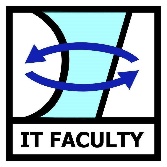
**ĐẠI HỌC ĐÀ NẴNG**

**ĐẠI HỌC BÁCH KHOA**

**KHOA CÔNG NGHỆ THÔNG TIN**

Tel. (84-511) 736 949, Fax. (84-511) 842 771

Website: [itf.ud.edu.vn](about:blank), E-mail: [cntt@edu.ud.vn](mailto:cntt@edu.ud.vn)

****

**BÁO CÁO CUỐI KỲ**

**KHOA HỌC DỮ LIỆU**

ĐỀ TÀI: THU THẬP DỮ LIỆU VÀ DỰ ĐOÁN GIÁ BÁN LAPTOP

**GIÁO VIÊN HƯỚNG DẪN:**

**ThS. Phạm Công Thắng**

**SINH VIÊN THỰC HIỆN:**

**Nhân đẹp trai**

**Huỳnh Trung Hiếu 102180204**

**Lê Viết Trung (Leader) 102180232**

**LỚP: 18TCLC – DT3**

**Đà Nẵng, Tháng 5 - 2021**

**MỤC LỤC**

[ĐỀ TÀI: THU THẬP DỮ LIỆU VÀ DỰ ĐOÁN GIÁ BÁN LAPTOP 1](#_Toc71305040)

[LỜI MỞ ĐẦU 4](#_Toc71305041)

[1 CƠ SỞ LÝ THUYẾT 5](#_Toc71305042)

[1.1 CƠ SỞ LÝ THUYẾT 5](#_Toc71305043)

[1.1.1 Hệ điều hành Linux 5](#_Toc71305044)

[2 TRIỂN KHAI VÀ ỨNG DỤNG 8](#_Toc71305045)

[2.1 THU THẬP DỮ LIỆU 8](#_Toc71305046)

[2.1.1 Tường lửa (Firewall) là gì? 8](#_Toc71305047)

[2.1.2 Tường lửa trên Linux Ubuntu 9](#_Toc71305048)

[2.1.3 Giao thức mạng máy tính 11](#_Toc71305049)

[2.1.4 Tấn công mạng máy tính 15](#_Toc71305050)

[2.2 TRÍCH XUẤT ĐẶC TRƯNG 20](#_Toc71305051)

[2.2.1 Sản phẩm đạt được 20](#_Toc71305052)

[2.2.2 Đánh giá kết quả 21](#_Toc71305053)

[2.3 MÔ HÌNH HÓA DỮ LIỆU 22](#_Toc71305054)

[3 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 23](#_Toc71305055)

[4 TÀI LIỆU THAM KHẢO 24](#_Toc71305056)

[5 PHỤ LỤC 25](#_Toc71305057)

DANH SÁCH HÌNH

[Figure 1 Cấu tạo của một HDH hoàn chỉnh 5](#_Toc71306219)

[Figure 2 Linus Tovards 6](#_Toc71306220)

[Figure 3 Biểu tượng tự sáng tạo mới của người dùng cho Linux 6](#_Toc71306221)

[Figure 4 Hình sau liệt kê lịch sử thời gian phát hành các phiên bản nhân Linux 7](#_Toc71306222)

[Figure 12 Minh họa tường lửa 9](#_Toc71306223)

[Figure 13 Kích hoạt tường lửa Linux 10](#_Toc71306224)

[Figure 14 Tắt tường lửa trên Linux 10](#_Toc71306225)

[Figure 15 Thêm quy tắc mới 11](#_Toc71306226)

[Figure 16 Từ chối một số cổng 11](#_Toc71306227)

[Figure 17 Giao diện của GUFW 11](#_Toc71306228)

[Figure 18 Tạo quy tắc trong GUFW 12](#_Toc71306229)

[Figure 19 Xóa một rule bất kì 12](#_Toc71306230)

[Figure 20 Các loại giao thức mạng máy tính 13](#_Toc71306231)

[Figure 21 Giao thức TCP 14](#_Toc71306232)

[Figure 22 GIao thức UDP 15](#_Toc71306233)

[Figure 23 Ví dụ về mã độc WannaCry 16](#_Toc71306234)

[Figure 24 Worm Conflicker (năm 2008) 17](#_Toc71306235)

[Figure 25 Ví dụ về một email lừa đảo 18](#_Toc71306236)

[Figure 26 Ví dụ minh họa về Man-in-the-middle Attack 18](#_Toc71306237)

[Figure 27 Minh họa về DoS và DdoS 19](#_Toc71306238)

[Figure 28 Mô phỏng SQLi 20](#_Toc71306239)

[Figure 41 Tình trạng Firewall đang tắt 21](#_Toc71306240)

[Figure 42 DoS server không có rule 21](#_Toc71306241)

[Figure 43 Anti DoS sau khi có rule 22](#_Toc71306242)

[Figure 44 Truy cập vào Server bị từ chối đối với ID bị DROP 22](#_Toc71306243)

**DANH SÁCH BẢNG**

[Table 2 Bảng so sánh TCP và UDP 15](#_Toc71306244)

**PHÂN CHIA NHIỆM VỤ**

|  |  |
| --- | --- |
| **Tên** | **Nhiệm vụ** |
| Trương Nhân | Quan sát, tìm kiếm và thu thập dữ liệu |
| Trung Hiếu | Tiền xử lý dữ liệu và trích xuất đặc trưng dữ liệu |
| Viết Trung | Mô hình hóa dữ liệu |

# LỜI MỞ ĐẦU

Chúng ta đang sống trong một thời đại mới, thời đại phát triển rực rỡ của Công Nghệ Thông Tin (CNTT). CNTT đã ở một bước phát triển cao trong xu thế hiện đại hóa thế giới khoa học và công nghệ, không thể kể đến chính là khả năng lập trình nhằm tăng khả năng xử lý các bài toán thực tiễn và trong nghiên cứu khoa học.

Những tiến bộ trong máy tính đã khiến con người ngày càng tính toán một cách nhanh chóng và linh hoạt hơn. Nếu máy tính là một công cụ thì lập trình ngôn ngữ chính là cách mà chúng ta sử dụng công cụ này hiệu quả nhất. Thông qua đó, chúng ta dần đạt được hiệu suất công việc cao hơn và lượng giá trị dữ liệu ngày càng lớn hơn. Nhờ có internet, lượng dữ liệu đó ngày càng nhiều và quý giá hơn. Chính vì lượng dữ liệu thông tin quý giá đó mà khoa học dữ liệu ra đời.

Khoa học dữ liệu giúp chúng ta hiểu và phân tích được các dữ liệu có sẵn nhằm phục vụ các mục đích khác nhau như thương mại, an ninh quốc phòng hay các mục đích cao cấp hơn như dự đoán, phân loại sử dụng trí tuệ nhân tạo. Đây chính xác là lĩnh vực quan trọng đang ngày càng nhiều công ty và doanh nghiệp quan tâm. Trong đó lĩnh vực thương mại đang vô cùng cần thiết để dự đoán giá trị của một sản phẩm nào đó. Qua đó, chúng em quyết định sẽ thu thập dữ liệu và dự đoán giá tiền của các sản phẩm laptop đang có và hiện hành trên thị trường.

Qua thời gian học tập và nghiên cứu, nhờ sự hỗ trợ và hướng dẫn tận tình của thầy Phạm Công Thắng, chúng em đã tiếp thu được rất nhiều kiến thức có ích phục vụ cho bài tập cuối kì này. Vì vậy, qua đây chúng em cũng xin gửi lời cảm ơn chân thành đến thầy Thắng của khoa Công Nghệ Thông Tin trường Đại Học Bách Khoa Đà Nẵng đã giúp chúng em được thực hiện bài tập lớn này.

Chúng em xin chân thành cảm ơn!

Lớp 18TCLC-DT3

# CƠ SỞ LÝ THUYẾT

## Thu thập dữ liệu

## Trích xuất và mô hình hóa dữ liệu

### Mô hình hóa dữ liệu

#### Khái niệm

Trong thực tế, sẽ có nhiều khái niệm về mô hình hóa dữ liệu. Trong công nghệ phần mềm, mô hình hóa dữ liệu (Data modeling) là quy trình tạo ra một mô hình dữ liệu cho một hệ thống thông tin bằng cách áp dụng một số kỹ thuật chính thức nhất định.

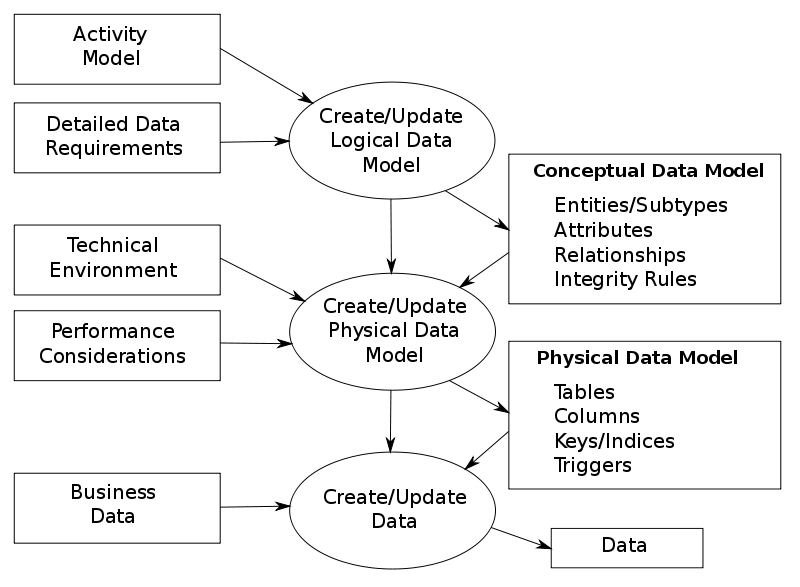


Figure 1 Quy trình mô hình hóa dữ liệu

Mô hình dữ liệu bao gồm các đối tượng dữ liệu và giá trị dữ liệu. Các đối tượng mục (item) và thể loại (category) là các đối tượng cốt lõi trong mô hình dữ liệu, được xác định bởi đối tượng đặc tả (spec). Tập hợp các đối tượng mục là một danh mục (catalog). Đối tượng phân cấp (hierarchy) xác định một hình thức thứ bậc của tập hợp các danh mục.

Mô hình hóa dữ liệu là quá trình tạo ra một mô hình dữ liệu. Khi tạo một mô hình dữ liệu, trước tiên phải xác định dữ liệu, các thuộc tính và mối quan hệ của nó với dữ liệu khác và xác định các ràng buộc hoặc giới hạn đối với dữ liệu. Ví dụ: có thể tạo mô hình dữ liệu cho sản phẩm trong đó thuộc tính nhà cung cấp của mặt hàng sản phẩm liên kết với id nhà cung cấp trong danh mục nhà cung cấp.

#### Lợi ích của mô hình hóa dữ liệu

* Cho phép nhân viên phát triển ứng dụng và cơ sở dữ liệu để xác định các yêu cầu hệ thống và giao tiếp với nhau và với người dùng cuối.
* Cung cấp một bức tranh rõ ràng về các ràng buộc toàn vẹn tham chiếu. Duy trì tính toàn vẹn tham chiếu là điều cần thiết trong mô hình quan hệ nơi các mối quan hệ được mã hóa hoàn toàn.
* Cho phép tạo một sơ đồ tóm tắt kết quả của các nỗ lực mô hình hóa dữ liệu và tạo một lược đồ cơ sở dữ liệu từ mô hình đó.

### Trích xuất đặc trưng dữ liệu

#### Khái niệm

Trích xuất đặc trưng (feature selection) là một quá trình chọn lọc một tập con chứa các thuộc tính liên quan để sử dụng trong quá trình xây dựng mô hình.

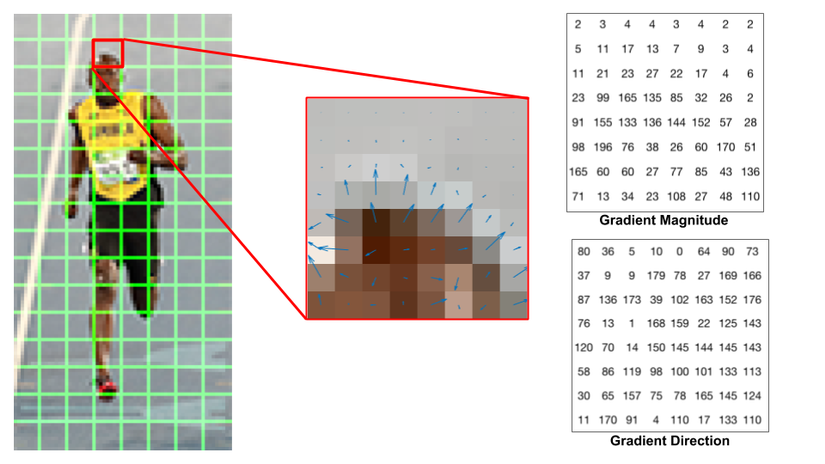


Figure 2 Trích xuất đặc trưng cho ảnh

Các kỹ thuật trích chọn đặc trưng được dùng cho một số lý do:

* Đơn giản hóa các mô hình để giúp các nhà nghiên cứu/người dùng diễn dịch dễ dàng hơn
* Giảm thời gian huấn luyện
* Tránh tình trạng dữ liệu nhiều chiều
* Tăng cường tổng quát hóa bằng cách giảm sự quá khớp (chính thức, giảm phương sai)

Trích xuất đặc trưng dữ liệu bao gồm hai phần là xây dựng các dữ liệu và lựa chọn các dữ liệu đặc trưng.

### Thư viện sử dụng

#### Numpy as np

Numpy (Numeric Python): là một thư viện toán học rât phổ biến và mạnh mẽ của Python. NumPy được trang bị các hàm số đã được tối ưu, cho phép làm việc hiệu quả với ma trận và mảng, đặc biệt là dữ liệu ma trận và mảng lớn với tốc độ xử lý nhanh hơn nhiều lần khi chỉ sử dụng Python đơn thuần.

NumPy thường được sử dụng cùng với các gói như SciPy (Python Scientific) và Mat-plotlib (thư viện vẽ đồ thị). Sự kết hợp này được sử dụng rộng rãi để thay thế cho MatLab, một nền tảng phổ biến cho tính toán kỹ thuật.

Các thao tác sử dụng:

* Các biến đổi Fourier và các quy trình để thao tác shape
* Thao tác xử lý ma trận

#### Seaborn as sns

Seaborn là một trong những thư viện Python được đánh giá cao nhất thế giới được xây dựng nhằm mục đích tạo ra các hình ảnh trực quan đẹp mắt. Nó có thể được coi là một phần mở rộng của một thư viện khác có tên là Matplotlib vì nó được xây dựng trên đó.

Các thao tác sử dụng:

* Trực quan hóa dữ liệu đơn biến và đa biến
* Vẽ biểu đồ thống kê phân tích

#### Pandas as pd

Pandas là một thư viện Python cung cấp các cấu trúc dữ liệu nhanh, mạnh mẽ, linh hoạt và mang hàm ý. Tên thư viện được bắt nguồn từ panel data (bảng dữ liệu). Pandas được thiết kế để làm việc dễ dàng và trực quan với dữ liệu có cấu trúc (dạng bảng, đa chiều, có tiềm năng không đồng nhất) và dữ liệu chuỗi thời gian.

Các thao tác sử dụng:

* Xử lý các dữ liệu bị mất mát
* Xử lý các dữ liệu lỗi, phân tách dữ liệu, kiểm tra shape ma trận

# TRIỂN KHAI VÀ ỨNG DỤNG

## THU THẬP DỮ LIỆU

## TRÍCH XUẤT ĐẶC TRƯNG

### Sản phẩm đạt được

Sau khi chúng em có được sản phẩm chúng em bắt đầu tiến hành kiểm tra lưu lượng mạng và đánh giá khả năng của sản phẩm.

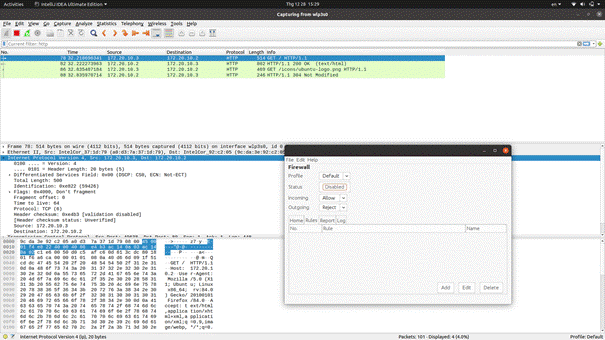


Figure 20 Tình trạng Firewall đang tắt

Sau khi mở Firewall và chúng em tiến hành DoS server. Hiện tại, trong cấu hình Profile này chưa có bất kì rule nào nên khi DoS đã xảy ra tắt nghẽn mạng, gói SYN/ACK được server gửi lại liên tục làm máy chủ mất nhiều tài nguyên.

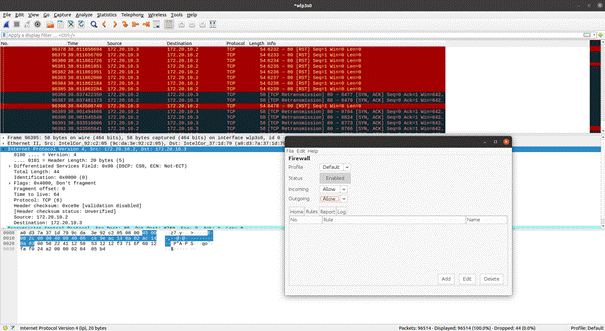


Figure 21 DoS server không có rule

Sau khi nhận ra có một ID đang tiến hành DoS Server. Chúng em đã thiết lập rule đối với ID đó để DROP. Sau đó sử dụng Wireshark để kiểm tra kết quả.

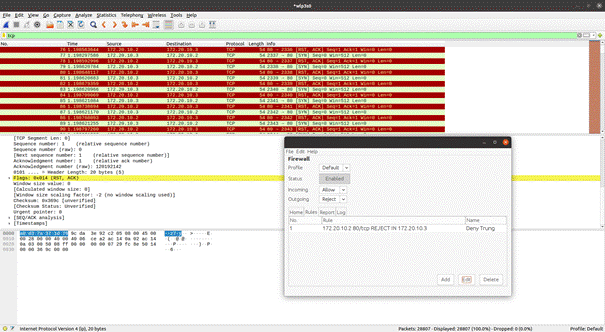


Figure 22 Anti DoS sau khi có rule

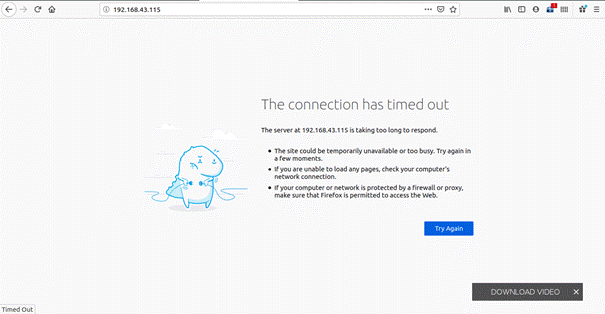


Figure 23 Truy cập vào Server bị từ chối đối với ID bị DROP

### Đánh giá kết quả

* **Ưu điểm:**
* Sản phẩm đã hoàn thiện được yêu cầu của đồ án và hoàn thành được 80% mục tiêu ban đầu đề ra của chúng em.
* Nhờ quan sát các lưu lượng mạng bằng Wireshark chúng em đã phát hiện được các máy đang DoS và DROP thành công ip mở đầu cho phát triển tính năng tự động chống DoS sau này
* **Nhược điểm:**
* Trình giao diện quản lý còn thô sơ vẫn còn hơi khó sử dụng.
* Đôi lúc edit rule và delete rule còn xảy ra lỗi, khi edit hay delete xong phải tắt/bật tường lửa lại để load lại list rule.

## MÔ HÌNH HÓA DỮ LIỆU

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

* **KẾT LUẬN**

Sau rất nhiều nỗ lực và thời gian, chúng em cũng đã hoàn thành được ứng dụng và chạy được một cách nhanh chóng và chính xác. Nhưng khi nhìn lại thực tế, chúng em cũng đã tự học được những bài học và những ưu điểm cũng như nhược điểm riêng của ứng dụng.

**Ưu điểm:**

* Hiểu rõ được bản chất vấn đề cũng như yêu cầu của đề tài
* Đáp ứng được hầu hết các yêu cầu của đề tài như xây dựng hoàn thiện ứng dụng tường lửa trên Linux.
* Áp dụng kiến thức của nguyên lý hệ điều hành và mạng máy tính để ứng dụng vào trong đồ án.

**Nhược điểm:**

* Khả năng lập trình còn chưa cao dẫn đến thuật toán còn chưa tối ưu nên ứng dụng còn mới ở mức cơ bản, chưa áp ứng được nhu cầu sử dụng thực tế cho server.
* Thời gian xử lý các dòng lệnh còn chậm.
* **HƯỚNG PHÁT TRIỂN**

Sau các kết quả thu được từ quá trình nghiên cứu và làm đồ án cơ sở này, chúng em đã có được các định hướng phát triển thêm như sau:

* Hoàn thiện và cải tiến về thuật toán sao cho đạt tốc độ xử lý nhanh chóng và hiệu quả hơn đối với các ma trận cấp cao hơn nữa.
* Thiết kế thành một ứng dụng cấp có độ tương tác có thể sử dụng được ở mức server.
* Thiết kế các chức năng tự động chống DoS và DDoS.
* Thiết kế giao diện thân thiện dễ sử dụng hơn nữa.
* Lập trình lại để thích ứng được với mọi hệ điều hành.

# TÀI LIỆU THAM KHẢO

**[1]** Đề tài “Tìm hiểu về tường lửa”, <https://www.slideshare.net/dungdtk9a/titmhiuvtnglapdf>, 30/9/2020

**[2]** TỔNG QUÁT VÀ PHÂN LOẠI GIẢI PHÁP TƯỜNG LỬA, <http://saigonlab.vn/blogs/news/tong-quat-va-phan-loai-giai-phap-tuong-lua> , 30/9/2020

**[3]** Toàn bộ kiến thức về Tấn Công Mạng (Cyber-attack), <https://resources.cystack.net/tan-cong-mang-cyber-attack/> , 30/9/2020

**[4]** Linux là gì, <https://wiki.matbao.net/linux-la-gi-tong-hop-moi-kien-thuc-ve-he-dieu-hanh-linux/> , 30/9/2020

**[5]** Tường lửa Linux, <https://code.tutsplus.com/vi/tutorials/the-linux-firewall--net-31748> , 30/9/2020

**[6]** Linux, hệ điều hành, <https://freetuts.net/linux-la-gi-he-dieu-hanh-linux-423.html> , 24/10/2020

**[7]** Tường Lửa [FireWall] Là Gì - Tính Năng Của FireWall, <https://vinahost.vn/tuong-lua-la-gi.html> , 25/10/2020

**[8]** Cách thiết lập tường lửa trên Linux, <https://quantrimang.com/cau-hinh-tuong-lua-trong-linux-bang-firestarter-55647>, 15/11/2020

**[9]** Basic Netword Summary, 26/12/2020

**[10]** Sự khác nhau giữa TCP và UDP,<https://www.softwaretestinghelp.com/tcp-vs-udp/>, 20/6/2020

**[11]** Tấn công từ chối dịch vụ DoS và DDoS là gì? Tác hại của chúng ra sao, <https://quantrimang.com/tim-hieu-ve-tan-cong-tu-choi-dich-vu-dos-34926>, 26/12/2020

# PHỤ LỤC