sẽ phải mất khá nhiều thời gian để xóa những thư "không mời mà đến", nếu vô ý còn có thể bị nhiễm virus và nặng nề hơn là mất thông tin như thẻ tín dụng, tài khoản ngân hàng qua các thư dạng phishing....

Theo báo cáo tình hình thư rác do Kaspersky Lab vừa công bố, tỷ lệ thư rác trong lưu lượng truy cập thư của quý 3/2014 tăng 1,7 % so với quý trước, đạt trung bình 66,9%. Ba nguồn phát tán thư rác hàng đầu gồm có Mỹ (14%) và Nga (6,1%) và Việt Nam đứng vị trí thứ 3 với 6%.

Để ngăn chặn spam, nhiều tổ chức, cá nhân đã nghiên cứu và phát triển những kỹ thuật phân loại thư thành các nhóm; từ đó xác định, nhận biết giữa thư rác và thư có giá trị. Tuy nhiên, những người tạo nên thư rác luôn tìm mọi cách vượt qua các bộ phân loại này và phát tán chúng. Vì vậy, cần có một hệ thống phân loại đâu là spam mail và đâu là mail tốt. Xuất phát từ thực trạng đó, tôi chọn hướng nghiên cứu "**Phân loại thư rác bằng phương pháp học máy**" với mục đích tìm hiểu, thử nghiệm một số phương pháp tiếp cận cho bài toán phân loại thư, từ đó ngăn chặn thư spam hiệu quả hơn.

Nội dung của luận văn được trình bày theo 3 chương. Tổ chức cấu trúc như sau:

1. Chương 1 Tổng quan về học máy và thư rác: Chương này giới thiệu tổng quát về học máy và thư rác bao gồm khái niệm, ứng dụng và phần trình bày chi tiết về học máy có giám sát, các kỹ thuật của học máy có giám sát dùng cho phân loại như Naïve Bayes, SVM, cây quyết định,… Chương cũng giới thiệu khái quát về thư rác, các đặc trưng của thư rác và biểu diễn thư rác dựa trên học máy có giám sát;
2. Chương 2 Phân loại thư rác bằng một số thuật toán có giám sát: Nội dung chính trong chương này là đi sâu nghiên cứu hai thuật toán học máy có giám sát là Naïve Bayes và phương pháp SVM (Support Vector Machine).
3. Chương 3 Cài đặt, thử nghiệm và đánh giá thuật toán: Phần đầu chương giới thiệu toán phân loại thư rác, bộ dữ liệu thử nghiệm và cài đặt chi tiết hai thuật toán đề cập ở chương 2. Phần cuối của chương trình bày kết quả thu được và đưa ra đánh giá về hai thuật toán được sử dụng trong bài toán lọc thư rác.

Cuối luận văn là phần kết luận và danh sách các tài liệu tham khảo. Phần thực nghiệm về phân loại thư rác được trình bày thêm trong phần phụ lục luận văn.