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

NGUYỄN THỊ HUYÊN

NGHIÊN CỨU MỘT SỐ KỸ THUẬT PHÁT HIỆN TRANG WEB GIẢ MẠO VÀ ỨNG DỤNG

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2016

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

NGUYỄN THỊ HUYÊN

NGHIÊN CỨU MỘT SỐ KỸ THUẬT PHÁT HIỆN TRANG WEB GIẢ MẠO VÀ ỨNG DỤNG

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 NGỌC CƯƠNG

THÁI NGUYÊN - 2016

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. Nguyễn Ngọc Cương. 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

Nguyễn Thị Huyên

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 giảng dạy cho 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. Nguyễn Ngọc Cương 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 05 năm 2016

Nguyễn Thị Huyên

MỤC LỤC

Trang

DANH SÁCH KÍ HIỆU, TỪ VIẾT TẮT

Viết tắt	**Viết đầy đủ**
|X|	Lực lượng của tập X
APWG	Anti Phishing Working Group
ARP	Address Resolution Protocol
CSDL	Cơ sở dữ liệu
Phishing	Giả mạo
DHCP	Dynamic Host Configuration Protocol
DNS	Domain Name System
DOM	Document Object Model
TF-IDF	Term Frequency – Inverse Document Frequency
WEKA	Waikato Environment for Knowledge Analysis
NB	Naïve Bayes
MAC	Media Access Control
LAN	Local Area Network
DoS	Dinal of Services
TCP/IP	Transmission Control Protocol / Internet Protocol
SMTP	Simple_Mail_Transfer_Protocol
URL	Uniform Resource Locator
XML	Extensible Markup Language

DANH MỤC CÁC BẢNG VÀ HÌNH VẼ

MỞ ĐẦU

1. Đặt vấn đề

Hiện nay, công nghệ thông tin hầu như được áp dụng rộng rãi trên toàn cầu, nước chúng ta cũng đang dần chuyển mình từ từ tiếp xúc với công nghệ vì thấy được lợi ích to lớn trong việc áp dụng công nghệ thông tin vào các lĩnh vực như kinh doanh, quản lý, mua sắm,... nói chung là tất cả nhu cầu của con người. Một trong những dịch vụ công nghệ hàng đầu được sử dụng phổ biến nhất là dịch vụ WEB. Với công nghệ WEB hiện tại thì có thể đáp ứng mọi nhu cầu của con người và hơn thế nữa.

Giả mạo (*phishing* biến thể từ *fishing* nghĩa là câu cá và *phreaking* nghĩa là nhử người dùng tiết lộ bí mật), trong lĩnh vực bảo mật máy tính là một hành vi giả mạo ác ý nhằm lấy được các thông tin nhạy cảm như tên người dùng, mật khẩu và các chi tiết thẻ tín dụng bằng cách giả dạng thành một chủ thể tin cậy trong một giao dịch điện tử.

Vấn đề giả mạo (phishing hay fake) nói chung và giả mạo web nói riêng là một loại tội phạm kỹ thuật xã hội đang có xu hướng gia tăng trên mạng. Giả mạo được báo cáo là vấn nạn web lần đầu tiên vào năm 2001 của hiệp hội bảo vệ khách hàng, hiệp hội thương mại liên bang của Mỹ và ngày nay nhóm làm việc chống giả mạo APWG (*Anti Phishing Working Group*) đã đưa ra thông số những trang web giả đang tăng khoảng 50% mỗi năm.

Hầu hết các tấn công lừa đảo hiện đại xảy ra bằng cách thu hút người sử dụng truy cập vào một trang web độc hại trông và hoạt động giống như bản gốc. Khi đó, người sử dụng nếu bị thuyết phục rằng trang này là xác thực có thể cung cấp thông tin cá nhân bao gồm cả thông tin xác thực hoặc thông tin ngân hàng. Những thông tin này thường được kẻ sử dụng để thực hiện một số hình thức của hành vi trộm cắp hay gian lận trong thực tế.

Do vậy, việc nghiên cứu và phát hiện các trang web giả mạo là một nhu cầu cấp thiết hiện nay.

Phát hiện trang web giả mạo là việc đầu tiên để ngăn chặn và xóa bỏ các trang web giả mạo. Hiện nay có rất nhiều các cách tiếp cận khác nhau để phát hiện các trang web giả mạo.

Một đặc tính nổi bật nhất của trang web giả mạo là nó phải tương tự như trang web gốc. Điều này có nghĩa là hai trang web gốc và web giả mạo có cấu trúc giống nhau đến mức tốt nhất để người dùng có đủ tự tin tiết lộ những thông tin nhạy cảm. Hầu hết các trang lừa đảo đều làm tốt việc tạo giao diện hợp lệ bằng cách sao chép bố trí trang, font, kiểu, logo và thậm chí các thông tin bảo mật của trang hợp lệ.

Có nhiều kỹ thuật và giải pháp để phát hiện trang web giả mạo:

1. *Hướng mở rộng các giải pháp từ thư rác*: *Thuật toán TF-IDF* (*Term Frequency/Inverse Document Frequency*) sử dụng những từ khóa duy nhất để xác định một trang cụ thể. Kỹ thuật này thường được dùng trong khai thác văn bản hoặc với các máy tìm kiếm để tìm các trang liên quan. Thuật toán TF-IDF sẽ xác định những từ khóa của một trang web, những từ khóa này được đưa vào một máy tìm kiếm chẳng hạn Google và lấy ra nhóm những URL trên cùng. Nếu trang web bị nghi ngờ nằm trong nhóm đó thì trang này được coi là hợp lệ, ngược lại nó sẽ bị cho là lừa đảo vì hầu hết các trang lừa đảo không có thứ hạng cao trong các kết quả của máy tìm kiếm.

Thuật toán này được ứng dụng trong giải pháp Cantina được phát triển bởi các nhà nghiên cứu của Đại học Carnegie Mellon với việc sử dụng năm từ khóa có tần suất xuất hiện cao nhất trong trang. Tuy nhiên giải pháp chỉ phù hợp khi có hai giả thiết sau:

- Thứ nhất, trang lừa đảo phải nhìn và hoạt động giống với trang hợp lệ thì mới cho kết quả từ khóa được xác định bởi TF-IDF giống nhau.

- Thứ hai, các máy tìm kiếm phải cho kết quả xếp hạng các trang web hợp lệ chính xác và cao hơn các trang lừa đảo.

2. Hướng sử dụng giải pháp Bayesian: Thuật toán lọc Bayesian vốn được phát triển để phát hiện thư rác nhưng các nhà nghiên cứu của Đại học Iowa đã sử dụng thuật toán này để phát triển thành công cụ chống lừa đảo được đặt tên là B-APT. Lợi thế chính của thuật toán này là có khả năng phát hiện được những đối tượng chưa từng nhìn thấy trước đó. Việc sử dụng phép lọc Bayesian là một giải pháp hứa hẹn cho việc phát hiện lừa đảo 0 ngày