Chappie Swarm: Persona-Driven Web Corpus Generation

Nicholas Kaufman
August 8, 2016
Outline

I. Introduction of the Tool
II. A Look Under the Hood
III. Use Cases
IV. Future Work
Chappie Introduction
Chappie Swarm: Scope of Application

■ Why build it?

■ What is it?
  • Tool for generating accurate and current web corpora.
  • Formulated to be systematic, reproducible, and offer the ability to quickly disseminate created datasets.
  • Intended to replicate enterprise or business type browsing.
  • Simulates users not the internet.

■ What is it not?
  • Intrusion Detection or Prevention System.
  • Repository of web surfing data.
  • Traffic Simulator.
    – Designed only to interface with the web.
Why should you use the application?
  • Rapidly generate web traffic data for use in Machine Learning systems.

Why focus on human emulation?
  • Increased validity of data in Machine Learning systems.

Why not just pay graduate students or interns to do this?
  • Combat observer effect – people will behave differently when they know they’re being watched.

What about existing web corpora?
  • Addresses question of currency – designed to be a better reflection of the web as it exists today.
A Look Under the Hood
Core technology behind Chappie Swarm is Docker, Pipework, Wireshark, and PhantomJS, Docker-NAT-Router.
A Chappie in Particular

■ Chappies browse the web in accordance with a Markov Model.
  • The “state” that a Chappie is in corresponds to its current interest.

■ Each Chappie is seeded at run time with a ”major” interest.
  • These might be things such as “sports” or “movies.”
  • While browsing, a Chappie may decide to change its interest with a certain probability, or to continue browsing according to its current interest.

■ Chappies interact with websites in multiple ways:
  • Search for terms associated with an interest.
  • Select a website from a “favorites” list.
  • Click links on a webpage.
  • Utilize the browser’s “back button.”
Chappie: A Visual Description

chappie_web_surfer topic transition process

Random transition between “thinking” states based on probabilities

- Select major topic during launch
- Enter state machine
- 70% stay major
- 75% break to major

Major topic

- 20% major to minor
- 55% minor to major

Any minor topic

- 35% stay minor
- 25% break to minor

Take a break

- 10% go to break
Chappie: A Visual Description

chappie_web_surfer process

Random transition between “thinking” states, with bias towards one specific topic (i.e., “news”)

“news” → “news” → “sports” → next topic?

Google

CNN

Google

CBS Sports

Default user homepage

1

“Depth first” (0-max depth) selection of links represents clicks

2

Random use of back button incorporated on each page

3

Customizable probability for time spent on each page and break time between sessions

1...N more sites

One script per container
State Transitions Explained

- Chappie always remembers its major topic.
  - Major topic is the topic selected during initial seeding.

- Before deciding on a new webpage, the Chappie asks itself: “Am I still interested in this topic?”
  - Random number chosen via uniform probability distribution. State change depends upon the value selected.
Browsing Features

Treading the line between automated browsing and human experience, we attempt to control for a few features that will look “more human.”

- Time between successive page calls.
  - Default behavior is a randomly chosen number in a given time window.
    - Utilizes uniform distribution.
- Reported User-Agent String.
  - Selected from a pool of up-to-date and out-of-date web browsers.
- Status codes of visited pages.
- Number of distinct webpages requested.
- “Types” of web pages visited.
  - Is there a cohesive “story” that can be put together given the web logs?
Use Cases
Population Estimation

- From an outside perspective, enterprise-level web traffic will be obfuscated.
  - Firewalls, Proxies, NATting, etc.

- Hard to tell how many users exist behind a particular obfuscation layer – much less several of them.

- Dual vantage of capture allows a user to create both "labelled" and "unlabeled" data sets for a machine learning approach.
Geographical Profiling

- Geolocation affects the ecological validity of web surfing samples.
  - Geofencing
  - CDNs
  - Transit points

- The same website might behave differently for people who are in geographically disparate areas.

- Chappie swarm might be used in conjunction with a service, such as AWS, to study differences in traffic patterns to better understand regional discrepancies.
Threat Detection

- There are classes of attacks which use web pages as means of infecting individual computers, or corporate networks.
  - Watering Hole
  - Drive-By

- Chappie Swarm Application can be used to generate data which can be considered both ‘normal/benign’ and malicious.

- This would allow researcher to be able to cheaply label network traces or proxy logs for use with a Multiple Instance Learning Algorithm.
Future Work
The basic model of how a Chappie behaves is simple. Better emulation of human behavior is a priority.

Better automation of topic material.
- More dynamic favorites list, home page, etc.
- Consideration of ideas such as language barriers and belief systems in influencing websites that a Chappie might visit.

Expand Application to utilize more of the Network Stack.
- Chappies which send emails would be useful for experiments in detecting threats such as (spear) phishing in addition to web based exploits.
Questions?