Определение ППО веб-приложения с помощью автоматического выделения характерных признаков

Самосадный Кирилл
ЛБИС ВМК МГУ

Рускрипто 2016

Введение

- Тестирование методом черного ящика:
  - Сбор информации о тестируемом веб-приложении.
- Получение информации об архитектуре ПО веб-приложения:
  - операционная система;
  - веб-сервер;
  - языки програмирования;
  - фреймворки, CMS и т.п.
- Проверка эксплуатируемости известных уязвимостей для компонент архитектуры веб-приложения.

Определение ОС и веб-сервера

- nmap:
  - сетевой (IP)
  - транспортный (TCP, UDP)
- httprint:
  - прикладной (HTTP)

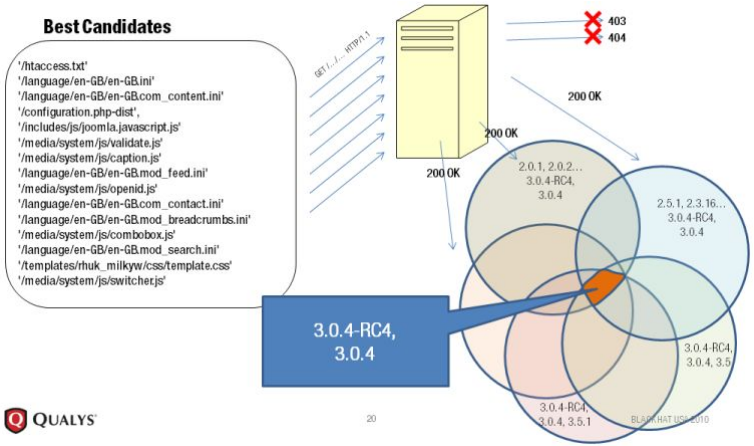
HTTP Test	What to expect
`HEAD / HTTP/1.0`	Normal HTTP header response
`DELETE / HTTP/1.0`	Response when operations such as DELETE aren't generally allowed
`GET / HTTP/3.0`	Response to a request with an improper HTTP protocol number
`GET / JUNK/1.0`	Response to a request with an improper protocol specification

Модель TCP/IP

Существующие подходы к определению ППО

- Набор характерных признаков на основе экспертного знания.
- Анализ на уровне HTTP-запросов и HTTP-ответов:
  - наличие или отсутствие некоторого ресурса;
  - анализ HTTP-ответа как текста.
- Средства:
  - whatweb;
  - Blind Elephant.

```
# Matches #
matches [

# HTML Comment
{ :text=>'<!--START: 3dcart stats-->' },

{ :text=>'<!--END: 3dcart stats-->' },

{ :search=>"headers[set-cookie]", :regexp=>/3dvisit/ },

]


end
```



Недостатки существующих подходов

- Анализ только статических признаков, присутствующих в исходном коде HTML-страницы, и невозможность проанализировать процесс отображения этой страницы.
- Отправка характерного набора несвязанных HTTP-запросов, что отличается от поведения обычного клиента.
- Проблема новизны при использовании экспертных знаний.
- Неустойчивость относительно методов защиты от ботов.

Предлагаемое решение

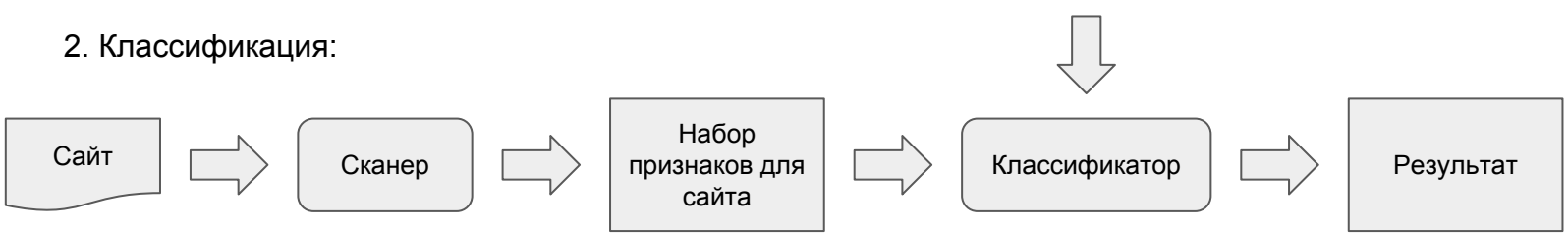
- Базовое средство - безоконный (headless) браузер.
- Отображение корня веб-сайта в безоконном браузере, что позволяет анализировать процесс отображения HTML-страницы.
- Список характерных признаков для ППО формируется на основе обработки набора веб-сайтов, использующих данное ППО, и выделения общих признаков для них.
- Для определения использования ППО некоторого класса на данном веб-сайте его характерные признаки проверяются на совпадение с характерными признаками этого класса.

Преимущества предлагаемого подхода

- Эмулируется заход обычного пользователя на главную страницу веб-сайта:
  - не подозрительно;
  - устойчиво к защите от ботов.
- Анализируется не только статический HTML, но и процесс его отображения:
  - больше возможных характеристик;
  - полнота данных.
- Решается проблема новизны.

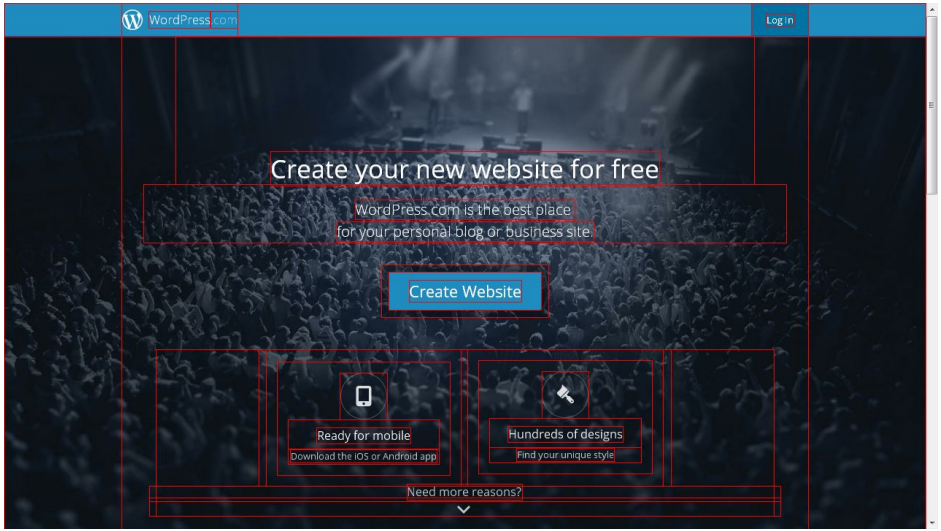
1. Обучение:

2. Классификация:



Реализация

- QtWebKit + Remote Debugger Protocol
- python (PyQt) + JavaScript
- Характеристики:
  - HTTP-заголовки, Cookies;
  - метаинформация;
  - css-классы;
  - URL;
  - хэш от js, css и img;
  - интерфейсы.



http://a.com/wp-content/uploads/2015/07/budgetBadge.js
http://b.com/wp-content/uploads/2014/11/Candara_400.font_.js

Загружается JavaScript-код из wp-content/uploads/

Тесты

- WordPress + Joomla.
- Обучающая выборка - примеры с официальных сайтов.
- Тестовая выбор - Alexa Top 500.
- Вспомогательное средство - whatweb:
  - 2% обучающей выборки - false negative
  - 2% тестовой выборки - WordPress
- Результаты относительно whatweb:
  - 3000 веб-сайтов
  - 9% false positive:
    - 251 - всего
    - 29 - настоящие false positive
  - 30% false negative
    - 19 - всего

Итоги

- Рассмотрены ограничения существующих методов определения ППО.
- Предложен новый метод определения ППО.
- Метод реализован и протестирован на некоторых классах ППО.