Malicious URL Detection

Department of Computer Science and Engineering

Malaviya National Institute of Technology, Jaipur

Supervisor: Dr. Smita Naval

Sumit (2017UCP1323)
Himanshu Rawat (2017UCP1230)
Rajat Gedam (2017UCP1254)

May 13, 2021

Aim of the Work

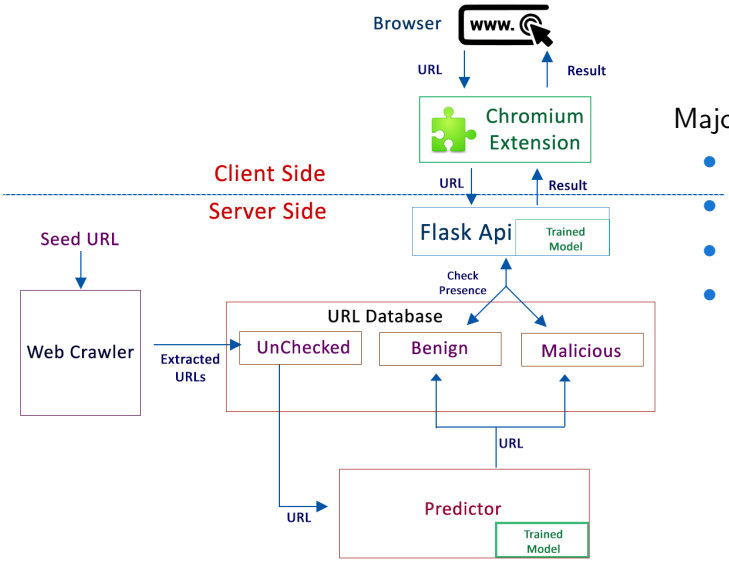
- To study and design the Machine Learning Model for predicting the URL behaviour as Malicious or Benign.
- To design a web crawler for initial list of URLs.
- To create a chrome extension as an application of the above work.

Literature Survey and Related Work

Author	Feature	Work
Kolari et al. [1]	Lexical Feature	Extracting words delimited by special characters - Bag of Words Model
Blum et al. [2]	Lexical Feature	Bi-gram Features
McGrath and Gupta [3]	Host-based Feature	Time to live from Registration of Domain
Justin et al. [4] , Ma et al. [5]	Host-based Feature	WHOIS information, location, domain-name properties
Hou et al. [6]	Content-based Feature	Length of document, word count, invisible object, links to remote script
Choi et al. [7], Canali et al. [8]	Content-based Feature	Number of iframe tags, Count of suspicious elements
Canali et al. [8]	Content-based Feature	Statistical Properties - Presence of double documents, Number of elements with small area

Table: Literature Survey and Related Work

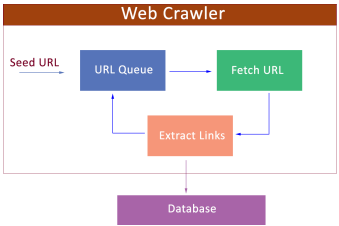
Architecture



Major Components

- Web Crawler
- URL Database
- Chrome Extension
- Classifier

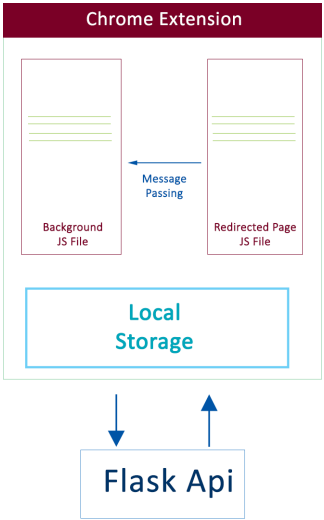
Major Component : Web Crawler



- Web Crawler is an internet bot that systematically browses the world wide web.
- Take seed URL as input, crawl webpages and collect links.
- store collected links in database for model to predict.

Major Component : Chrome Extension

- The interaction of the browser with our server-side components happens through the chrome extension.
- Warns the user in case of malicious URL and gives a chance to whitelist it.
- Communicates with the database and ML Model through Flask API.



Feature Representation : Lexical Features

Obtained from the properties of the URL name.

- Malicious sites tend to "look different" in the eyes of the users who see them.
- It can be used to infer patterns in malicious URLs that we would otherwise miss through ad-hoc approach.
- Safer and Faster since it does not involve execution

https://www.bfuduuioo1fp.mobi/ws/file.dll

Features	Value
URL Length	41
Count of Special Characters	3
Count of Sub-Directories	2
Count of Query Parameters	0
Count of '-'	0
Count of '.'	3
Digit to Letter ratio	0.03
Count of Sub-Domains	1

Table: Lexical Feature Representation

Feature Representation : Host-based features

- They are obtained from the host-name properties of the URL
- They allow us to know the following about the malicious host -
  - Location
  - Identity
  - Management style and properties
- Key Observations :
  - Phishers exploited Short URL services
  - The time-to-live from the registration of the domain was almost immediate for the malicious URLs
  - Many used botnets to host themselves on multiple machines across several countries
- Features Used
  - We have used "duration" as a host-based feature for our classifier

Feature Representation : Content Based feature

- Content based feature are extracted using content(i.e. HTML, JS, CSS code) of webpage.
- HTML features are based both on statistical information about the raw content of a page (e.g., the page length or the percentage of whitespaces) and on structural information derived from parsing the HTML code (e.g., the location of specific elements in the page).
- Some of the feature suggested in Canali et al. [8] and Choi et al. [7] are:-
  - Number of hidden elements.
  - Number of included URLs.
  - Presence of double documents.
  - Presence of meta refresh tag.

Learning Algorithm

After converting URLs into feature vectors, many of the learning algorithms can be generally applied to train a predictive model

- Batch Learning - SVM, Random Forest, Naive Bayes
- Online Learning
- Representation Learning - Deep Learning

Online Learning

- It represents a family of efficient and scalable learning algorithm that learn from data sequentially
- Much more scalable than batch learning algorithm
- Learning and forecasting are computationally efficient
- Why Online Learning ?
  - Most of the malicious websites are not active - So host-based and content-based features are not available
  - Online Learning algorithm will continuously accept stream of data over a period of time
  - In this way we will encounter active websites
  - Active websites will provides us their host-based and content-based features
  - Its accuracy will increase with time - Because all types of features can be employed for use
  - We may use "river" library for online learning

Dataset Creation

- We have used dataset from kaggle.com, But the ratio of malicious to benign was very small.
- Used phishtank.org for malicious URLs.
- we have used 134,000 URLs(with nearly equal amount of malicious and benign URLs).
- Divided URLs in ratio of 70:30 for training and test set.

Experimental Results

1. Lexical Features
2. Host-based Features
3. Content-based Features

Features	Accuracy	Precision	Recall
(1)	95%	96%	95%
(2)	80%	57.14%	74.22%
(3)	75%	70.54%	97.84%
(1) and (2)	82.85%	92%	76.67%
(1) and (3)	86.19%	86.53%	88.53%
(2) and (3)	77.14%	79.41%	75%
(1), (2) and (3)	90.47%	85.57%	90.81%

Table: Observations on combinations of Features

Conclusion

- Lexical features are more readily available as compared to host-based and content-based features
- In long run online learning may give better results
- For immediate results (as in our chrome extension), lexical features are more reliable
- True positives were very high and false positives were very low for model using only lexical features

Limitations of the work and Future Scope

- Ratio of malicious to benign URLs is very low.
- Unavailability of host-based and content-based features.
- Honeypots may be used in order to increase the accuracy and reliability of the system.

References

[1] P. Kolari, T. W. Finin, and A. Joshi, "Svms for the blogosphere: Blog identification and splog detection," 2006.

[2] A. Blum, B. Wardman, T. Solorio, and G. Warner, "Lexical feature based phishing url detection using online learning," 2010.

[3] D. K. McGrath and M. Gupta, "Behind phishing: An examination of phisher modi operandi," 2008.

[4] M. Justin, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious urls," 2009.

[5] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying suspicious urls: An application of large-scale online learning," 2009.

[6] Y.-T. Hou, Y. Chang, T. Chen, C.-S. Laih, and C.-M. Chen, "Malicious web content detection by machine learning," 2010.

[7] H. Choi, B. B. Zhu, and H. Lee, "Detecting malicious web links and identifying their attack types," 2011.

[8] D. Canali, M. Cova, G. Vigna, and C. Kruegel, "Prophiler: A fast filter for the large-scale detection of malicious web pages," 2011.