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

Trần Huy Phong

TÌM HIỂU, NGHIÊN CỨU HỆ THỐNG PHÁT HIỆN XÂM NHẬP DỰA TRÊN KHAI PHÁ DỮ LIỆU

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2015

LỜI CAM ĐOAN

Tôi xin cam đoan số liệu và kết quả nghiên cứu trong luận văn này là trung thực và chưa được sử dụng để bảo vệ học hàm, học vị nào.

Tôi xin cam đoan: Mọi sự giúp đỡ cho việc thực hiện luận văn này đã được cám ơn, các thông tin trích dẫn trong luận văn này đều đã được chỉ rõ nguồn gốc.

Thái nguyên, ngày tháng năm

TÁC GIẢ LUẬN VĂN

Trần Huy Phong

LỜI CẢM ƠN

Trong thời gian nghiên cứu và thực hiện luận văn này, em đã may mắn được các thầy cô chỉ bảo, dìu dắt và được gia đình, bạn bè quan tâm, động viên. Em xin bày tỏ lời cảm ơn sâu sắc nhất tới tất cả các tập thể, cá nhân đã tạo điều kiện giúp đỡ em trong suốt quá trình thực hiện nghiên cứu luận văn này.

Trước hết em xin trân trọng cảm ơn Ban giám hiệu trường Đại học Công nghệ thông tin và truyền thông, Phòng Đào tạo và Khoa Sau đại học của nhà trường cùng các thầy cô giáo, những người đã trang bị kiến thức cho em trong suốt quá trình học tập.

Với lòng biết ơn chân thành và sâu sắc nhất, em xin trân trọng cảm ơn thầy giáo- TS. Trần Đức Sự, người thầy đã trực tiếp chỉ bảo, hướng dẫn khoa học và giúp đỡ em trong suốt quá trình nghiên cứu, hoàn thành luận văn này.

Xin chân thành cảm ơn tất cả các bạn bè, đồng nghiệp đã động viên, giúp đỡ nhiệt tình và đóng góp nhiều ý kiến quý báu để em hoàn thành luận văn này.

Do thời gian nghiên cứu có hạn, luận văn của em chắc hẳn không thể tránh khỏi những sơ suất, thiếu sót, em rất mong nhận được sự đóng góp của các thầy cô giáo cùng toàn thể bạn đọc.

Xin trân trọng cảm ơn!

Thái nguyên, ngày……tháng….năm……

TÁC GIẢ LUẬN VĂN

Trần Huy Phong

MỤC LỤC

DANH MỤC VIẾT TẮT

Ký Hiệu	Tiếng Anh	Ý Nghĩa
IDS	Intrusion Detection System	Hệ thống phát hiện xâm nhập
NIDS	Network-base IDS	
HIDS	Host-based IDS	
KDD	Knowledge Discovery and Data Mining	Phát hiện tri thức
AAFID	Autonomous Agents for Intrusion Detection	Tác nhân tự trị cho việc phát hiện xâm phạm
CSDL		Cơ sở dữ liệu
OLAP	On Line Analytical Processing	Công cụ phân tích trực tuyến
DARPA	Defense Advanced Research Projects Agency	Cơ quan dự án phòng thủ tiên tiến
CPU	Central Processing Unit	Đơn vị xử lý trung tâm
DoS	Denial-of-Service	Tấn công từ chối dịch vụ
MADAMID	Mining Audit Data for Automated Models for Instruction Detection	Khai phá dữ liệu được sử dụng trong mô hình tự động để phát hiện xâm nhập
RIPPER		Thuật toán phân lớp dựa vào luật
WEKA	Waikato Enviroment for krowledge Analysis	

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

DANH MỤC BIỂU ĐỒ

MỞ ĐẦU

1. Lý do chọn đề tài:

Kể từ khi mạng Internet ra đời đến nay, thế giới đã chứng kiến sự thay đổi vô cùng to lớn và kì diệu về nhiều mặt của đời sống con người. Nền kinh tế thế giới và đời sống xã hội đã có nhiều sự biến đổi và ngày càng phụ thuộc vào công nghệ thông tin nói chung cũng như công nghệ Internet nói riêng. Điều đó cũng dẫn đến một mặt trái, đó là càng ngày càng nhiều các thông tin quan trọng của các cơ quan, tổ chức hay cá nhân lưu trữ trên các mạng máy tính, mà đa số các mạng máy tính này lại không đảm bảo độ an toàn, bảo mật thông tin tuyệt đối.

Đi cùng với sự phát triển đó là những nguy cơ tấn công và xâm nhập mạng không ngừng gia tăng. Các đối tượng tấn công và hình thức tấn công mạng ngày một đa dạng, tinh vi và phức tạp hơn.

Vấn đề bảo mật, an toàn cho các hệ thống thông tin nói chung và hệ thống mạng nói riêng là một vấn đề cấp bách và rất đáng được quan tâm. Bởi vậy, để bảo vệ các hệ thống thông tin người ta sử dụng nhiều các giải pháp kỹ thuật khác nhau như hệ thống tường lửa, mã hoá, mạng riêng ảo (VPN), phòng chống virus…Trong đó phát hiện xâm nhập trái phép (IDS) là một trong những công nghệ quan trọng nhất nhằm giúp các tổ chức phát hiện và ngăn chặn kịp thời các tấn công trong thời gian thực, cũng như dự đoán được các nguy cơ tấn công trong tương lai [3], [5]. Chính vì vậy, nghiên cứu về hệ thống IDS sẽ giúp chúng ta nâng cao khả năng xây dựng hệ thống phòng thủ cho việc giám sát an ninh mạng.

Hai phương pháp cơ bản để phát hiện xâm nhập trái phép là dựa trên tập luật và dựa trên các dấu hiệu bất thường [1], [2], [6], [7]. Phương pháp dựa trên tập luật có thể phát hiện các tấn công dựa trên một cơ sở dữ liệu các dấu hiệu đã được định nghĩa trước. Phương pháp này thường có độ chính xác cao cũng như ít đưa ra các cảnh báo nhầm. Tuy nhiên, vấn đề của phương pháp này là không thể phát hiện được các tấn công mới chưa được định nghĩa hoặc cập nhật trong cơ sở dữ liệu. Phương pháp dựa trên các dấu hiệu bất thường có thể giúp xác định các tấn công mới nhưng thường cho độ chính xác thấp hơn so với phương pháp dựa trên tập luật.

Hiện nay, Khai phá dữ liệu đã có nhiều bước phát triển vượt bậc và có nhiều ứng dụng kỹ thuật bằng các thuật toán khác nhau trong thực tế. Khai phá dữ liệu là một phương pháp tiếp cận mới trong việc phát hiện xâm nhập. Xây dựng mô hình hệ thống phát hiện xâm nhập dựa trên khai phá dữ liệu là một hướng phát triển mới và hiệu quả trong xây dựng hệ thống IDS.

Xuất phát từ những yêu cầu và lý do trên, em lựa chọn đề tài luận văn là: "**Tìm hiểu, nghiên cứu hệ thống phát hiện xâm nhập dựa trên khai phá dữ liệu".**

Luận văn nghiên cứu khai phá dữ liệu và nghiên cứu ứng dụng mô hình hệ thống phát hiện xâm nhập trái phép dựa trên khai phá dữ liệu; Từ đó đánh giá hiệu năng của hệ thống phát hiện xâm nhập đối với các thuật toán phân lớp khác nhau trong thực tế.

2. Mục tiêu nghiên cứu:

- Nghiên cứu tổng quan về hệ thống phát hiện xâm nhập.
- Nghiên cứu một số thuật toán khai phá dữ liệu.
- Ứng dụng một số thuật toán khai phá dữ liệu trong phát hiện xâm nhập, so sánh sự hiệu quả của các thuật toán.
- Đánh giá hiệu năng cho mô hình đó bằng các thuật toán phân lớp khác nhau như: Naïve Bayes, Decision Tree.

3. Đối tượng và phạm vi nghiên cứu:

- Nghiên cứu mô hình hệ thống IDS hiện nay và đánh giá ưu, nhược điểm của IDS.
- Nghiên cứu các bài toán, kỹ thuật khai phá dữ liệu.
- Ứng dụng của khai phá dữ liệu trong hệ thống phát hiện xâm nhập.
- Một số thuật toán phân lớp dữ liệu.
- Đánh giá hiệu năng các kỹ thuật phân lớp cho hệ thống phát hiện xâm nhập dựa trên khai phá dữ liệu.

4. Ý nghĩa thực tiễn của luận văn:

- Nghiên cứu ứng dụng mô hình hệ thống phát hiện xâm nhập dựa trên khai phá dữ liệu giải quyết các vấn đề tồn tại của hệ thống IDS hiện nay.
- Đánh giá hiệu quả phân lớp cho mô hình. Đồng thời đề xuất lựa chọn các kỹ thuật phân lớp phù hợp với từng loại tấn công cụ thể cho hệ thống phát hiện xâm nhập dựa trên khai phá dữ liệu đã đề xuất.

5. Phương pháp nghiên cứu:

Việc giám sát các hành động trên mạng có thể thu thập và phân tích để phát hiện ra các tấn công mạng. Các hành động này có thể tìm thấy trong các tệp log của ứng dụng như tạo, xóa file, truy cập vào tệp có mật khẩu, gọi các lệnh của hệ thống...

Việc phân tích phát hiện các tấn công dựa trên tập dữ liệu về các hành động này có thể thực hiện thông qua các thuật toán phân lớp dữ liệu, để phân lớp thành các lớp tấn công đã biết trước hoặc lớp truy cập bình thường.

Nghiên cứu các tài liệu liên quan trong lĩnh vực khai phá dữ liệu và phát hiện xâm nhập. Tìm hiểu, nghiên cứu các kỹ thuật phát hiện xâm nhập dựa trên phương pháp thống kê và khai phá dữ liệu.

Trên cơ sở nghiên cứu và phân tích tập dữ liệu DARPA [15]. Phân tích bằng lý thuyết và thực nghiệm để xác định các thuộc tính quan trọng của tập dữ liệu có ảnh hưởng đến một hành động tấn công cụ thể, từ đó trích rút và chuyển đổi thành định dạng phù hợp cho các thuật toán học phân lớp.

Nghiên cứu xây dựng các thực nghiệm sử dụng phần mềm Weka [14], đánh giá hiệu quả của các thuật toán học phân lớp trên tập dữ liệu DARPA.