HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Khuất Thị Ngọc Ánh

PHƯƠNG PHÁP PHÁT HIỆN TẤN CÔNG WEB ỨNG DỤNG DỰA TRÊN KỸ THUẬT PHÂN TÍCH HÀNH VI

Chuyên ngành: Hệ thống thông tin
Mã số: 8.48.01.04

Hà Nội 2020

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. Đỗ Xuân Chợ

Phản biện 1: ……

Phản biện 2: ……

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .....năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Các nguy cơ mất an toàn thông tin trên thế giới nói chung và Việt Nam nói riêng liên tục ra tăng và phát triển về cả số lượng cũng như mức độ nguy hiểm của các cuộc tấn công. Theo ghi nhận của một số công ty bảo mật trên thế giới, trong vài năm trở lại đây Việt Nam luôn được coi là điểm nóng của mã độc và các cuộc tấn công website trái phép. Hàng loạt các cuộc tấn công website diễn ra với quy mô lớn vào các website của các doanh nghiệp, tổ chức chính phủ… đã gây mất an toàn thông tin và ảnh hưởng nghiêm trọng đến uy tín và doanh nghiệp, tổ chức chính phủ. Hiện nay, các cơ quan nhà nước, các tổ chức chính phủ đã và đang có nhiều biện pháp tích cực trong việc phòng chống và phát hiện tấn công website. Rất nhiều biện pháp đã được ứng dụng và triển khai trong thực tế. Tuy nhiên, các kỹ thuật tấn công website ngày càng được biến đổi tinh vi và phức tạp, đặc biệt là các truy cập thể hiện các hành vi bất thường của người dùng website rất dễ dàng để vượt qua được sự giám sát của các sản phẩm an toàn web.

Website của Trường Đại học Công nghệ Giao thông vận tải được sử dụng cho phép nhiều user bao gồm cả sinh viên, giảng viên và cán bộ công nhân viên chức sử dụng để làm việc và tra cứu thông tin. Hàng ngày có hàng trăm nghìn giao dịch, của người dùng truy cập vào website của trường nhằm khai thác và thực hiện mục đích của mình. Trong số các truy cập này đã có nhiều truy cập bất thường người của người dùng web được ghi nhận, gây mất an toàn thông tin và uy tín của nhà trường. Chính vì vậy, vấn đề phát hiện và ngăn chặn các truy nhập bất thường của người dùng web lên Website của Trường Đại học Công nghệ Giao thông vận tải đang rất được quan tâm hiện nay. Từ những lý do trên, học viên với sự giúp đỡ của TS. Đỗ Xuân Chợ lựa chọn đề tài: “Phương pháp phát hiện tấn công web ứng dụng kỹ thuật phân tích hành vi”.

2. Tổng quan vấn đề cần nghiên cứu

Hiện nay việc tăng trưởng và phát triển nhanh chóng của Internet dẫn đến nhu cầu

bảo mật và đảm bảo an toàn thông tin đang được các doanh nghiệp ngày càng chú trọng.

Theo Báo cáo an ninh website Q3/2018 của CyStack [17], trong quý 3 năm 2018 trên thế giới đã có 129.722 website bị tin tặc tấn công và chiếm quyền điều khiển. Như vậy, cứ mỗi phút trôi qua lại có một website bị tin tặc kiểm soát. Bằng việc chiếm quyền điều khiển website tin tặc có thể gây ra rất nhiều vấn đề rắc rối cho các chủ website: đánh cắp dữ liệu, cài đặt mã độc, phá hoại website, tạo trang lừa đảo (phishing), tống tiền…Theo thống kê, Việt Nam đứng thứ 19 (chiếm 0.9%) trong số các quốc gia có website bị tin tặc tấn công. Cụ thể trong quý 3 năm 2018 đã có 1.183 website của Việt Nam bị tin tặc tấn công và kiểm soát. Các website giới thiệu sản phẩm và dịch vụ của Doanh nghiệp là đối tượng bị tin tặc tấn công nhiều nhất, chiếm tới 71,51%. Vị trí thứ hai là các website Thương mại điện tử chiếm 13,86%. Các website có tên miền .gov.vn của chính phủ chiếm 1.9% trong danh sách với tổng số 23 website bị tấn công.

Ngoài việc sử dụng các phương pháp phòng chống tấn công truyền thống, xu hướng hiện nay là xử dụng trí tuệ nhân tạo, học máy để áp dụng trong lĩnh vực an toàn thông tin để phát hiện nhanh chóng và tăng độ chính xác. Có 2 hướng tiếp cận chính là dựa vào dấu hiệu và hành vi để phát hiện tấn công web nói chung và hành vi bất thường người dùng web nói riêng. Mỗi phương pháp đều có những ưu điểm và nhược điểm nhất định. Trong luận văn, tác giả sẽ đi sâu vào việc nghiên cứu về phương pháp phát hiện hành vi bất thường người dùng web dựa trên kỹ thuật phân tích hành vi. Để luận văn đạt được những kết quả trên, cần nghiên cứu và làm rõ các nội dung:

- Tìm hiểu một số lỗ hổng, điểm yếu và các cuộc tấn công lên web ứng dụng;
- Nghiên cứu và tìm hiểu về một số phương pháp và công nghệ phát hiện tấn công web ứng dụng;
- Nghiên cứu phương pháp phát hiện tấn công web bằng kỹ thuật phân tích hành vi trên cơ sở thuật toán học máy và hành vi người dùng.

3. Mục đích nghiên cứu

- Tìm hiểu về thuật toán phân loại học máy;
- Tìm hiểu về hành vi bất thường người dùng web;

- Nghiên cứu phương pháp phân loại hành vi bất thường của người dùng web dựa trên các thuật toán học máy.

4. Đối tượng và phạm vi nghiên cứu

- Đối tượng nghiên cứu: Dữ liệu Truy cập web, dữ liệu truy cập web ứng dụng của trường Đại học Công nghệ Giao thông vận tải.
- Phạm vi nghiên cứu: Hệ thống website và phương pháp phát hiện hành vi của người dùng web.

5. Phương pháp nghiên cứu

Dựa trên các thuật toán học máy có giám sát từ đó phân loại người dùng và xác định người dùng bất thường.

Cấu trúc nội dung luận văn gồm 3 chương với các nội dung như sau:

Chương 1: Nguy cơ mất an toàn thông tin web và biện pháp phòng chống

Nội dung chương 1 của luận văn sẽ trình bày về một số kỹ thuật tấn công website bao gồm: một số phương pháp tấn công, các công cụ hỗ trợ tấn công... Bên cạnh đó, trong chương 1 luận văn sẽ trình bày một số phương pháp và công cụ phòng chống tấn công web.

Chương 2: Phương pháp phát hiện tấn công trên web dựa trên kỹ thuật phân tích hành vi

Nội dung chương 2 của luận văn sẽ nghiên cứu về một số phương pháp phát hiện tấn công web bao gồm kỹ thuật phát hiện và các công cụ mã nguồn mở hỗ trợ phát hiện tấn công web. Ngoài ra, trong chương 2 sẽ trình bày về phương pháp phát hiện tấn công web dựa trên kỹ thuật phân tích hành vi.

Chương 3: Thực nghiệm và đánh giá

Nội dung chương 3 của luận văn sẽ thực hiện thực nghiệm phát hiện tấn công web dựa trên kỹ thuật phân tích hành vi trên cơ sở thuật toán và hành vi đã được lựa chọn và phân tích ở chương 2

Kết luận.

CHƯƠNG 1: NGUY CƠ MẤT AN TOÀN THÔNG TIN WEB VÀ BIỆN PHÁP PHÒNG CHỐNG

1.1. Kỹ thuật tấn công web

Ngày nay nguy cơ mất an toàn thông tin ngày càng xảy ra nhiều và dẫn đến các hậu quả nghiêm trọng mà người quản trị website không thể lường trước được. Đặc biệt là đối với các cuộc tấn công web ngày càng tinh vi và khó lường. Chính vì vậy, trong mục này luận văn sẽ khảo sát các phương thức tấn công lỗ hổng bảo mật Website dựa trên khuyến nghị của OWASP (The Open Web Application Security Project- dự án mở về bảo mật ứng dụng Web).

1.1.1. Tấn công SQL injection

1.1.2. Tấn công kiểu Broken Authentication And Session Management

1.1.3. Tấn công Cross Site Scripting (XSS)

1.1.4. Kiểu tấn công Insecure Direct Object References

1.1.5. Tấn công Sensitive Data Exposure

1.1.6. Tấn công Missing Function Level Access Control

1.1.7. Tấn công Using Components with Known Vulnerabilities

1.1.8. Tấn công Unvalidated Redirects and Forwards

1.1.11. Tấn công APT

1.2. Phương pháp phòng chống tấn công trên web

1.2.1. Các phương pháp phòng chống tấn công web phổ biến

❖ Phương pháp phòng chống tấn công SQL injection

❖ Phương pháp phòng chống tấn công Cross Site Scripting (XSS)

❖ Phương pháp phòng chống tấn công Cross-Site Request Forgery (CSRF)

1.2.2. Một số phương pháp nâng cao bảo mật hệ thống máy chủ website

Kết luận chương 1

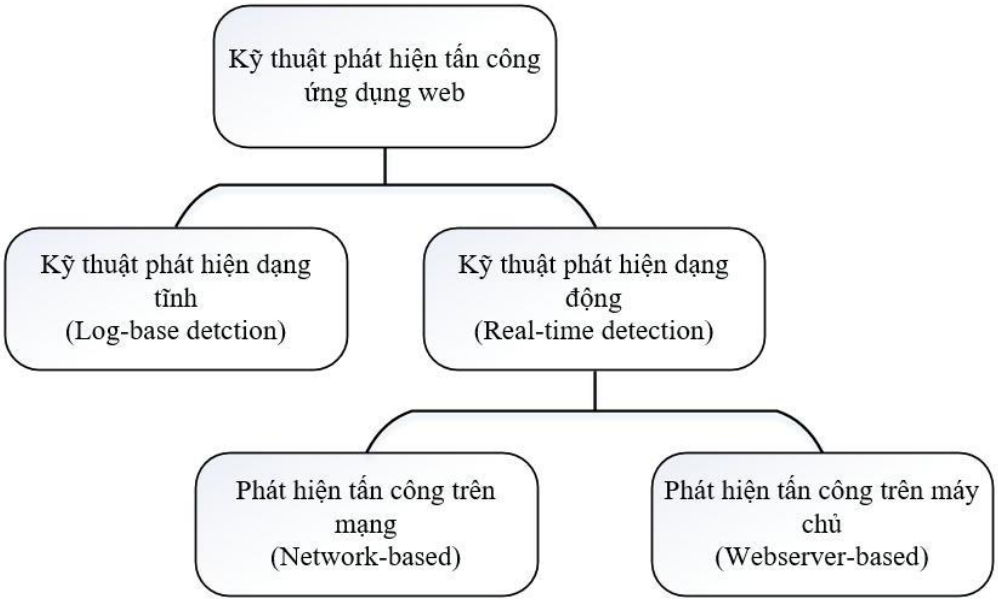
Trong chương 1, luận văn đã khảo sát về các nguy cơ mất an toàn thông tin Website cũng như tìm hiểu về kĩ thuật tấn công vào các lỗ hổng phổ biến hiện nay (Top 10 OWAPS). Từ đó đưa ra một số phương pháp phòng chống tấn công khi xây dựng Website.

Vấn đề phát hiện sớm các cuộc tấn công Website để có các biện pháp phòng ngừa hữu hiệu đóng một vai trò hết sức quan trọng. Chương tiếp theo, luận văn sẽ nghiên cứu các phương pháp phát hiện tấn công trên Website dựa trên kĩ thuật phân tích hành vi.

CHƯƠNG 2: PHƯƠNG PHÁP PHÁT HIỆN TẤN CÔNG TRÊN WEB DỰA TRÊN KỸ THUẬT PHÂN TÍCH HÀNH VI

2.1. Giới thiệu về phương pháp phát hiện tấn công web

2.1.1. Một số phương pháp phát hiện tấn công web



Hình 2.1: Phân loại phương pháp phát hiện tấn công web

2.1.2. Công cụ phát hiện tấn công web

2.1.2.1. Sử dụng tường lửa WAF

2.1.2.2. Sử dụng hệ thống phát hiện xâm nhập

❖ **Chức năng của IDS**

❖ **Kiến trúc của hệ thống phát hiện xâm nhập IDS**

❖ **Cách thức làm việc của IDS**

Ưu điểm của Network-Based IDSs:

- Quản lý được cả một network segment (gồm nhiều host).

- "Trong suốt" với người sử dụng lẫn kẻ tấn công.
- Cài đặt và bảo trì đơn giản, không ảnh hưởng tới mạng.
- Tránh tấn DOS ảnh hưởng tới một host nào đó.
- Có khả năng xác định lỗi ở tầng Network (trong mô hình OSI).
- Độc lập với OS.

Hạn chế của Network-Based IDSs:

- Có thể xảy ra trường hợp báo động giả (false positive), tức không có intrusion mà NIDS báo là có intrusion.
- Không thể phân tích các traffic đã được mã hóa (vd: SSL, SSH, IPSec…).
- NIDS đòi hỏi phải được cập nhật các signature mới nhất để thực sự an toàn.
- Có độ trễ giữa thời điểm bị attack với thời điểm phát báo động. Khi báo động được phát ra, hệ thống có thể đã bị tổn hại.
- Không cho biết việc attack có thành công hay không.
- Giới hạn băng thông.

*Ưu điểm của Host - Based IDS***:**

Hạn chế của Host - Based IDS:

2.1.2.3. Công cụ phần mềm dò quét

2.2. Phương pháp phát hiện hành vi bất thường người dùng web sử dụng học máy

2.2.1. Một số thuật toán phát hiện tấn công web

2.2.1.1. Phương pháp học có giám sát sử dụng SVM (SVM- Support vector machine)

2.2.1.2. Decision Tree

2.2.1.3. Random Forest

2.2.1.4. KNN

2.2.2. Lựa chọn và trích xuất hành vi người dùng web

2.2.2.1. Mô tả bộ dữ liệu

Trong luận văn, tác giả trích xuất hành vi bất thường từ bộ dữ liệu về tấn công web CSIC 2010.

Bảng 2.1: Mô tả các trường dữ liệu trong bộ dữ liệu CSIC

Cột dữ liệu	**Mô tả**
index	Số thứ tự
method	Phương thức cho HTTP/1.1 như GET, HEAD, POST, PUT, …
url	Đường dẫn hay địa chỉ dùng để tham chiếu đến các tài nguyên trên mạng Internet
userAgent	Là một chuỗi nhận dạng của trình duyệt web khi gửi yêu cầu đến máy chủ web
cacheControl	Tối ưu tốc độ tải trang, tăng tính bảo mật
accept	Là kiểu dữ liệu mà sẽ nhận được từ response, response mà dại trả về khác kiểu thì sẽ bị ban ngay. Thường thấy nhất là các kiểu text/html, application/xhtml+xm
acceptEncoding	Khai báo kiểu mã hóa nội dung mà request chấp nhận
acceptCharset	Sử dụng để chỉ các bộ thiết lập ký tự nào được chấp nhận

Cột dữ liệu	Mô tả
acceptLanguage	Sử dụng để chỉ ngôn ngữ nào được chấp nhận
host	Địa chỉ IP máy chủ
contentLength	Chỉ dẫn kích cỡ của phần thân đối tượng, trong số thập phân của hệ 8, được gửi tới người nhận
contentType	Là kiểu thông tin mà server trả về cho client, nó phải phù hợp với cái accept mà client request tới
cookie	Chứa thông tin được mã hóa dùng để gửi lên server, giúp xác định phiên giữa client-server
payload	Chứa dữ liệu và các tham số của người dùng gửi lên

Thông thường trong bài toán phân tích hành vi người dùng để xác định bất thường, sẽ tập trung chủ yếu vào các trường dữ liệu người dùng nhập vào. Đối với tập dữ liệu CSIC đã thu thập luận văn sẽ tập trung vào trường payload, url và cookie để xây dựng bộ feature.

2.2.2.2. Trích chọn thuộc tính sử dụng kỹ thuật TF-IDF (Term Frequency – Inverse Document Frequency)

❖ Ứng dụng N-Gram trong trích xuất kí tự và từ trong văn bản

Mô hình ngôn ngữ thống kê cho phép gán (ước lượng) xác suất cho một chuỗi m phần tử (thường là từ) P($w_1 w_2 \dots w_m$) tức là cho phép dự đoán khả năng một chuỗi từ xuất hiện trong ngôn ngữ đó. Theo công thức Bayes:

$$\mathbf{P(AB) = P(B|A) * P(A)}$$

Trong đó:

- P(A): Xác suất xảy ra sự kiện A

- P(B): Xác suất xảy ra sự kiện B
- P(B|A): Xác suất (có điều kiện) xảy ra sự kiện B nếu biết rằng sự kiện A đã xảy ra.

Từ đó ta được:

$$P(w_1 w_2 \dots w_m) = P(w_1) * P(w_2|w_1) * P(w_3|w_1 w_2) * \dots * P(w_m|w_1 w_2 \dots w_{m-1})$$

Theo công thức này thì bài toán tính xác suất của mỗi chuỗi từ quy về bài toán tính xác suất của một từ với điều kiện biết các từ trước nó (có thể hiểu $P(w_1 = P(w_1|start)$ là xác suất để w_1 đứng đầu chuỗi hay nói cách khác người ta có thể đưa thêm ký hiệu đầu dòng start vào mỗi chuỗi).

Trong thực tế, dựa vào giả thuyết Markov người ta chỉ tính xác suất của một từ dựa vào nhiều nhất n từ xuất hiện liền trước nó, và thông thường n= 0,1,2,3. Vì vậy, nhiều người gọi mô hình ngôn ngữ là mô hình N-gram, trong đó n là số lượng từ (bao gồm cả từ cần tính và các từ ngữ cảnh phía trước).

- Với n = 1, unigram.

- Với n = 2, ta có khái niệm bigram.

- Với n = 3, ta có trigram.

Nhưng vì n càng lớn thì số trường hợp càng lớn nên thường người ta chỉ sử dụng với n = 1,2 hoặc đôi lúc là 3.

Theo công thức Bayes, mô hình ngôn ngữ cần phải có một lượng bộ nhớ vô cùng lớn để có thể lưu hết xác suất của tất cả các chuỗi độ dài nhỏ hơn m. Rõ ràng, điều này là không thể khi m là độ dài của các văn bản ngôn ngữ tự nhiên (m có thể tiến tới vô cùng). Để có thể tính được xác suất của văn bản với lượng bộ nhớ chấp nhận được, ta sử dụng xấp xỉ Markov bậc n:

$$P(w_1 w_2 \dots w_m) = P(w_1) * P(w_2|w_1) * P(w_3|w_1 w_2) * \dots$$

$$* P(w_{m-1}|w_{m-n-1} w_{m-n} \dots w_{m-2}) * P(w_m|$$

$$w_{m-n} w_{m-n+1} \dots w_{m-1})$$

Với công thức này, ta có thể xây dựng mô hình ngôn ngữ dựa trên việc thống kê các cụm có ít hơn n+1 từ. Các mô hình N-gram được hình dung thông qua ví dụ sau:

❖ **TF-IDF**

Term Frequency – Inverse Document Frequency (TF-IDF) là giải pháp đánh trọng số kết hợp tính chất quan trọng của một từ trong tài liệu chứa nó (TF- tần suất xuất hiện của từ trong tài liệu) với tính phân biệt của từ trong tập tài liệu nguồn (IDF- nghịch đảo tần suất tài liệu). Đây là một kỹ thuật cơ bản và thường được sử dụng kết hợp với các thuật toán khác để xử lý văn bản. Mục đích của kỹ thuật này là tính trọng số của một từ, qua đó đánh giá mức độ quan trọng của từ đó trong văn bản. Trong đó:

- **TF** được tính theo công thức:

$$\mathbf{tf}(t,d) = \frac{\mathbf{f}(\mathbf{t},\mathbf{d})}{\mathbf{max}\{\mathbf{f}(\mathbf{w},\mathbf{d}) : \mathbf{w} \epsilon \mathbf{d}}$$

Trong đó:

- tf(t, d): tần suất xuất hiện của từ t trong văn bản d
- f(t, d): Số lần xuất hiện của từ t trong văn bản d
- max({f(w, d) : w ∈ d}): Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d

- **IDF** được tính theo công thức:

$$\mathbf{idf}(t,D) = \mathbf{log}\ \frac{|\mathbf{D}|}{|\{\mathbf{d} \epsilon \mathbf{D} : \mathbf{t} \epsilon \mathbf{d}\}|}$$

Trong đó:

- idf(t, D): giá trị idf của từ t trong tập văn bản
- |D|: Tổng số văn bản trong tập D
- |{d ∈ D : t ∈ d}|: thể hiện số văn bản trong tập D có chứa từ t.

- Giá trị **TF-IDF**:

$$\textbf{tfidf(t, d, D) = tf(t, d) x idf(t, D)}$$

Ví dụ trích chọn thuộc tính sử dụng kết hợp N-Gram và TF-IDF cho request người dùng: http://localhost:8080?id=abc';+drop+table+usuarios;. Thu được kết quả trong bảng 2.2.

Bảng 2.2: Kết quả trích chọn thuộc tính sử dụng kết hợp N-Gram và TF-IDF

	tfidf		tfidf		tfidf
tab	0.23375	**rop**	0.138057	**=ab**	0.138057
ble	0.23375	**st:**	0.138057	**?id**	0.138057
abl	0.23375	**t:8**	0.138057	**alh**	0.138057
abc	0.23375	**tp:**	0.138057	**bc'**	0.138057
';+	0.138057	**dro**	0.138057	**bct**	0.138057
lho	0.138057	**cal**	0.138057	**c';**	0.138057
e+a	0.138057	**cta**	0.138057	**ttp**	0.138057
hos	0.138057	**+ab**	0.138057		
htt	0.138057	**+dr**	0.138057		
id=	0.138057	**+ta**	0.138057		
le+	0.138057	**//l**	0.138057		
le;	0.138057	**/lo**	0.138057		
loc	0.138057	**80**	0.138057		
d=a	0.138057	**0?i**	0.138057		
oca	0.138057	**808**	0.138057		
op+	0.138057	**80?**	0.138057		
ost	0.138057	**://**	0.138057		
p+t	0.138057	**:80**	0.138057		
p:/	0.138057	**;+d**	0.138057		

Kết luận chương 2

Trong chương 2 luận văn đã giới thiệu tổng quát về các phương pháp phát hiện tấn công web và một số công cụ hỗ trợ phát hiện tấn công. Từ hạn chế của việc sử dụng các công cụ tấn công, luận văn đã đề xuất phương pháp phát hiện hành vi bất thường của người dùng web sử dụng học máy thông qua các thuật toán: SVM, Random Forest, KNN. Luận văn đã đưa ra phương pháp tính và sử dụng kĩ thuật trích chọn thuộc tính trong văn bản TF-IDF để lựa chọn và trích xuất hành vi người dùng đưa ra cảnh báo trước về các cuộc tấn công web cho người quản trị.

Trên cơ sở các kết quả đã đạt được của chương 2, trong chương tiếp theo luận văn sẽ tiến hành thực nghiệm phát hiện tấn công web dựa trên kỹ thuật phân tích hành vi trên cơ sở các thuật toán (SVM, Random Forest, KNN) và hành vi đã trích xuất- lựa chọn.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. Một số yêu cầu cài đặt

3.1.1. Yêu cầu chung cho cài đặt thử nghiệm

- *Phần cứng:* Bộ xử lý 32bit (x86) hoặc 64bit (x64) có tốc độ 2 gigahertz (GHz) hoặc nhanh hơn; RAM 4GB trở lên; Đĩa cứng có dung lọợng trống 10 GB (64 bit).

- *Phần mềm:* Cài đặt trên hệ thống Windows/Linux (Centos 7.2); Công cụ lập trình: Phần mềm Python 2.7 trở lên hoặc phần mềm Pycham Professional 2020.1.

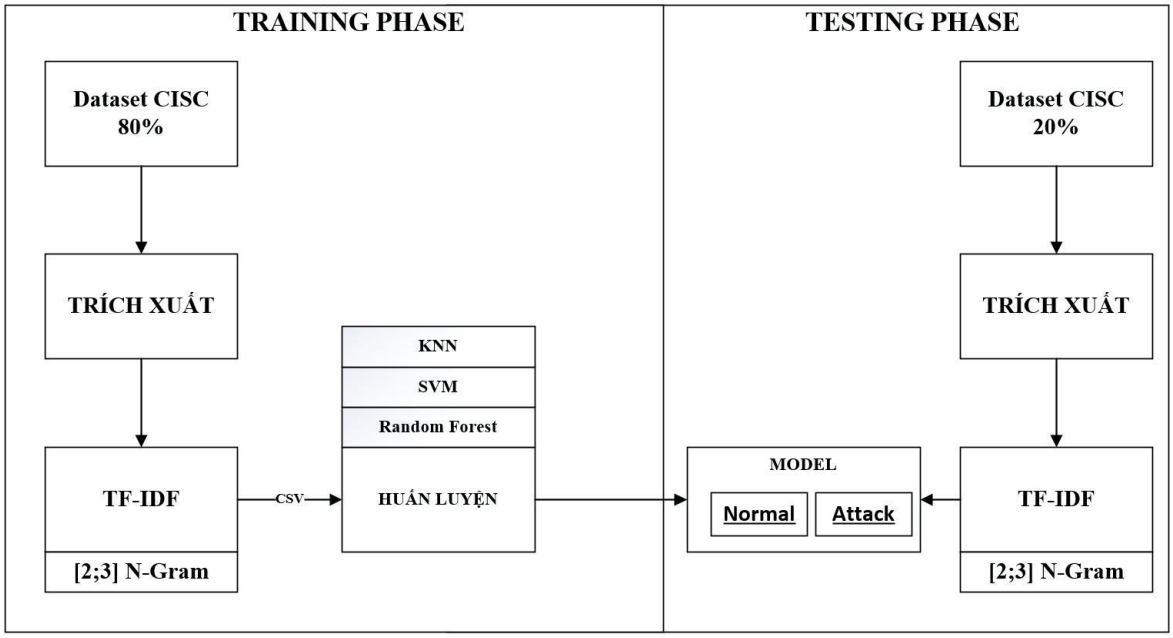
- *Dữ liệu:* CSIC 2010, bộ dữ liệu tấn công thu thập tại website trường Đại học Công nghệ Giao thông vận tải.

3.1.2. Giới thiệu chung về Python

3.2. Kịch bản thực nghiệm

Bộ dữ liệu CSIC đầu vào sẽ được chia thành nhiều tập khác nhau để kiểm nhiệm mô hình. Quá trình xây dựng mô hình bao gồm hai giai đoạn chính:

- Giai đoạn 1: Huấn luyện mô hình (Training phase)
- Giai đoạn 2: Kiểm thử mô hình (Testing phase).



Hình 3.1: Quá trình xây dựng mô hình

❖ **Giai đoạn huấn luyện mô hình (bao gồm 3 bước chính):**

- Bước 1: Bộ dữ liệu các request bình thường từ người dùng trong tập dữ liệu CSIC. Tại bước này, thực hiện tính toán sự xuất hiện của các ký tự quan trọng mới và lưu chúng trong cơ sở dữ liệu.
- Bước 2: Mô-đun không gian vectơ được sử dụng để chuyển đổi dữ liệu chuỗi thành các vectơ. Sử dụng kĩ thuật trích chọn thuộc tính TF-IDF kết hợp N-Gram.
- Bước 3: Mô-đun xử lý dữ liệu sử dụng thuật toán học máy (lần lượt thay thế các thuật toán khác nhau để xác định mô hình tối ưu nhất cho bài toán: KNN, SVM, Random Forest).

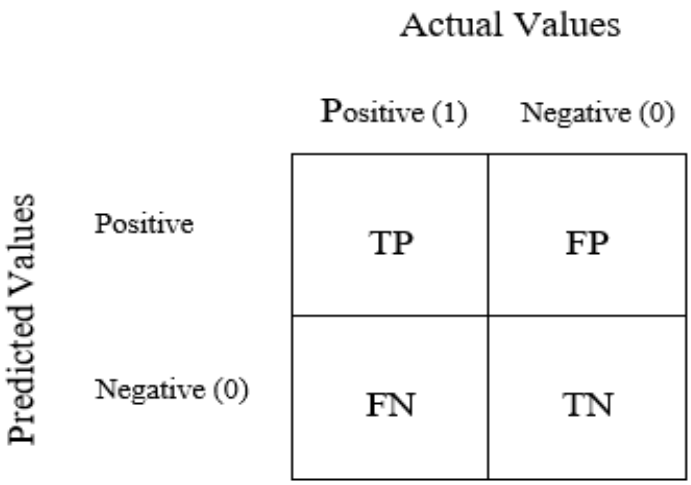
❖ **Giai đoạn kiểm thử mô hình:**

- Bước 1: Phần dữ liệu thử nghiệm được tiến hành loại bỏ nhãn dữ liệu.
- Bước 2: Thực hiện quá trình trích xuất đặc trưng dữ liệu tương tự bước 2 ở giai đoạn 1.
- Bước 3: Thử nghiệm các mô hình ứng với các thuật toán học máy đã được xây dựng ở giai đoạn 1. Tác giả lựa chọn phương pháp đánh giá độ chính xác bằng cách sử dụng ma trận độ đo (confusion matrix) và $F1_{score}$ được mô tả như sau:

Confusion Matrix là một phương pháp đánh giá kết quả của những bài toán phân loại với việc xem xét cả những chỉ số về độ chính xác và độ bao quát của các dự đoán cho từng lớp. Một confusion matrix gồm 4 chỉ số sau đối với mỗi lớp phân loại:

- **TP (True Positive)**: mẫu mang nhãn dương được phân lớp đúng vào lớp dương
- **TN (True Negative)**: mẫu mang nhãn âm được phân lớp đúng vào lớp âm.
- **FP (False Positive - Type 1 Error)**: mẫu mang nhãn âm bị phân lớp sai vào lớp dương.

- **FN (False Negative - Type 2 Error)**: mẫu mang nhãn dương bị phân lớp sai vào lớp âm.



Hình 3.2: Ma trận độ đo (Conusion matrix)

Ký hiệu TP là True Positive; TN là True Negative; FP là False Positive và FN là False Negative. Thực hiện phép đo Precision – Recall, trong đó, Precision là tỉ lệ số điểm TP trong những điểm được phân loại Positive, còn Recall là tỉ lệ số điểm TP trong số điểm thực sự là Positive. Công thức như sau:

$$\text{precision} = \frac{TP}{TP+FP}$$

$$\text{recall} = \frac{TP}{TP+FN}$$

Ta thấy rằng, Precision và Recall phủ càng cao thì càng tốt. Nhưng trong thực tế, hai giá này không thể đạt được cực đại cùng một lúc và thông thường phải tìm kiếm sự cân bằng. Thước đo $F1_{score}$ là trung bình hài hòa giữa Precision và Recall. Nó có xu hướng bằng không nếu hai giá trị này có xu hướng bằng không.

$$F1_{score} = 2 * \frac{precision \times recall}{precision + recall}$$

❖ **Kịch bản thực nghiệm chi tiết:**

Trong mô hình này đã sử dụng bộ dữ liệu bao gồm 25065 liên kết bất hợp pháp của một số loại tấn công (XSS, SQL injection) và 36000 liên kết hợp pháp. Bộ

dữ liệu từ một số nguồn dữ liệu của các công cụ bảo vệ hệ thống như tệp nhật ký của hệ thống phát hiện và ngăn chặn xâm nhập, yêu cầu HTTP (phương thức GET, phương thức POST) của tường lửa ứng dụng Web.

Các bộ dữ liệu ban đầu đã được thực hiện phân chia thành hai phần riêng biệt với 80% các liên kết để đào tạo và 20% các liên kết để thử nghiệm. Trong quá trình thử nghiệm thêm một số phương pháp học máy để so sánh phương pháp đã đề xuất.

3.3. Một số kết quả thực nghiệm

Thực hiện thử nghiệm trên dữ liêu bao gồm:

- 36000 request bình thường;
- 25065 request bất thường;
- Tỉ lệ chia dữ liệu Traning/Testing là 8/2;
- Số lớp dữ liệu cần phần là 2 lớp: Bình thường/Bất thường.

Từ việc thực hiện phân chia dữ liệu đầu vào của tập bình thường thành các đoạn với tỉ lệ như trên, ta được bảng kết quả:

Bảng 3.1: Kết quả thực nghiệm xây dựng bộ phân lớp bình thường/bất thường theo kịch bản

	KNN				SVM				Rừng ngẫu nhiên			
	F1_Score	**Confusion Matrix**	**Precision**	**Recall**	**F1_Score**	**Confusion Matrix**	**Precision**	**Recall**	**F1_Score**	**Confusion Matrix**	**Precision**	**Recall**
N=2	0.9695	[7133 56] [313 4711]	0.9922	0.9579	0.9807	[7090 99] [136 4888]	0.9862	0.9811	0.9837	[7090 99] [136 4888]	0.9862	0.9811
N=3	0.9767	[3594 11] [131 2371]	0.9969	0.9648	**0.9975**	[3598 7] [8 2494]	0.9980	0.9977	0.9862	[3598 7] [8 2494]	0.9980	0.9977
N=4	0.9692	[7148 41] [326 4698]	0.9942	0.9563	0.9967	[7171 18] [22 5002]	0.9974	0.9969	0.9752	[7171 18] [22 5002]	0.9974	0.9969

Nhận xét: Kết quả sau khi chạy với 3 thuật toán học máy ta thu được mô hình tốt nhất với thuật toán SVM và Ngram = 3.

Phát hiện bất thường từ hành vi người dùng web là một vấn đề khó trong phòng chống tấn công ứng dụng web. Thuật toán phân loại được đề xuất để phát hiện các liên kết bất hợp pháp dựa trên ứng dụng phương pháp học máy với việc trích chọn các đặc trưng thuộc tính dữ liệu của người dùng. Thuật toán phát hiện liên kết bất hợp pháp phân tích các liên kết theo một chuỗi các bước để xác định xem liên kết đó là hợp pháp hay độc hại. Mặc dù thuật toán đề xuất cải thiện độ chính xác phân loại của các liên kết bất hợp pháp, nhưng với sự gia tăng số lượng tham số có trong các yêu cầu, độ chính xác phân loại sẽ giảm. Do đó, trong thời gian tới, cần tìm sự kết hợp của các phương pháp phát hiện bất thường dựa trên học sâu nhằm cải thiện độ chính xác phân loại không chỉ của các liên kết đáng ngờ mới đặc trưng các loại tấn công chưa được định danh.

Kết luận chương 3

Trong chương 3 luận văn đã xây dựng ba kịch bản thử nghiệm phân loại hành vi người dùng web. Với mỗi kịch bản đã xây dựng được mô hình học máy như: KNN, SVN, Random Forest.

Các kết quả thử nghiệm ban đầu cho thấy giải pháp phát hiện tấn công web ứng dụng dựa trên kỹ thuật phân tích hành vi đề xuất có tính khả thi cao và phù hợp với các yêu cầu đề ra.

KẾT LUẬN

1. Những đóng góp của luận văn

Với mục tiêu nghiên cứu các phương pháp phát hiện tấn công web ứng dụng dựa trên kỹ thuật phân tích hành vi và thử nghiệm, luận văn đã đi sâu nghiên cứu các vấn đề xung quanh đề tài nghiên cứu, các thuật toán học máy phát hiện tấn công web để ứng dụng vào phát hiện hành vi bất thường của người dùng.

Những kết quả chính đã đạt được trong luận văn:

- Khảo sát một số nguy cơ mất an toàn thông tin thông qua các kĩ thuật tấn công web, đưa ra các phương pháp phòng chống tấn công web phổ biến cũng như đưa ra một số phương pháp nhằm nâng cao bảo mật hệ thống.
- Tìm hiểu phương pháp phát hiện tấn công web dựa trên kỹ thuật phân tích hành vi. Thực hiện trích xuất hành vi bất thường từ bộ dữ liệu về tấn công web (bộ dữ liệu CSIC 2010) sử dụng kĩ thuật trích chọn TF-IDF kết hợp N-Gram.
- Lựa chọn và ứng dụng thuật toán học máy nhằm phân loại hành vi tấn công và hành vi bình thường lên web, sử dụng các thuật toán học máy có giám sát: KNN, SVM, Random forest.
- Thực nghiệm xây dựng mô hình bộ phân lớp bình thường/bất thường theo từng kịch bản để đưa ra mô hình tốt nhất khi sử dụng N-Gram với n=3.

2. Hướng phát triển của luận văn

Một số hướng phát triển tiếp theo của luận văn:

- Mặc dù thuật toán đề xuất cải thiện độ chính xác phân loại của các liên kết bất hợp pháp, nhưng với sự gia tăng số lượng tham số có trong các yêu cầu, độ chính xác phân loại sẽ giảm. Do đó, cần tìm sự kết hợp của các phương pháp phát hiện bất thường dựa trên học sâu nhằm cải thiện độ chính xác phân

loại không chỉ của các liên kết đáng ngờ mới đặc trưng các loại tấn công chưa được định danh.

- Thực hiện nghiên cứu phương pháp phát hiện tấn công web dựa trên kỹ thuật phân tích hồ sơ hành vi.