HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

VONGSAVANH VANPHATH

NGHIÊN CỨU PHƯƠNG PHÁP PHÁT HIỆN THAY ĐỔI NỘI DUNG BẢNG KẾT QUẢ CỦA TRANG TIN XỔ SỐ KIẾN THIẾT

Chuyên ngành : HỆ THỐNG THÔNG TIN
Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2020

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học:

PGS.TSKH. HOÀNG ĐĂNG HẢI

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 2020

Có thể tìm hiểu luận văn tại:

1. Thư viện Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Trong những năm gần đây, công nghệ thông tin và truyền thông có vai trò lớn đối với sự phát triển của mỗi quốc gia, mỗi doanh nghiệp. Ứng dụng CNTT&TT cũng có tác động không nhỏ đến đời sống kinh tế, xã hội của đại bộ phận người dân trên thế giới. CNTT&TT cũng góp phần quan trọng trong vấn đề an ninh và phát triển bền vững của mỗi quốc gia. Do vậy, ứng dụng CNTT&TT trở thành một phần không thể thiếu trong chiến lược phát triển của các doanh nghiệp và các quốc gia trên thế giới.

Với tốc độ phát triển và ứng dụng của CNTT&TT ngày càng nhanh như hiện nay, hàng ngày có một lượng lớn thông tin được lưu trữ, truyền tải thông qua các trang thông tin điện tử (TTĐT) cũng kéo theo nhiều rủi ro về sự mất an toàn thông tin. Thiệt hại do mất an ninh an toàn trên các trang TTĐT đã tăng rất nhanh và sẽ ảnh hưởng nghiêm trọng đến sự phát triển kinh tế- xã hội, nếu công tác đảm bảo an ninh an toàn không được triển khai đúng mức. Bởi các kỹ thuật của tội phạm mạng ngày càng cao và tinh vi hơn, số lượng điểm yếu an ninh ngày càng tăng, số vụ xâm phạm an toàn mạng ngày càng nhiều.

Trước những nguy cơ tấn công mạng ngày càng gia tăng vào các trang TTĐT, việc bảo đảm an toàn cho trang TTĐT là hết sức cần thiết. Một nguy cơ có thể xảy ra là nội dung thông tin trên trang có thể bị tin tặc tấn công, giả mạo bằng cách thay đổi thông tin. Ví dụ giả mạo kết quả trên trang tin kết quả xổ số có thể gây ra những tác hại rất lớn.

Do vậy, việc nghiên cứu phương pháp phát hiện thay đổi nội dung trang thông tin điện tử, cụ thể là cho một trang TTĐT về kết quả xổ số là hết sức cần thiết. Đó cũng là lý do học viên xin chọn đề tài: "**Nghiên cứu phương pháp phát hiện thay đổi nội dung bảng kết quả của trang tin xổ số kiến thiết"** làm đề tài cho luận văn nghiên cứu của mình.

Luận văn bao gồm 3 chương, bố cục các chương và các mục đi kèm như sau:

Chương 1: Tổng quan về vấn đề nghiên cứu

Khái niệm an toàn thông tin nhằm mục đích chính bảo vệ các khía cạnh tính bí mật, toàn vẹn và sẵn sàng của thông tin. Trong đó tính toàn vẹn chính là khía cạnh mà luận văn này muốn nghiên cứu, để xác định các nguy cơ thay đổi, giả mạo nội dung trang TTĐT.

Chương 2: Nghiên cứu phương pháp kiểm tra phát hiện thay đổi nội dung trang tin xổ số

Đảm bảo tính toàn vẹn của thông tin, tức là thông tin chỉ được phép xóa hoặc sửa đổi bởi những đối tượng được phép và phải đảm bảo rằng thông tin vẫn còn chính xác khi được lưu trữ hay truyền đi. Ví dụ trường hợp tính toàn vẹn của thông tin bị phá vỡ: thay đổi kết quả xổ số trên trang xổ số kiến thiết từ một đối tượng không được phép dẫn đến nhiều hệ lụy. Chương này trình bày cụ thể về kiến trúc, cơ chế hoạt động của trang TTĐT cùng với mô hình tổng quát cho phương pháp kiểm tra phát hiện giả mạo nội dung trang tin. Bằng

cách trình bày cụ thể phương pháp thu thập thông tin, chọn lọc nội dung thông tin cần kiểm tra đối với trang TTĐT, phân tích các công cụ thu thập thông tin sẽ đưa ra phương pháp kiểm tra phát hiện giả mạo nội dung trang kết quả xổ số.

Chương 3: Cài đặt và thử nghiệm

Chương này nhằm hiện thực hóa các kết quả đã nghiên cứu, sẽ tiến hành triển khai thử nghiệm thu thập nội dung thông tin, ghi thông tin, kiểm tra phát hiện thay đổi nội dung trang tin kết quả xổ số.

CHƯƠNG 1. TỔNG QUAN VỀ VẤN ĐỀ NGHIÊN CỨU.

Khái niệm an toàn thông tin nhằm mục đích chính bảo vệ các khía cạnh tính bí mật, toàn vẹn và sẵn sàng của thông tin. Trong đó tính toàn vẹn chính là khía cạnh mà luận văn này muốn nghiên cứu, để xác định các nguy cơ thay đổi, giả mạo nội dung trang TTĐT.

1.1. Vấn đề an toàn thông tin: cần nêu các mối nguy cơ, tác động đến trang thông tin điện tử nói chung

Các nguy cơ đe dọa an toàn thông tin:

- An toàn thông tin được đánh giá bằng hai chỉ số: xác suất ngăn chặn các nguy cơ và thời gian đảm bảo mức độ an toàn xác định.
- Vì thông tin được chứa trong các tham số thông tin của vật mang, nên để đảm bảo an toàn thông tin, các tham số này phải giữ được giá trị của nó trong khoảng thời gian nhất định.
- Thông tin thường bị đe dọa lấy cắp, thay đổi hay bị xóa một cách vô tình hay cố ý.
- Để bảo vệ thông tin có hiệu quả, cần ước lượng giá trị của nguy cơ đe dọa an toàn thông tin. Giá trị của một nguy cơ cụ thể đối với thành phần thông tin xem xét đầu tiên trong mọi trường hợp có thể biểu thị dưới dạng tích của các thiệt hại tiềm ẩn do thực trạng nguy cơ về yếu tố thông tin đầu tiên với xác suất thực tế thể hiện nó.
- Việc nhận giá trị định lượng tương đối chính xác và khách quan của các thành phần là phức tạp.

Từ những phân tích trên đây có thể thấy rằng, việc đánh giá một cách đầy đủ các nguy cơ về an toàn thông tin đối với nguồn tài nguyên thông tin của mỗi cơ quan, tổ chức là bước đi cần thiết để có thể xây dựng các chính sách, giải pháp bảo vệ thông tin một cách hữu hiệu

1.2. Vấn đề bảo đảm an toàn trang TTĐT nói chung

Trong cổng/trang TTĐT thường có các thành phần cho người dùng nhập dữ liệu vào như mục đăng nhập, tìm kiếm, bình luận, liên kết đến bài viết, v.v. Ngoài việc giúp cho người dùng dễ dàng tương tác với cổng/trang TTĐT, các mục này nếu không được kiểm soát chặt chẽ sẽ trở thành một nguy cơ lớn để tin tặc thực hiện các cuộc tấn công. Bởi vậy, trước khi đưa cổng/trang TTĐT vào hoạt động chính thức cần sử dụng các công cụ phần mềm để tìm và kiểm tra tất cả các lỗ hổng có thể bị kẻ xấu khai thác. Từ đó tìm cách khắc phục những lỗ hổng trên cổng/trang TTĐT của mình để đảm bảo an ninh an toàn.

Ngoài ra có thể sử dụng biểu thức chính quy áp dụng cho tất cả các ngôn ngữ lập trình để thực hiện các công việc này.

Sau khi đã xác định được các lỗi trên cổng/trang TTĐT của mình, cũng cần phân loại để đưa ra những giải pháp phòng chống thích hợp. Việc phân loại các lỗi và các kiểu tấn

công thành các nhóm khác nhau sẽ giúp người quản trị dễ dàng xác định các nguy cơ cũng như biện pháp đối phó. Sau đây là một số lỗi phổ biến trên các cổng/trang TTĐT nói riêng và ứng dụng web nói chung, có thể bị khai thác để tấn công.

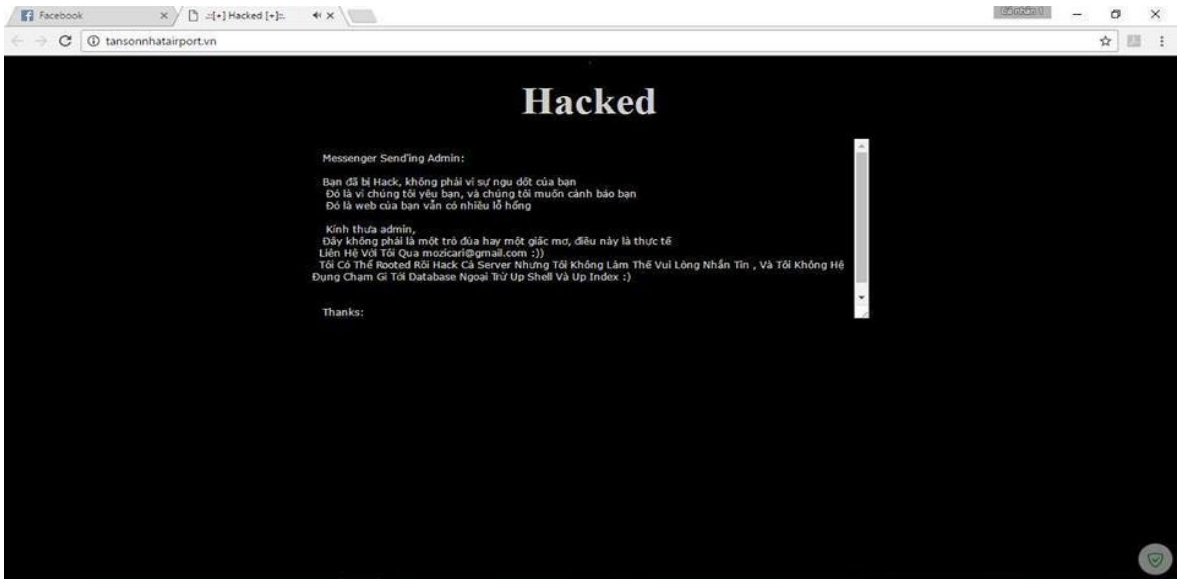
- Các lỗi Injection
- Các lỗi Cross-Site-Scripting (XSS)
- Các lỗi quản lý xác thực và quản lý phiên làm việc
- Các lỗi đối tượng tham chiếu không an toàn
- Các lỗi cấu hình thiếu an toàn
- Các lỗi lưu trữ dữ liệu thiếu an toàn.
- Các lỗi Cross Site Request Forgery (CSRF)
- Các lỗi do ứng dụng sử dụng những thành phần chứa lỗi bảo mật.
- Các lỗi trong việc kiểm soát quyền truy cập
- Một số hình thức tấn công nhằm vào SSO.

1.3. Nguy cơ thay đổi, giả mạo nội dung trang TTĐT nói chung.

Tấn công Deface là tấn công thay đổi nội dung, hacker sẽ thông qua một điểm yếu nào đó để thay đổi nội dung trang TTĐT của nạn nhân.

Có rất nhiều nguyên nhân trang TTĐT bị Deface, chủ yếu là do trang TTĐT tồn tại nhiều điểm yếu bảo mật nghiêm trọng mà hacker có thể upload file lên server hoặc có quyền đăng nhập vào trang quản trị trang TTĐT (Ví dụ : SQL Injection). Thậm chí nếu trang TTĐT trên hosting an toàn thuộc server bị tấn công thì cũng sẽ bị tấn công Deface luôn (Local Attack).

Các trường hợp trang TTĐT bị tấn công Deface: lỗi SQL injection, lỗi XSS (Cross Site Scripting), lỗ hổng Remote File Include, lỗ hổng Local file inclusion, không cập nhật phiên bản, mật khẩu quản trị yếu



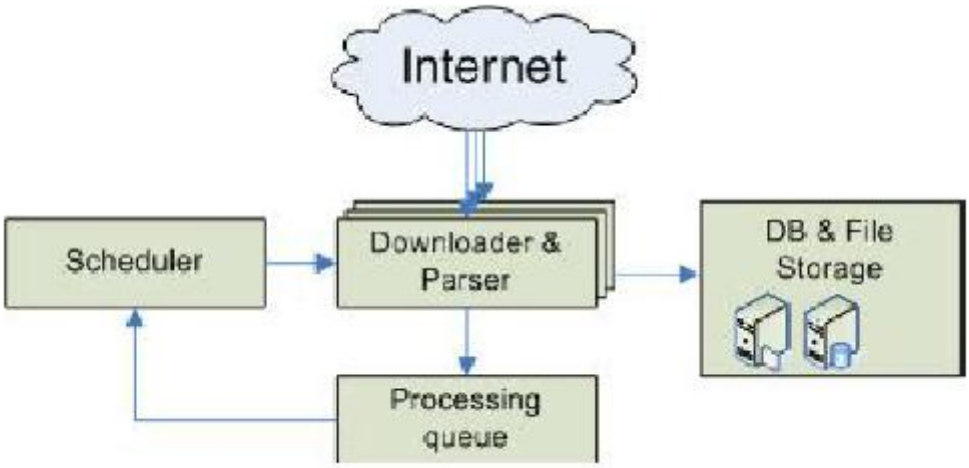
Hình1.1: Màn hình một trang TTĐT bị tấn công

1.4. Các mô hình, phương pháp, kỹ thuật liên quan đến thu thập thông tin, trích chọn dữ liệu.

Hiện nay có 2 phương pháp chính dùng để thu thập dữ liệu: API và Trang (Sites).

1.4.1. Web Crawler

Một Web Crawler là một chương trình máy tính có thể "duyệt web" một cách tự động và theo một phương thức nào đó được xác định trước. Vì là một chương trình nên quá trình " duyệt web" của các web crawler không hoàn toàn giống với quá trình duyệt web của con người (Web crawler phải sử dụng các phương thức dựa trên HTTP trực tiếp chứ không thông qua web browser như con người).



Hình 1.2. Sơ đồ hoạt động của một web crawler đơn giản.

Về bản chất, web crawling chính là quá trình duyệt đệ quy một đồ thị cây có các node là các web page. Tùy thuộc vào chiến lược của crawler, các node có thể được duyệt theo chiều sâu hoặc duyệt theo chiều rộng. Trong thực tế, quá trình crawling web sẽ phải đối diện với rất nhiều vấn đề khó khăn như: kích thước khổng lồ của word wide web, các trang web HTML được viết không chuẩn, hạn chế ghé thăm một URL đã được ghé thăm trước đó, các trang web động, nội dung các trang web được cập nhật thường xuyên,...

1.4.2. Web Scraper

Web Scraper là một thuật ngữ để chỉ các phần mềm có khả năng bóc tách và trích xuất thông tin chứa trên các web page một cách tự động. Công việc này được gọi là web scraping, web harvesting hoặc web data extraction. Các web scraper khác với web crawler ở chỗ, trong khi web crawler tập trung vào việc duyệt các trang web thông qua các liên kết hyperlink, thì web scraper lại tập trung vào việc chuyển đổi nội dung có cấu trúc, sau đó bóc tách, trích xuất phần thông tin mong muốn và lưu trữ lại vào các cơ sở dữ liệu hoặc spreadsheet.

Một số kỹ thuật được sử dụng trong web scraping có thể kể ra như: so trùng, lập trình HTTP, phân tích cấu trúc DOM.

Một số ứng dụng quan trọng của Web Scraping: E-commerce Websites (Website thương mại điện tử), content Aggregators (Bộ tổng hợp nội dung), Marketing and Sales Campaigns (Chiến dịch tiếp thị và bán hàng), search Engine Optimization- SEO (Tối ưu hóa công cụ tìm kiếm), Data for Machine Learning Project (Dữ liệu cho các dự án máy học).

1.4.3. Phân biệt Web Crawling và Web Scraping

Web Crawling là quá trình thu thập thông tin từ các Website trên mạng Internet theo các đường links cho trước. Các Web Crawler sẽ truy cập các links này để download toàn bộ nội dung của trang web cũng như tìm kiếm thêm các đường links bên trong để tiếp tục truy cập và download nội dung từ các đường links này. Dữ liệu sau khi được tải về sẽ được đánh chỉ số (indexing) rồi lưu vào cơ sở dữ liệu.

Web Scraping cũng thực hiện việc tìm kiếm và thu thập thông tin nhưng khác với Web Crawling, Web Scraping không thu thập toàn bộ thông tin của một trang web mà chỉ thu thập những thông tin cần thiết, phù hợp với mục đích của người dùng. Trong Web Scraping chúng ta cũng phần nào sử dụng WebCrawler để thu thập dữ liệu, kết hợp với Data Extraction (trích xuất dữ liệu) để tập trung vào các nội dung cần thiết.

Ví dụ như đối với trang amazon.com, Web Crawling sẽ thu thập toàn bộ nội dung của trang web này (tên các sản phẩm, thông tin chi tiết, bảng giá, hướng dẫn sử dụng, các reviews và comments về sản phẩm,…). Tuy nhiên Web Scraping có thể chỉ thu thập thông tin về giá của các sản phẩm để tiến hành so sánh giá này với các trang bán hàng online khác.

1.5. Một số thuật toán kiểm tra phát hiện thay đổi nội dung trang TTĐT

1.5.1. Hàm băm

1.5.1.1. Giới thiệu hàm băm

Hàm băm (hash function) là giải thuật với đầu vào là những khối dữ liệu và kết quả đầu ra là các giá trị băm tương ứng với mỗi giá trị đầu vào. Ở đây giá trị băm có thể được coi như một khóa để phân biệt các dữ liệu với nhau, tuy vẫn còn hiện tượng trùng khóa hay còn gọi là đụng độ nhưng điều này vẫn được chấp nhận và mọi người vẫn đang tìm cách để cải thiện giải thuật nhằm giảm thiểu sự đụng độ đó. Để giảm chi phí tính toán khi tìm một khối dữ liệu trong một tập hợp, người ta sử dụng bảng băm.

1.5.1.2. Tính một chiều của hàm băm

Hàm băm được xem là hàm một chiều khi cho trước giá trị băm, khó có thể tái tạo lại thông điệp ban đầu, hay còn gọi là "tiền ảnh" ("pre-image"). Thật vậy, với bài toán tìm "tiền ảnh" tương ứng với một giá trị băm, trong trường hợp lý tưởng, cần phải thực hiện hàm băm cho khoảng 2^n thông điệp.

Cách tấn công nhằm tạo ra một thông điệp khác với thông điệp ban đầu nhưng có cùng giá trị băm gọi là tấn công "tiền ảnh thứ hai" (second pre-image attack).

Hàm băm mật mã phải có khả năng chống lại các loại tấn công mật mã, tối thiểu phải đảm bảo có 3 tính chất sau:

+ Kháng tiền ảnh (Pre-image resistance).

+ Kháng tiền ảnh thứ hai (Second pre-image resistance).

+ Kháng xung đột (Collision resistance).

1.5.1.3. Cấu trúc hàm băm

Các hàm băm hầu hết đều có chung cấu trúc giải thuật như sau:

+ Cho dữ liệu đầu vào *M* có độ dài bất kỳ. Có thể thêm vào M một số bit để nhận được dữ liệu có độ dài là bội của hằng số cho trước. Chia nhỏ thông điệp thành từng khối có kích thước bằng nhau: M_1, M_2, …M_s

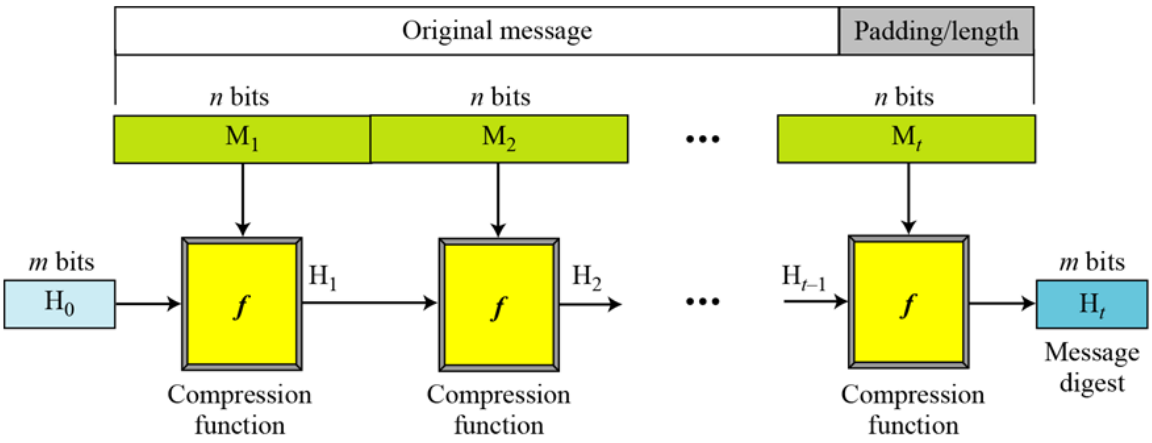
+ Gọi *H* là trạng thái có kích thước *n* bit,

+ Gọi *f* là hàm dùng để trộn khối dữ liệu với trạng thái hiện hành

➢ - Khởi tạo, gán H_0 bằng một vector khởi tạo nào đó

➢ - $H_i = f(H_{i-1}, M_i)$ với $i = 1, 2, 3, \ldots, s$

+ H_s chính là thông điệp rút gọn của thông điệp *M* ban đầu



Hình 1.3 Sơ đồ Merkel-Damgard

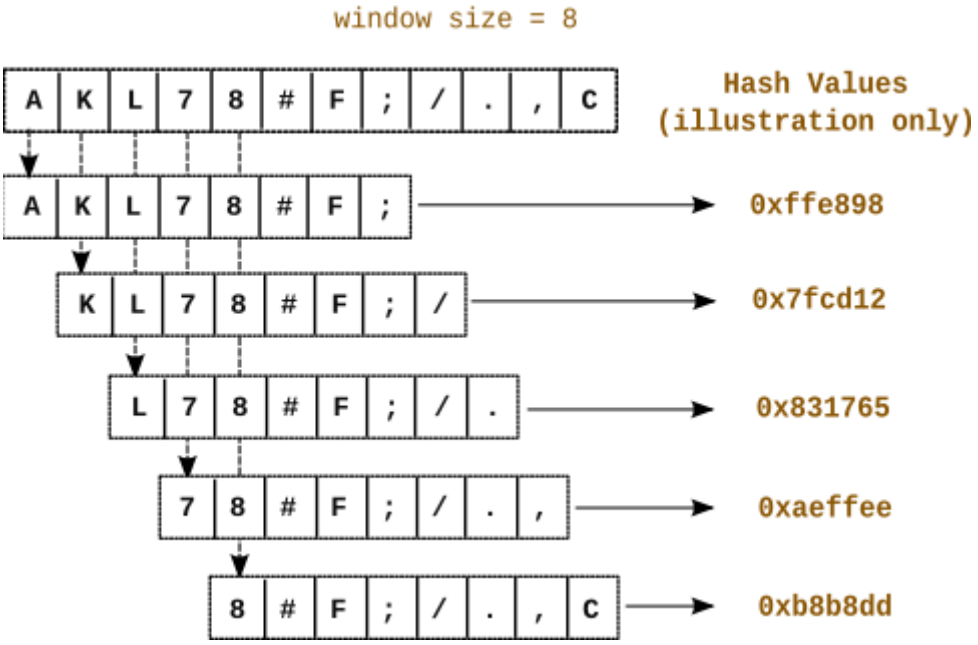
1.5.2. Thuật toán đối sánh chuỗi

Đối sánh chuỗi là việc so sánh một hoặc vài chuỗi (thường được gọi là mẫu hoặc pattern) với toàn bộ văn bản để tìm ra nơi và số lần xuất hiện của chuỗi đó trong văn bản.

1.5.3. Dấu vân tay tài liệu (Document Fingerprint)

Trong khoa học máy tính, dấu vân tay nhận dạng duy nhất dữ liệu gốc cho tất cả các mục đích thực tiễn giống như là việc nhận dạng duy nhất dấu vân tay người trong thực tế. Dấu vân của tài liệu là tập hợp các mã được sinh ra tù các khóa nội dung của tài liệu đó. Mỗi mã đó được gọi là một giá trị băm.

1.5.4. Thuật toán Rabin Fingerprint



Hình 1.4 Mô tả thuật toán Rabin Fingerprint

1.5.5. Thuật toán Rabin Fingerprint cải tiến

Thuật toán Rabin Fingerprint cải tiến áp dụng xây dựng hệ thống giám sát website nhằm phát hiện kịp thời các cuộc tấn công để đảm bảo tính toàn vẹn của trang web đồng thời tạo ra thông điệp cảnh báo có ý nghĩa khi trang web đã bị tấn công.

1.5.6. Thuật toán tìm sự khác nhau của hai văn bản "An O(ND) Difference Algorithm"

Cốt lõi của thuật toán được xây dựng bằng hai phương pháp:

LCS (Longest common subsequence).

SMS (Shortest Middle Snake).

1.5.7. Thuật toán tìm sự khác nhau của hai hình ảnh

Việc tìm sự khác nhau của hai hình ảnh cơ bản là sự so sánh trực tiếp các điểm ảnh của hai ảnh.

+ Cải tiến:

Việc lấy thông số các điểm ảnh trong C# thường sử dụng 2 phương thức set và get, tuy nhiên khi bạn gọi 2 phương thức này hệ thống sẽ Lock ảnh lại đến khi kết thúc phương thức vừa gọi tự động sẽ UnLock ảnh đó cho việt truy cập lần sau. Chính việc Lock rồi Unlock liên tục đã làm đã làm cho việc xử lý ảnh chậm, nhất là với ảnh có kích thước lớn.

Vì vậy thuật toán có thể cải tiến bằng cách sử dụng kỹ thuật LockBits, lưu các thông tin của ảnh vào mảng để xử lý.

1.6. Kết luận chương

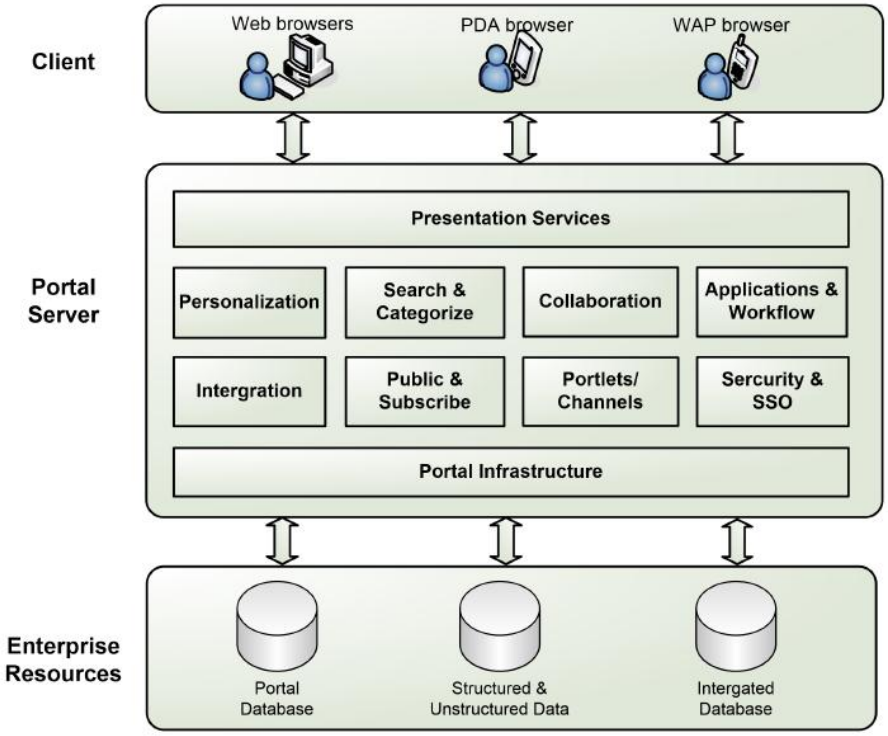
Trong chương 1 luận văn nêu khái niệm tấn công thay đổi nội dung, nguyên nhân và cách khắc phục cùng với một số thuật toán phát hiện sự thay đổi đó.

Việc bị hacker tấn công là điều không thể tránh khỏi vì ngay cả những ông lớn như Google lẫn Facebook cũng đã từng phải chao đảo vì "những vị khách không mời mà đến" này. Tuy nhiên với những kiến thức trên, chúng ta có thể hạn chế được tới 99% các cuộc xâm lăng ngoài ý muốn đó. Suy cho cùng thì tấn công giao diện Deface cũng không quá ghê gớm.

CHƯƠNG 2. NGHIÊN CỨU PHƯƠNG PHÁP KIỂM TRA PHÁT HIỆN THAY ĐỔI NỘI DUNG TRANG TIN XỔ SỐ

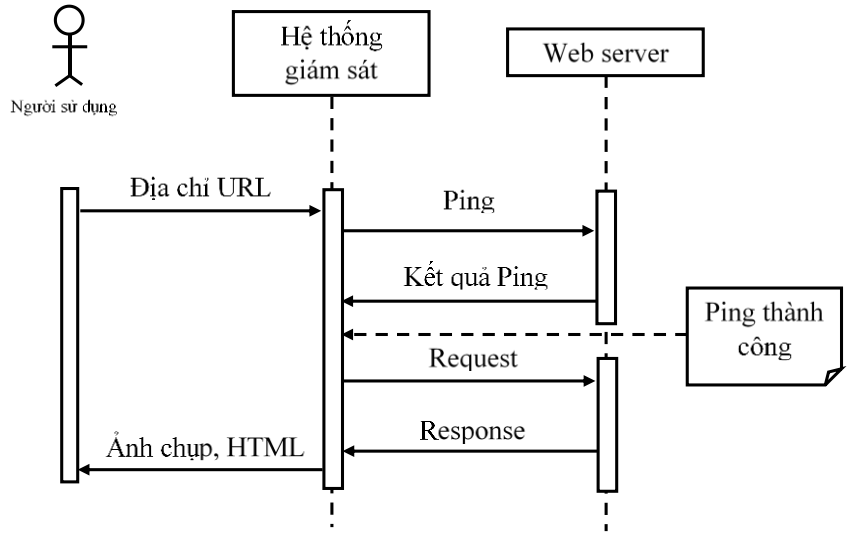
Đảm bảo tính toàn vẹn của thông tin, tức là thông tin chỉ được phép xóa hoặc sửa đổi bởi những đối tượng được phép và phải đảm bảo rằng thông tin vẫn còn chính xác khi được lưu trữ hay truyền đi. Ví dụ trường hợp tính toàn vẹn của thông tin bị phá vỡ: thay đổi kết quả xổ số trên trang xổ số kiến thiết từ một đối tượng không được phép dẫn đến nhiều hệ lụy. Chương này trình bày cụ thể về kiến trúc, cơ chế hoạt động của trang TTĐT cùng với mô hình tổng quát cho phương pháp kiểm tra phát hiện giả mạo nội dung trang tin. Bằng cách trình bày cụ thể phương pháp thu thập thông tin, chọn lọc nội dung thông tin cần kiểm tra đối với trang TTĐT, phân tích các công cụ thu thập thông tin sẽ đưa ra phương pháp kiểm tra phát hiện giả mạo nội dung trang kết quả xổ số.

2.1. Khái quát về kiến trúc chung, cơ chế hoạt động của các trang TTĐT.

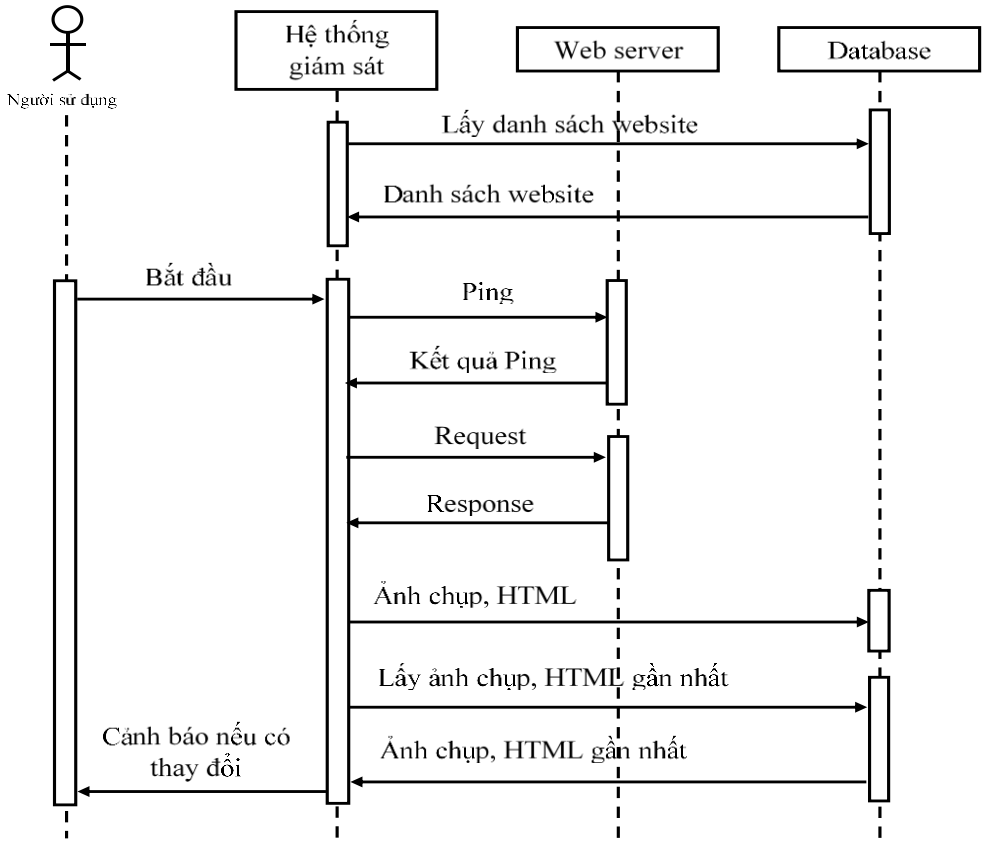


Hình 2.1 Mô hình kiến trúc Portal

2.2. Mô hình tổng quát cho phương pháp kiểm tra phát hiện thay đổi nội dung bảng kết quả của trang tin xổ số.



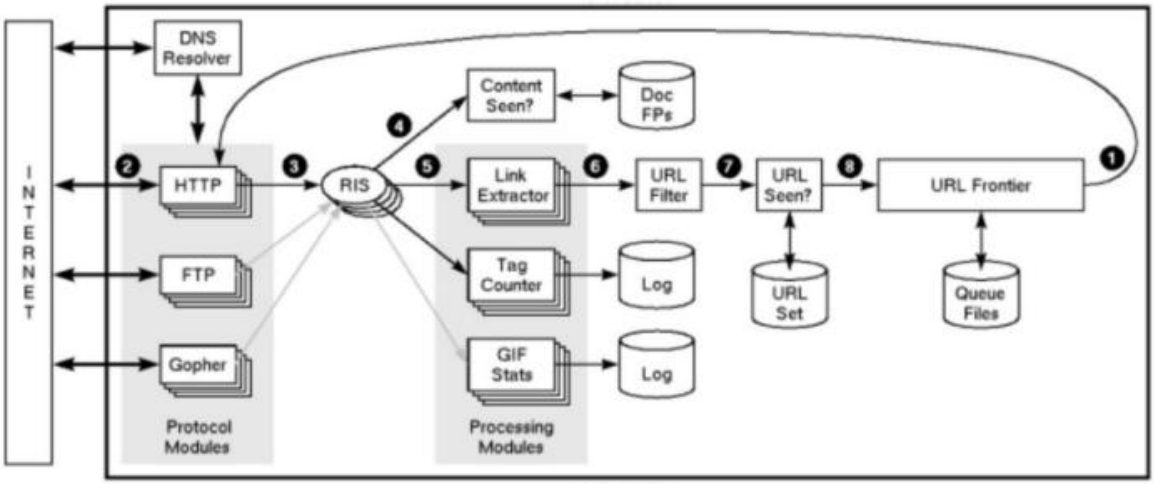
Hình 2.2 Biểu đồ trình tự kiểm tra trang TTĐT



Hình 2.3 Biểu đồ trình tự so sánh nội dung

2.3. Phân tích, đánh giá một số công cụ thu thập thông tin. Chọn một công cụ thu thập thông tin.

2.3.1. Hệ thống thu thập dữ liệu Mercator



Hình 2.4 Các thành phần chính của Mercator.

2.3.2. Hệ thống thu thập dữ liệu từ Twitter- TwitterEcho

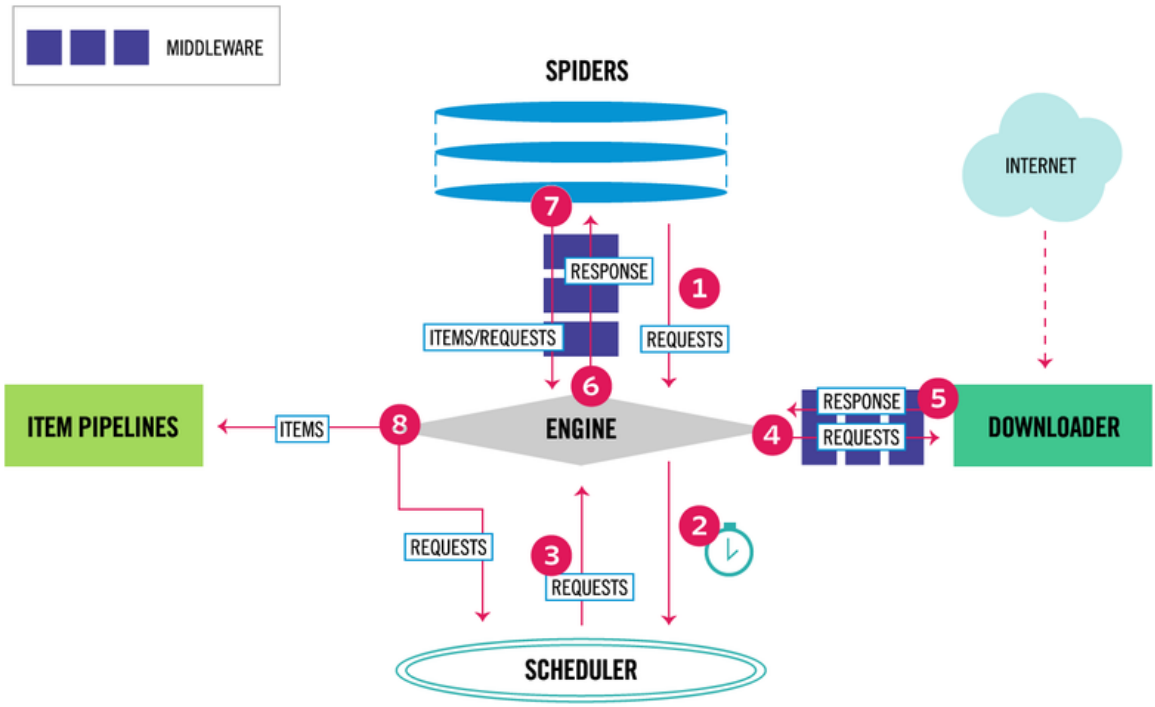
Các dịch vụ truyền thông đa phương tiện xã hội (social media) đã nổi lên trong vài thập kỷ gần đây, thay đổi cách mà chúng ta thông tin với nhau. Do đó những nhà nghiên cứu cần xây dựng hệ thống cho việc thu thập dữ liệu đó hoặc là sử dụng các API được cung cấp bởi mạng xã hội, hoặc là thu thập dữ liệu thông qua Web Crawler.

Đặc biệt mạng xã hội Twitter chứa đựng nguồn thông tin cho việc nghiên cứu, từ việc phân tích tương tác của người sử dụng, phân tích việc sử dụng hashtag, và trích dẫn URL, phân tích nội dung cụ thể nào đó (Ví dụ: phân tích sự lan truyền của dịch cúm, điều tra số lượng người nước ngoài nói tiếng Pháp).

2.3.3. Công cụ HTTrack

HTTrack là công cụ miễn phí cho phép download WWW từ Internet tới thư mục nằm trên máy tính. HTTrack sắp xếp cấu trúc liên kết của site gốc.

2.3.4. Công cụ Scrapy:



Hình 2.5 Các thành phần của công cụ Scrapy

2.4. So sánh thay đổi nội dung mã nguồn web

Việc so sánh thay đổi nội dung mã nguồn, có thể so sánh toàn bộ mã nguồn hoặc chỉ so sánh một phần nội dung (ví dụ: những nội dung xuất hiện trên giao diện, bỏ qua các thẻ...). Hai phần này đều có chung quy trình, chỉ khác so sánh một phần nội dung cần có thêm bước tiền xử lý.

Sau khi có phần văn bản cần so sánh sử dụng thuật toán Rabin Fingerprint cải tiến để lấy giá trị băm của văn bản để so sánh chúng với nhau, nếu giá trị băm khác nhau thì hai văn bản khác và đã có sự thay đổi.

2.5. Chuyển đổi Trang web thành hình ảnh

Trong C# có hỗ trợ công cụ giúp chuyển đổi Trang web thành hình ảnh

2.6. So sánh thay đổi nội dung hình ảnh trang web

Sau khi đã có hình ảnh trang web, sử dung Thuật toán tìm sự khác nhau của hai hình ảnh đã cải tiến để tìm sự khác nhau giữa hai ảnh, giá tri trả về là một ảnh được bôi đỏ những chỗ thay đổi

2.7. Quản lý thời gian thực

Sử dụng công cụ Timer trong C# để liên tục kiểm tra sự thay đổi

2.8. Lưu dữ liệu

2.9. Kết luận chương

Như đã giới thiệu ở Chương 1, những cuộc tấn công thay đổi trang TTĐT được thực hiện để xâm phạm tính toàn vẹn của nó bằng nhiều hình thức.

Có nhiều biện pháp để giữ cho trang TTĐT được an toàn hơn, nhưng không có biện pháp nào hoàn toàn tối ưu, bởi vì các cuộc tấn công như vậy không thể được ngăn chặn ở các lớp (layer) mạng cao hơn, do đó những cơ chế an ninh tốt hơn cần được cung cấp.

Chương 2 đã đề xuất nghiên cứu phương pháp kiểm tra phát hiện thay đổi nội dung trang tin xổ số nhằm phát hiện kịp thời các cuộc tấn công (như đã nêu) bằng phương pháp đa kiểm tra dựa trên nhiều thuật toán nhằm phát hiện thay đổi để đảm bảo tính toàn vẹn của trang TTĐT.

CHƯƠNG 3. CÀI ĐẶT VÀ THỬ NGHIỆM

Chương này nhằm hiện thực hóa các kết quả đã nghiên cứu, sẽ tiến hành triển khai thử nghiệm thu thập nội dung thông tin, ghi thông tin, kiểm tra phát hiện thay đổi nội dung trang tin kết quả xổ số.

3.1. Cài đặt công cụ thu thập thông tin.

Vì scrapy là một công cụ tạo web spider cực mạnh. Rất nhiều dự án và ứng dụng sử dụng scrapy, ví dụ như lấy toàn bộ hình ảnh của một website, các bài viết theo danh mục và theo chủ đề, tạo bot lấy dữ liệu người dùng như số điện thoại và email trên facebook.. hoặc đơn giản hơn là lấy kết quả sổ xố kiến thiết ... Nên học viên đã lựa chọn công cụ này để thu thập nội dung trang TTĐT trong khuôn khổ luận văn này.

Để chuẩn bị cho scrapy chúng ta cần cài đặt những package sau

Cài đặt

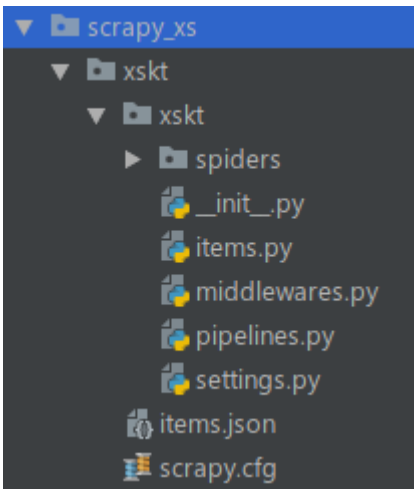
```
pip install scrapy
```

Bắt đầu project

Trong luận văn này sẽ sử dụng scrapy để lấy dữ liệu kết quả sổ xố từ trang TTĐT http://xskt.com.vn, nếu muốn tạo trang web mở đại lý có thể sử dụng source này. Trên virtual environment command line, chạy dòng sau:

```
1. scrapy startproject xskt
```

Scrapy sẽ tạo một folder và các file như sau:



3.2. Phương pháp thu thập thông tin từ trang TTĐT về kết quả xổ số.

3.3. Xây dựng một kịch bản thử nghiệm.

Kịch bản: Chạy chương trình 1 giờ, 5 giờ, 1 ngày với thời gian kiểm tra là 10 phút/lần, vùng kiểm tra là vùng chứa các nội dung chính, có lưu dữ liệu mã nguồn, ảnh chụp trang TTĐT với trang TTĐT: http://xskt.com.vn

3.4. Kết quả thử nghiệm thu thập nội dung thông tin, ghi thông tin, kiểm tra phát hiện thay đổi nội dung trang tin kết quả xổ số.

Lần 1, chạy 1 giờ từ 19h00 đến 20h05 ngày 15/2/2020.

STT	Website	Số lần KT	Số lần phát hiện thay đổi	Tình trạng bất thường
1	xskt.com.vn	7	0	Không

Bảng 3-1. Kết quả thử nghiệm lần 1

Lần 2, chạy 5 giờ từ 14h00 đến 19h05 ngày 16/02/2018

STT	Website	Số lần KT	Số lần phát hiện thay đổi	Tình trạng bất thường
1	xskt.com.vn	31	1	Không

Bảng 3-2. Kết quả thử nghiệm lần 2

Lần 3, chạy 1 ngày từ 19h00 ngày 17/02/2020 đến 19h05 ngày 18/02/2020

STT	Website	Số lần KT	Số lần phát hiện thay đổi	Tình trạng bất thường
1	xskt.com.vn	145	1	Không

Bảng 3-3. Kết quả thử nghiệm lần 3

3.5. Phân tích, đánh giá kết quả thử nghiệm.

Hệ thống chạy ổn định, không bị lỗi, cảnh báo chính xác bằng âm thanh khi phát hiện thay đổi, mức độ chiếm bộ nhớ RAM ổn định, không tăng khi hệ thống chạy lâu dài, dung lượng lưu trữ dữ liệu kiểm tra trang TTĐT trên ổ cứng trung bình 200KB/lần kiểm tra (gồm ảnh chụp, dữ liệu lưu trong database). Nếu tiến độ kiểm tra 10 phút/lần thì 1 ngày 1 trang TTĐT lưu dữ liệu tốn 30MB dung lượng.

3.6. Kết luận chương

Sau khi hoàn thành demo đã đạt được kết quả như sau:

- Phát hiện được tất cả các thay đổi xảy ra của website
- Gửi cảnh báo về email cho quản trị viên mỗi khi có sự thay đổi.
- Giao diện ứng dụng khá thuận tiện
- Dễ dàng cho quản trị viên kiểm tra và phát hiện vị trí cần khắc phục khi có sự cố.
- Tốc độ chương trình tương đối ổn định .

KẾT LUẬN

❖ Các kết quả đạt được:

Nghiên cứu về các giải thuật chính được sử dụng để phát hiện sự thay đổi về nội dung của website, giúp tăng cường khả năng giám sát, phát hiện và cảnh báo, nhằm hỗ trợ cho người quản trị có thể phản ứng nhanh hơn trong các trường hợp trang TTĐT của mình bị tấn công.

Nắm rõ các nguy cơ mất ATTT đối với các trang TTĐT, đặc biệt là thay đổi nội dung. Từ đó nghiên cứu các phương pháp thu thập thông tin, các phương pháp kiểm tra tính toàn vẹn của thông tin để phân tích, thử nghiệm, kiểm tra phát hiện thay đổi nội dung trang TTĐT về kết quả xổ số.

❖ Hướng phát triển:

Tìm hiểu thêm về các tấn công hiện đại, có nguy cơ gây tổn thương trang TTĐT, và tìm cách khắc phục nhằm đảm bảo tính an toàn của trang TTĐT.

Tiếp tục nghiên cứu và phân tích bộ công cụ Scrapy và những công cụ thu thập thông tin khác nhằm phát hiện các điểm yếu khác để khắc phục