Collecting “Typical” Domain Names for Web Servers

Paul Hoffman

MAPRG meeting at IETF 110

March 2021

Motivation

- When researchers measure the properties of the authoritative DNS servers on the Internet, they first need to define the types of authoritative servers they are sampling
- The current collections of domain names against which one can do research are not that good for assessing things about “typical” domain names
  - Most popular web sites
  - Extracts from the zone files of gTLDs
  - Dumps from passive DNS collection systems

External links on Wikipedia pages

- Wikipedia has wikis in almost every language
- External links go to a large variety of real but not popular web pages
  - governments of small cities
  - colleges and universities of all sizes
  - obscure sports teams
  - small regional music and movie studios
  - personal sites of academics
- Worldwide coverage

Collection and analysis

- Retrieve the database of external links for each language Wikipedia from a mirror of the main Wikipedia site
- Extract all the external links
- Clean up the list of external links, limit to “http:” and “https:”
- For each remaining URL, strip off the scheme and everything after the domain name
- Cull the list of domain names so that only one copy of each domain name remains

The dataset

- 750 databases were from 2020-01-01
- After culling, 7.35 million unique domain names in the dataset
- Use a random sample of 100,000 from that dataset for testing
- Needed to start with more than 100,000 because many names could not be resolved to an IPv4 address
- Analyzed for things like how many had IPv6 addresses, how many were DNSSEC signed, and so on

Full report

- Collecting “Typical” Domain Names for Web Servers
- OCTO-023
- https://www.icann.org/octo-023-en.pdf