Abstract

In this work, we present and evaluate a machine learning framework that takes as input a domain name (based on the respective DNS request) and outputs the content category it belongs to. We evaluate several options for feature engineering and classification to find the most optimal setup for the specific problem domain. We also address the problem of data collection and preprocessing. We propose a SERP (Search Engine Response Pages)-mining approach to collect and label an appropriate dataset. Our experimental evaluation uncovers several interesting insights and forms the basis for further work into this interesting domain. The problem we addressed is summarized in the High-level architecture diagram.

System Architecture

The system architecture of the overall framework, containing the DNS Classification module, data preprocessing, feature engineering and classification steps.

Motivation and Contribution

- There exists several categories of web pages that belong to “borderline” categories (e.g., websites selling illegal substances or weapons) and might be of interest for any public or private organization to monitor as outgoing traffic.
- We built a machine learning framework for classifying DNS requests into topic categories, including data collection, preprocessing, and classification through various configurations.

SERP Dataset

- A total of 112 categories to be classified, with 11,278 instances
- Of those categories, 82 fall under “general” content and 30 fall under “borderline” categories to be monitored.

Conclusions

- Considering the multiple configurations used, Random forest, logistic regression, and SVM were the best performing classifiers and LDA performed less than expected, reinforcing the saying that “simpler is better” in machine learning applications.
- We also observed that the borderline instance classification does not follow the same patterns as the regular ones, with the title of a URL being a more weak indicator of the class label than its description.