

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

Phùng Thị Thu Trang

HỆ MIỄN DỊCH NHÂN TẠO VÀ ỨNG DỤNG
CHO LỌC THƯ RÁC

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

Phùng Thị Thu Trang

HỆ MIỄN DỊCH NHÂN TẠO VÀ ỨNG DỤNG
CHO LỌC THƯ RÁC

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

HƯỚNG DẪN KHOA HỌC: TS. VŨ MẠNH XUÂN

Thái Nguyên - 2015

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này của tự bản thân tôi tìm hiểu, nghiên cứu dưới sự hướng dẫn của TS Vũ Mạnh Xuân. Các chương trình thực nghiệm do chính bản thân tôi lập trình, các kết quả là hoàn toàn trung thực. Các tài liệu tham khảo được trích dẫn và chú thích đầy đủ.

TÁC GIẢ LUẬN VĂN

Phùng Thị Thu Trang

LỜI CẢM ƠN

Tôi xin bày tỏ lời cảm ơn chân thành tới tập thể các thầy cô giáo Viện công nghệ thông tin – Viện Hàn lâm Khoa học và Công nghệ Việt Nam, các thầy cô giáo Trường Đại học Công nghệ thông tin và truyền thông - Đại học Thái Nguyên đã dạy dỗ chúng tôi trong suốt quá trình học tập chương trình cao học tại trường..

Đặc biệt tôi xin bày tỏ lòng biết ơn sâu sắc tới thầy giáo TS Vũ Mạnh Xuân đã quan tâm, định hướng và đưa ra những góp ý, gợi ý, chỉnh sửa quý báu cho tôi trong quá trình làm luận văn tốt nghiệp. Cũng như các bạn bè, đồng nghiệp, gia đình và người thân đã quan tâm, giúp đỡ và chia sẻ với tôi trong suốt quá trình làm luận văn tốt nghiệp.

Dù đã có nhiều cố gắng nhưng chắc chắn sẽ không tránh khỏi những thiếu sót vì vậy rất mong nhận được sự đóng góp ý kiến của các thầy, cô và các bạn để luận văn này được hoàn thiện hơn.

Tôi xin chân thành cảm ơn!

Thái Nguyên, tháng 08 năm 2015

Phùng Thị Thu Trang

MỤC LỤC

DANH MỤC CÁC HÌNH VẼ.....	vi
DANH MỤC CÁC BẢNG.....	vii
MỞ ĐẦU.....	8
Chương 1 NGHIÊN CỨU TỔNG QUAN VỀ THƯ RÁC	3
1.1. Giới thiệu về thư rác.....	3
1.1.1. Lịch sử.....	3
1.1.2. Định nghĩa.....	4
1.1.3. Mục đích chính gửi thư rác	5
1.1.4. Các đặc tính của thư rác	6
1.1.5. Các kỹ thuật tạo thư rác	7
1.2. Các kỹ thuật phát hiện và ngăn chặn thư rác	8
1.2.1. Kỹ thuật blacklisting	8
1.2.2. Kỹ thuật whitelisting	9
1.2.3. Kỹ thuật heuristic filtering	11
1.2.4. Kỹ thuật challenge/ response	13
1.2.5. Phân tích và định hướng phát triển ứng dụng thử nghiệm.....	14
1.3. Kỹ thuật dùng học máy	15
1.3.1. Support vector Machine (SVM).....	16
1.3.2. K–Nearest Neighbor (kNN)	18
1.3.3. Naïve Bayes (NB)	19
1.3.4. Mạng Neural (Nnet)	21
Chương 2 HỆ MIỄN DỊCH NHÂN TẠO.....	23
2.1. Hệ miễn dịch	23
2.1.1. Hệ miễn dịch sinh học.....	23
2.1.2. Hệ miễn dịch nhân tạo.....	27
2.2. Một số định nghĩa.....	31
2.2.1. Self	32
2.2.2. NonSelf	32
2.2.3. Bộ dò	32
2.2.4. Tập bộ dò ChunkD(S, r) và ContD(S, r)	32
2.2.5. Khả năng phát hiện của tập bộ dò	33

2.2.6. Hole	34
2.3. Thuật toán sinh tập bộ dò r-chunk	34
2.3.1. Thuật toán.....	35
2.3.2. Độ phức tạp thuật toán	36
2.4. Thuật toán sinh tập bộ dò dạng r – contiguous	38
2.4.1. Thuật toán.....	38
2.4.2. Độ phức tạp thuật toán	39
2.5. Các nghiên cứu gần đây	42
Chương 3 CÀI ĐẶT THỬ NGHIỆM	44
3.1. Ứng dụng hệ miễn dịch nhân tạo trong lọc thư rác.....	44
3.1.1. Phát biểu bài toán	44
3.1.2. Cơ sở dữ liệu TREC'07	45
3.1.3. Thiết kế phần mềm	46
3.1.4. Phân tích thuật toán	47
3.1.5. Giao diện chương trình và kết quả	48
3.1.6. Đánh giá.....	51
3.2. So sánh với các thuật toán trên WEKA	51
3.2.1. Phát biểu bài toán	51
3.2.2. Cơ sở dữ liệu SpamBase.....	52
3.2.3. Phần mềm WEKA	54
3.2.4. Thiết kế phần mềm.....	57
3.2.5. Phân tích thuật toán.....	58
3.2.6. Giao diện chương trình và kết quả.....	61
3.2.7. Đánh giá	65
KẾT LUẬN.....	67
DANH MỤC CÁC CÔNG TRÌNH CÔNG BỐ	69
TÀI LIỆU THAM KHẢO.....	70

DANH SÁCH KÍ HIỆU, TỪ VIẾT TẮT

Viết tắt	Viết đầy đủ
$ X $	Lực lượng của tập X
CSDL	Cơ sở dữ liệu
HAM	Thư bình thường
HMD	Hệ miễn dịch
ISP	Internet Service Provider
MHC	Major Histocompatibility Complex
NSA	Negative Selection Algorithm-Thuật toán chọn lọc tiêu cực
SPAM	Thư rác
SVM	Support vector Machine
WEKA	Waikato Environment for Knowledge Analysis

DANH MỤC CÁC HÌNH VẼ

Hình 1.1: So sánh thư rác với các thư điện tử khác.	5
Hình 1.2. Mô hình SVM	17
Hình 1.3. Siêu phẳng h phân chia dữ liệu	17
Hình 2.1. Các tầng miễn dịch sinh học	24
Hình 2.2. Một số cơ quan của hệ miễn dịch sinh học	26
Hình 2.3. Cấu trúc phân tầng của HMD nhân tạo.....	27
Hình 2.4. Sơ đồ khối thuật toán chọn lọc tích cực.....	29
Hình 2.5. Sơ đồ khối thuật toán chọn lọc tiêu cực.....	31
Hình 3.1. Giao diện chương trình lọc spam trên bộ CSDL TREC'07	49
Hình 3.2. Kết quả quá trình test	50
Hình 3.3. Giao diện phần mềm Weka.....	55
Hình 3.4. Giao diện Weka Explorer.....	55
Hình 3.5. Giao diện Weka Explorer sau khi chọn CSDL Spambase.....	56
Hình 3.6. Phân loại dữ liệu.....	57
Hình 3.7. Giao diện chương trình	61

DANH MỤC CÁC BẢNG

Bảng 1.1. Các phần mềm chống thư rác	14
Bảng 2.1. Sự tương quan giữa hệ miễn dịch với môi trường mạng.....	31
Bảng 3.1. Kết quả khi chạy chương trình với 9 bộ test	50
Bảng 3.2. So sánh kết quả	51
Bảng 3.3. Kết quả thử nghiệm trên WEKA và NSA	63
Bảng 3.4. So sánh NSA với một số phương pháp cho kết quả tốt hơn.....	63
Bảng 3.5. So sánh NSA với một số phương pháp cho kết quả thấp hơn.....	64
Bảng 3.6. Kết quả so khớp với giá trị tham số r thay đổi	65

MỞ ĐẦU

Email là một trong những phương tiện truyền thông phổ biến nhất hiện nay, mỗi ngày trên thế giới có hàng tỉ các email được gửi đi nhưng trong số đó hơn một nửa là dạng thư rác (email spam). Email spam là các email được gửi đi với số lượng lớn nhằm mục đích quảng cáo, ăn cắp thông tin, lây lan virus... Đây là những email mà người nhận không hề mong đợi.

Có nhiều kỹ thuật lọc thư rác khác nhau như: Blacklisting, Whitelisting, Heuristic filtering, Challenge/ Response, Throttling, Address obfuscation, Collaborative filtering,... Tuy nhiên, đa số các phần mềm chống thư rác là dựa trên header của thư hoặc địa chỉ từ người gửi nhằm tăng tốc độ. Các phương pháp học máy gần đây được quan tâm nhiều vì chúng có khả năng thích nghi cao với sự tiến hóa của thư rác như phương pháp dựa trên xác suất Naïve Bayes, phương pháp học máy vector hỗ trợ (Support vector machine), phương pháp phân loại dựa trên láng giềng gần nhất (k-nearest neighbors).

Hệ miễn dịch nhân tạo là một hệ thống thích nghi lấy ý tưởng của học thuyết miễn dịch và những chức năng, nguyên tắc, mô hình miễn dịch quan sát được, áp dụng giải các bài toán thực tế. Nó có nhiều ứng dụng khác nhau, nhưng chủ yếu được áp dụng vào lĩnh vực an ninh mạng và an ninh máy tính.

Với lý do trên, đề tài “Hệ miễn dịch nhân tạo và ứng dụng cho lọc thư rác” tập trung nghiên cứu về hệ miễn dịch nhân tạo và ứng dụng một số kỹ thuật cải tiến thuật toán miễn dịch áp dụng cho bài toán lọc thư rác để có thể áp dụng cho không gian dữ liệu lớn.

Đối tượng và phạm vi nghiên cứu

- Nghiên cứu lý thuyết về hệ miễn dịch sinh học và hệ miễn dịch nhân tạo