A First Look at Web Browsing Predictions using DNS Logs

Utkarsh Goel, Clint Cooper, Brittany T. Fasy, and Mike P. Wittie
SEDE 2016
Motivation

- Users want **enjoyable experience** with the Web.
- Expectation that **webpages should load within 1-2 seconds**.
- Developers **use Web proxies** to deliver Web content in a **timely manner**.
- But **users still complain** about poor Web performance.
Motivation

- Web requests are processed only after clients request content.
- Generating response may take significant time, when content is not in cache.
- But, web servers don’t know what clients will want in the future and thus cannot proactively process future requests.
- The problem is: what to pre-cache?
“What-if” We Could Pre-Cache

- **Reduce latency** to speed up page loads.

- Push **webpage structure** to the web browser, **before it is requested**.

- Improve **advertisements** based on users’ next browsing interests.
Data Collection

- Use DNS logs instead of HTTP
  - HTTP contains user-sensitive information
- Collect DNS logs from ITC at MSU.
  - Timestamps
  - User IP addresses
  - Website domain names
- About 12 million DNS entries.
  - 1.2 M during morning hours
  - 6.7 M during school hours
  - 4.3 M during evening hours
Data Significance

![Graph showing #DNS Requests (thousands) over Time of Day (24 Hour Format)]
Website Category Popularity

The chart shows the popularity of different website categories during various hours:
- **OSN**: High popularity during Morning Hours.
- **Video**: Moderate popularity during Work Hours.
- **News**: Low popularity across all hours.
- **E-Commerce**: Moderate popularity during Night Hours.
- **Adult** and **Education**: Low popularity across all hours.
Length of Web Browsing Sessions

![Graph showing the cumulative distribution function (CDF) of web browsing sessions over time, categorized by morning hours, work hours, and night hours. The x-axis represents web sequence length in minutes, ranging from 0 to 20, and the y-axis represents the CDF of web browsing sessions, ranging from 0 to 1. The graph compares the distribution of session lengths across different time periods.](image-url)
Web Browsing Patterns

- Google ←→ Facebook
- Google ←→ Yahoo!
- Facebook ←→ Apple
- Google ←→ Twitter
- Facebook ←→ Twitter
- Twitter ←→ LinkedIn
- Facebook ←→ FBCDN
Takeaways

- DNS can be used, **in place of user-sensitive HTTP requests**, to identify patterns in Web browsing.

- Pre-caching could
  - Speedup webpage load times
  - Improve advertisements

- **Improve user enjoyment** with the Web.
Thank you

Code available at:

https://github.com/msu-netlab/magic
Background Slides
Data: \((T, S)\) where \(T\) is a time and \(S\) is tuple size.

Result: \(H = (K, V)\) where \(H\) is a HashMap with keys \(K\) as a sequence (seq) of domains and values \(V\) as the number of times that sequence occurs.

\[
data = DataAtTime(T)
\]

\[
IPLList = [x[0] \text{ for } x \text{ in } data]
\]

for \(ip \in IPLList:\)

\[
Domains = DomainsWithIP(ip)
\]

for \(i \in \text{range}(0, \text{len}(Domains) - S)\):

\[
seq = (domain_i, domain_{i+1}, ..., domain_S)
\]

if \(seq \in H:\)

\[
H[seq] = H[seq] + 1
\]

else:

\[
H[seq] = 1
\]