WEB PHISHING DETECTOR

Phạm Đại Hoàng An, Trần Hoàng Công Toại, Nguyễn Trọng Anh, Đỗng Hoàng Sơn

Giới thiệu đề tài

Hiện nay, có rất nhiều website lừa đảo trên không gian mạng. Các website này đã và đang gây thiệt hại rất lớn cho người dùng và các doanh nghiệp. Trong năm 2018, theo nhóm Công tác Chống lừa đảo (Anti-Phishing Working Group – APWG), có khoảng 51,401 trang web lừa đảo. Một báo cáo khác của RSA chỉ ra rằng các tổ chức và hiệp hội trên thế giới chịu thiệt hại lên đến 9 tỷ đô bởi các cuộc tấn công lừa đảo trong năm 2016

Chính bởi các hành vị lừa đảo ngày càng tinh vi và khó nhận biết, và cũng để bảo

vệ cho nhiều người dùng chưa có được những kiến thức cần thiết, là lý do ra đời cho đề tài này.

Mô tả dữ liệu

Các website giả mạo có các tính năng gần có thể rất giống với các trang web thật đối với mắt người, nhưng chúng khác nhau về IP. Môt URL cơ bản bao gồm :

- **Domain name:**đây là phần bắt buộc vì nó phải được đăng ký với tên miền
- **Subdomain name và Path:** hoàn toàn có thể bị những kẻ lừa đảo kiểm soát và đánh lừa người dùng

Từ dữ liệu của các website bao gồm cả website lừa đảo và hợp pháp, mục tiêu của dữ liệu này là xét **trạng thái** để làm tập training cũng như làm tập kiểm tra để xem độ chính xác của thuật toán:

THUỘC TÍNH	THÔNG SỐ
Số lượng các đặc điểm của website	25
Số lượng website được khảo sát	11430
Số lượng website giả mạo	5715
Tỷ lệ phần trăm website giả mạo	50%
Số lượng website hợp pháp	5715
Tỷ lệ phần trăm website hợp pháp	50%

Bảng 1. Mô tả dataset trong nghiên cứu
https://drive.google.com/file/d/1vBNy4ysnQ06t5F8TTnlbeZaiWDpLGyrp/view

Bằng việc sử dụng kỹ thuật filter để lấy các thông tin cần thiết từ dữ liệu tổng như length url, prefix suffix,... và trực quan hóa lên biểu đồ thông qua python, từ đó đưa ra được đặc điểm cụ thể của từng feature trong legit website và phishing website.

Từ các biểu đồ đã được trực quan trong mỗi trường hợp, ta đưa ra được các kết luận về đặc điểm của một phishing website. Từ đó tạo nên nguồn dữ liệu để train và test trong quá trình train model sau này. Nếu một website có nhiều yếu tố của một website lừa đảo thì ứng dụng sẽ thông báo đây là website lừa đảo.

Phương pháp phát hiện

Để có thể phân loại trang Web giả mạo, nhóm đã phân ra 2 giai đoạn thực hiện gồm: **Training mô hình** và **Phát hiện trang Web giả mạo**.

Training mô hình: Dữ liệu đưa vào sẽ được xử lý thông qua các đặc tính được liệt kê, qua đó ta trực quan dữ liệu và phân ra thành 2 tập test và training. Với tập dữ liệu dùng để training, ta đưa vào một mô hình phân loại cụ thể để huấn luyện mô hình đó, sau đó dùng tập test để kiểm tra độ chính xác.

Phát hiện trang Web giả mạo: Khi người sử dụng đưa vào một địa chỉ URL, các đặc tính từ thanh địa chỉ đó sẽ được mang ra so sánh với tập training để từ đó dự đoán trang Web đó là giả mạo hay không

Các thuật toán sử dụng

Sử dụng các thuật toán học có giám sát điển hình cho bài toán phân loại.
Support Vector Machine: là thuật toán khá mạnh và phổ biến trong bài toán phân lớp, thuật toán tìm siêu phẳng (hyperplane) tốt nhất phân chia hai lớp.
Logistic Regression: là thuật toán đơn giản phát triển từ Linear Regression, thuật toán này phân loại khá tốt cho các tập dữ liệu linear separable.
Decision Tree: là thuật toán phân loại dựa trên cây phân cấp, có thể mô tả bằng các luật nên mô hình dễ hiểu và dễ giải thích.
Random Forest: là thuật toán dựa trên Decision Tree, nhưng thay vì dùng một cây để phân loại thì Random Forest sử dụng nhiều cây và kết quả dựa trên bỏ phiếu. Các cây được xây dựng dựa trên số lượng mẫu và số thuộc tính ngẫu nhiên.

Mỗi thuật toán đều có ưu nhược điểm riêng, có thể lựa chọn thuật toán phù hợp nhất dựa trên việc đánh giá mô hình đã huấn luyện thông qua các thông số Precision, Recall và Accuracy.

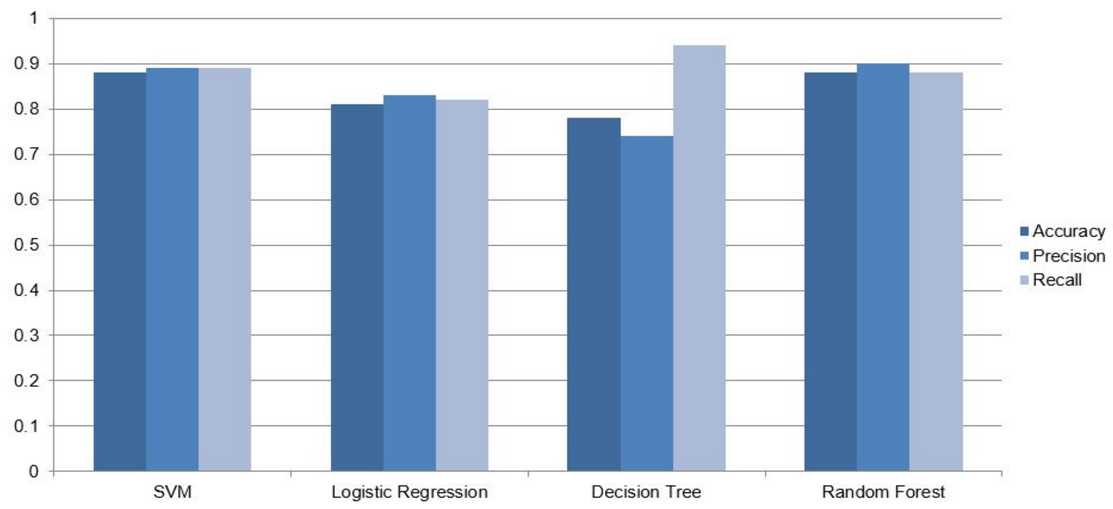
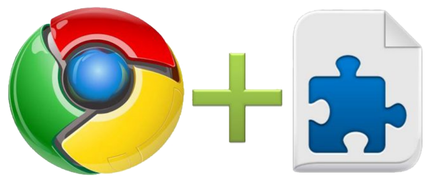


Chart 1. Model Evaluation

Hướng phát triển tương lai

Nghiên cứu này là cơ sở để phát triển mô hình dự đoán lừa đảo được tích hợp vào tiện ích của trình duyệt web (hay còn gọi là extension). Việc đưa vào extension của trình duyệt web giúp cho người dùng có thể dễ dàng cài đặt và sử dụng, đem lại tiện ích và hiệu quả tối đa. Tuy nhiên để tạo extension cũng gặp rất nhiều khó khăn như phái có sự cấp phép của bên trình duyệt hay sự tin tưởng từ phía người dùng. Do đó trong tương lai dự án này cần phải cải thiện nhiều hơn về tính bảo mật, tiện lợi, quảng cáo đến người dùng...

Ngoài phát triển thành extension, ta có thể mô hình hóa nghiên cứu này thành phần mềm được cài đặt trên các thiết bị. Điều này sẽ giúp người dùng có thể dự đoán lừa đảo ở mọi trình duyệt web thay vì chỉ xài được ở một trình duyệt như extension. Tuy vậy thì hướng phát triển này vẫn cần được cải thiện về tính an toàn để có thể được cài đặt trên thiết bị của người dùng mà không xảy ra xung đột với hệ điều hành.



Hình 1 : TÍch hợp extension trong Google Chorme

Kết luận

Thông qua việc kiểm nghiệm các phương pháp khác nhau, đề tài đã chọn thuật toán SVM. Trong thời gian tới, đề tài sẽ ngày càng được hoàn thiện tốt hơn, cho được kết quả phát hiện ngày càng chính xác hơn.

Vẫn còn những nhược điểm: chưa có extension cụ thể (giới hạn thời gian và chính sách bảo mật của google).

Contact

Phạm Đại Hoàng An
Big Data Club _ Group 3
Email: an.phambk19@hcmut.edu.vn
Website: https://anhoangbk19.github.io/BDC-Assignment/
Phone: 0833270501

References

1. Dalia Shihab Ahmed, Assist. Prof. Dr. Karim Q. Hussein, Hanan Abed Alwally Abed Allah. (2022). Turkish Journal of Computer and Mathematics Education. Vol. 13 No. 01. 100 - 107.
2. Jian Mao, Jingdong Bian, Wenqian Tian, Shishi Zhu, Tao Wei, Aili Liand Zhenkai Liang. (2019). Phishing page detection via learning classifiers from page layout feature. Truy cập từ: https://iwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-019-1361-0
3. Abdelhakim Hannousse, Salima Yahiouche. (2021). Web page phishing detection. Truy cập từ: https://data.mendeley.com/datasets/c2gw7fy2j4/3
4. Rami M. Mohammad, Fadi Thabtah, Lee McCluskey. (2022). Phishing Websites Features.
5. Huynh Chi Trung. (2020). Giới thiệu về Support Vector Machine (SVM). Truy cập từ: https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB