

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

NGUYỄN VĂN DIỄN

NGHIÊN CỨU GIẢI PHÁP PHÁT HIỆN XÂM NHẬP MẠNG MÁY
TÍNH BẤT THƯỜNG DỰA TRÊN KHAI PHÁ DỮ LIỆU

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2014

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài “Nghiên cứu Giải pháp phát hiện xâm nhập mạng máy tính bất thường dựa trên Khai phá dữ liệu” là công trình nghiên cứu của riêng tôi. Đề tài được hoàn thành dưới sự hướng dẫn của Thầy TS. Nguyễn Ngọc Cương. Những kết quả nghiên cứu, thử nghiệm được thực hiện hoàn toàn khách quan và trung thực. Các số liệu, kết quả trình bày trong luận văn là hoàn toàn trung thực và chưa từng được công bố trong bất cứ công trình nào.

Các tài liệu tham khảo sử dụng trong luận văn đều được dẫn nguồn (có bảng thống kê các tài liệu tham khảo) hoặc được sự đồng ý trực tiếp của tác giả.

Nếu xảy ra bất cứ điều gì không đúng như những lời cam đoan trên, tôi xin chịu hoàn toàn trách nhiệm.

Hà Nội, ngày 18 tháng 07 năm 2014

TÁC GIẢ

Nguyễn Văn Diễn

LỜI CẢM ƠN

Em xin chân thành cảm ơn Thầy TS. Nguyễn Ngọc Cương người đã trực tiếp hướng dẫn tận tình em trong suốt quá trình thực hiện Luận văn tốt nghiệp.

Em xin chân thành cảm ơn Quý thầy, cô Trường Đại học Công nghệ thông tin & Truyền thông Thái Nguyên, Viện Công nghệ Thông Tin, những người đã nhiệt tình giảng dạy và truyền đạt những kiến thức quý báu trong suốt thời gian em học tập và nghiên cứu tại trường. Với vốn kiến thức tiếp thu được trong quá trình học tập và nghiên cứu không chỉ là nền tảng cho quá trình nghiên cứu luận văn mà còn là hành trang quý báu trong quá trình hoạt động chuyên môn của em.

Cuối cùng, em xin kính chúc Quý thầy cô, đồng nghiệp, gia đình dồi dào sức khỏe và thành công.

Trân trọng cảm ơn!

MỤC LỤC

LỜI CAM ĐOAN.....	i
LỜI CẢM ƠN.....	ii
DANH MỤC TỪ VIẾT TẮT	vi
DANH MỤC BẢNG.....	vii
DANH MỤC HÌNH	viii
MỞ ĐẦU	ix
TỔNG QUAN VỀ NHIỆM VỤ CỦA LUẬN VĂN	xi
CHƯƠNG 1: HỆ THỐNG PHÁT HIỆN XÂM NHẬP MẠNG VÀ PHƯƠNG PHÁP PHÁT HIỆN XÂM NHẬP MẠNG.....	1
1.1 Hệ thống phát hiện xâm nhập mạng IDS (<i>Intrusion Detection System</i>).....	1
1.1.1. Định nghĩa	1
1.1.2. Vai trò, chức năng của IDS	1
1.1.3. Mô hình IDS mức vật lý	2
1.1.4. Kiến trúc và hoạt động bên trong mô hình hệ thống IDS	3
1.1.5. Phân loại IDS.....	6
1.1.6. Một số kiểu tấn công cơ bản vào hệ thống mạng	8
1.2 Một số phương pháp phát hiện bất thường trong hệ thống IDS	11
1.2.1 Phương pháp tiếp cận dựa trên xác suất thống kê	11
1.2.2 Phương pháp tiếp cận dựa trên trạng thái	12
1.2.3 Phương pháp tiếp cận dựa trên hệ chuyên gia.....	12
1.2.4 Phương pháp tiếp cận dựa trên khai phá dữ liệu	13
1.3 Khai phá dữ liệu trong IDS	14
1.3.1 Định nghĩa khai phá dữ liệu.....	14

1.3.2	Nhiệm vụ của khai phá dữ liệu	16
1.3.3	Các loại dữ liệu được khai phá	17
1.3.4	Quy trình khai phá dữ liệu.....	18
1.3.5	Một số phương pháp khai phá dữ liệu	19
1.3.6	Một số kỹ thuật dùng trong khai phá dữ liệu.....	21
CHƯƠNG 2: PHƯƠNG PHÁP PHÁT HIỆN BẤT THƯỜNG DỰA TRÊN KỸ THUẬT KHAI PHÁ DỮ LIỆU		26
2.1.	Phát hiện bất thường dựa trên khai phá dữ liệu	26
2.1.1.	Phương pháp phát hiện bất thường dựa trên khai phá dữ liệu.....	26
2.1.2.	Kỹ thuật phát hiện xâm nhập dựa trên khai phá dữ liệu.....	26
2.2.	Bài toán phát hiện phần tử dị biệt trong khai phá dữ liệu	28
2.2.1.	Một số thuật toán phát hiện dị biệt trong khai phá dữ liệu	30
2.2.2.	Mô hình phát hiện bất thường dựa trên kỹ thuật khai phá dữ liệu	36
CHƯƠNG 3: ĐỀ XUẤT TRIỂN KHAI THỬ NGHIỆM HỆ THỐNG PHÁT HIỆN XÂM NHẬP MẠNG.....		42
3.1.	Bài toán phân cụm dữ liệu trong CSDL kết nối mạng	42
3.2.	Thuật toán sử dụng cho bài toán ứng dụng	42
3.3.	Đánh giá Thuật toán phân cụm ứng dụng trong bài toán.....	44
3.4.	Ứng dụng thuật toán phân cụm K-medoids trong KPDL	48
3.4.1.	Quy trình xử lý bài toán ứng dụng:	48
3.4.2.	Tập hợp dữ liệu	49
3.4.3.	Tiền xử lý	49
3.4.4.	Tiến trình khai phá dữ liệu	51
3.5.	Chương trình Demo.....	54

3.6. Nhận xét bài toán KPDL.....	59
KẾT LUẬN VÀ HƯỚNG PHÁP TRIỂN.....	61
TÀI LIỆU THAM KHẢO	62

DANH MỤC TỪ VIẾT TẮT

ADAM	Audit Data Analysis Mining
CSDL	Cơ sở dữ liệu
DDoS	Distributed Denial of Services
DOS	Denial of Services
HIDS	Host Intrusion Detection System
HTTP	Hypertext Markup Language
ICMP	Internet Control Message Protocol
IDS	Intrusion Detection System
IDDM	Intrusion Detection Data Mining
IPS	Intrusion Prevention System
IP	Internet Protocol
KPDL	Khai phá dữ liệu
LOF	Local Outlier Factor
LSC	Local Sparsity Ratio
NIDS	Network Intrusion Detection System
MAC	Media Access Controllers
SQL	Structured Query Language
VPN	Virtual Private Network
TCP	Transmission Control Protocol
UDP	User Datagram Protocol

DANH MỤC BẢNG

Bảng 2.1: Danh sách các cảnh báo chưa rút gọn

Bảng 2.2: Danh sách các cảnh báo sau khi rút gọn

Bảng 3.1: Bảng thuộc tính CSDL mạng

Bảng 3.2: Thông tin chương trình cài đặt ứng dụng

DANH MỤC HÌNH

Hình 1.1: Mô hình IDS vật lý

Hình 1.2: Kiến trúc Modul trong IDS

Hình 1.3: Mô hình thu thập dữ liệu ngoài luồng

Hình 1.4: Mô hình thu thập dữ liệu trong luồng

Hình 1.5: Modul phân tích, phát hiện tấn công

Hình 1.6: Quá trình khám phá tri thức

Hình 2.1: Gán giá trị để lượng hóa các cuộc tấn công trên sơ đồ

Hình 2.2: Minh họa bài toán phát hiện phần tử dị biệt

Hình 2.3: Khoảng cách Reach – dist

Hình 2.4: Phương pháp LOF

Hình 2.5: Thuật toán LSC – Mine

Hình 2.6: Mô hình phát hiện bất thường sử dụng kỹ thuật KPDL

Hình 2.7: Mô hình Modul tổng hợp

Hình 3.1: Lưu đồ thuật toán K-Medoids

Hình 3.2: Tiến trình phát hiện xâm nhập mạng sử dụng kỹ thuật phân cụm

Hình 3.3: Biểu diễn CSDL mạng

Hình 3.4: Biến đổi dữ liệu trong CSDL

Hình 3.5: Gom cụm dữ liệu trong CSDL

Hình 3.6: Biểu diễn kết quả mẫu bất thường

Hình 3.7: Giao diện Menu chính

Hình 3.8: Giao diện khai phá trên giao thức HTTP

Hình 3.9: Giao diện khai phá dữ liệu tự động

Hình 3.10: Giao diện tiền xử lý

Hình 3.11: Giao diện khai phá dựa trên ngưỡng kết nối

MỞ ĐẦU

Ngày nay, Công nghệ thông tin nói chung và Ngành mạng máy tính nói riêng đã được ứng dụng trong hầu hết các lĩnh vực quan trọng của đời sống, nó tác động trực tiếp đến sự tồn tại và phát triển của nền kinh tế tri thức và công nghệ. Chính vì vậy, việc áp dụng Công nghệ thông tin đã trở thành một yêu cầu không thể thiếu cho tất cả các tổ chức, doanh nghiệp. Với tầm quan trọng như vậy, cần phải có một hệ thống mạng doanh nghiệp ổn định, hoạt động liên tục, đảm bảo tính tin cậy, nguyên vẹn, sẵn sàng và không thể từ chối để đáp ứng được mọi yêu cầu kết nối và xử lý của công việc.

Tuy nhiên, bên cạnh yêu cầu cấp thiết đó thì mạng máy tính luôn phải đối diện với rất nhiều nguy cơ mất an toàn như các cuộc “viếng thăm” bất hợp pháp hoặc các cuộc tấn công từ bên ngoài mạng luôn luôn có thể xảy ra với mức độ ngày càng phức tạp và tinh vi hơn. Do đó, yêu cầu phải có một hệ thống có thể phát hiện tự động những hành vi thâm nhập không được phép để cảnh báo nguy cơ và ngăn chặn đã trở nên cấp thiết.

Đã có nhiều hướng nghiên cứu và xây dựng hệ thống cảnh báo và thâm nhập dựa trên các phương pháp thâm nhập như: phát hiện thâm nhập dựa vào luật; kỹ thuật phân biệt ý định người dùng, phân tích trạng thái phiên, phương pháp phân tích thống kê ... Tuy nhiên đây là các phương pháp phát hiện xâm nhập dựa trên các dấu hiệu bất thường. Tức là dựa trên các dấu hiệu của các vụ tấn công đã biết, các phương pháp này phát hiện ra xâm nhập mạng bằng cách so sánh các giá trị đặc tả với một dãy các ký tự tấn công được cung cấp bởi chuyên gia và được cập nhật lại trong cơ sở dữ liệu. Điểm hạn chế của các phương pháp trên là chúng không thể phát hiện ra các cuộc tấn công mới không có trong cơ sở dữ liệu. So với các phương pháp trên thì phương pháp phân tích dựa trên kỹ thuật khai phá dữ liệu có nhiều ưu điểm rõ rệt hơn. Phương pháp này có thể sử dụng với cơ sở dữ liệu chứa nhiều nhiều, dữ liệu không đầy đủ, biến đổi liên tục, đặc biệt phương pháp này đòi hỏi mức độ sử dụng các chuyên gia không quá thường xuyên. Các ưu điểm này đem lại