Investigating Distribution of Data of HTTP Traffic: An Empirical Study

IEEE Innovations in Information Technology 2006


Presenter: A. Ziad Hatahet
Outline

- Motivation
- Previous Work
- Methodology
- Architecture
- Results
- Recommendations
- Future Directions
Motivation

- HTTP is the most common type of traffic on the Internet today
- Generating the proper model/profile will help in:
  - Conducting proper simulation
  - Conducting proper mathematical modeling
  - Conducting better statistical analysis
  - Better design of networking algorithms and equipments
- Improve the performance of HTTP traffic
Previous Work

- Several attempts have been made
- Several approaches
  - Goal: varying goals—not all intended to generate a model/template for HTTP data
  - Datasets: mostly relying on traces of requests to web servers collected from clients logs in certain institutions
  - Retrieved data: belongs to stored files
  - Only one is based on an empirical study
- Not all studies concur on the same model
- Inapplicability: All studies were conducted between 1995 and 1999. The web is very different today (doubling of Internet traffic every year since 1997 and a wide range of new applications)
  - Goal: generate a model/template of a “typical” website
  - Data sets should be based on the most popular accesses not traces from certain institutions
  - Empirical studies produce results closer to reality
  - Retrieved data today belongs to both dynamically-generated and stored files
Methodology

- Luxury of current environment
  - Highly empowered programming languages with rich network-programming features
  - Highly available and economical over-the-counter high computational power
  - High bandwidth connectivity
- Conducting a complex experiment on a large dataset and collecting empirical data on the Internet is achievable today in an acceptable time
Methodology

- Identify a database of the most popularly accessed websites
- Build the tools to automatically
  - Crawl the websites
  - Recursively parse the underlying webpages
- Obtain
  - Sizes
  - Types
  - Other metadata statistics
- Analyze obtained data
Architecture

WWW Space

Website DB

Crawler

MetaGetter

HTML Parser

Collected Data

Tools

Analysis

Distribution Model

σ₁
Architecture

- The database of websites
  - 500 from the Global Top 500 list of Alexa Inc.
  - 33,974 of the Stanford WebBase project
  - 180 duplicate entries  ➔ 34,294 total entries

- Tools
  - Development tool: Sun’s Java 2 Standard Edition 5.0
  - Data Analysis: MATLAB

- Computational Environment: experiment divided among 20 machines running simultaneously each with a 3.2 GHz Pentium 4
Architecture

- Two datasets
  - Home
  - Home + Level 1

- The split was conducted
  - In order to observe differences of data distribution between Home, and Home + Level 1
  - If no significant differences found, future analysis can be conducted on the smaller Home dataset
## Results

### Summary of Data Collection Phase

<table>
<thead>
<tr>
<th></th>
<th>Home</th>
<th>Home + L1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of web pages</td>
<td>34,294</td>
<td>114,838</td>
</tr>
<tr>
<td>Retrieved objects</td>
<td>499,714</td>
<td>1,895,776</td>
</tr>
<tr>
<td>Avg. num. of objects per page</td>
<td>14.6</td>
<td>16.5</td>
</tr>
<tr>
<td>Execution time (hours)</td>
<td>163</td>
<td>1428</td>
</tr>
</tbody>
</table>
Results: Size Distribution

Most occurring object at 43 B

Home

Home + L1
Results: Size Distribution

Size distribution for Home

[Graphs showing size distribution with 50%, 25%, and 25% divisions]
Results: Size Distribution

Size distribution for Home + L1

Graphs showing size distribution for Home + L1 with 50%, 25%, and 25% distributions.
Results: Type Distribution

Type distribution based on occurrence

Home

Home + L1
Results: Type Distribution

Type distribution based on traffic size

Home

- Image: 56.66%
- Text: 29.70%
- Application: 13.07%
- Video: 0.01%
- Others: 0.12%

Home + L1

- Image: 37.32%
- Text: 47.39%
- Application: 13.92%
- Audio: 0.81%
- Video: 0.20%
- Others: 0.35%
### Results: Object Size

Average object size based on type (in KB)

<table>
<thead>
<tr>
<th>Type</th>
<th>Home</th>
<th>Home + L1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
<td>14.7</td>
<td>16.4</td>
</tr>
<tr>
<td>Audio</td>
<td>230.9</td>
<td>512.5</td>
</tr>
<tr>
<td>Image</td>
<td>3.9</td>
<td>3.6</td>
</tr>
<tr>
<td>Text</td>
<td>15.6</td>
<td>15.9</td>
</tr>
<tr>
<td>Video</td>
<td>21.4</td>
<td>654.8</td>
</tr>
<tr>
<td>Other</td>
<td>3.3</td>
<td>6.2</td>
</tr>
</tbody>
</table>
Results: Common Objects

- Most common 3 objects

<table>
<thead>
<tr>
<th>Size (Bytes)</th>
<th>Content type/subtype</th>
<th>Probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>image/gif</td>
<td>3.8</td>
</tr>
<tr>
<td>4,040</td>
<td>text/html</td>
<td>1.2</td>
</tr>
<tr>
<td>49</td>
<td>image/gif</td>
<td>0.7</td>
</tr>
</tbody>
</table>

- 5.7% of traffic from web servers to clients is due to spacer images or non-existing files!
Results: Web-Server Distribution

Web-server software distribution

- Apache: 49.6%
- Microsoft-IIS: 30.2%
- Unknown: 7.9%
- Others: 5.0%
- Zeus: 0.3%
- Lotus Domino: 0.4%
- Sun ONE: 0.7%
- Rapidsite: 0.7%
- IBM HTTP Server: 0.8%
- Netscape Enterprise: 4.5%
Recommendations

- Introduce pixel-based or vertical separator in HTML standard
- Make the web-editing software manufacturer aware of the extra traffic being caused by the “nudge” operation in order to influence a different method for the same procedure
- Automatic offline utilities to remove stale links (to non existing files)
Future Directions

- Produce a probability distribution function that best fits the distribution of HTTP traffic
- Analyze the results in terms of actual TCP packets generated rather than just flat data
Thank You
Previous Work (Detailed)

  - NSFNET traffic: six datasets from web servers
    - Three from academic institutions
    - Two from research institutes
    - One from a commercial Internet service provider
  - Goal: Studying workload characterization of web servers to improve performance
  - Inapplicability
    - Datasets do not apply well to today’s environment
    - The goal is not to generate a model
  - Datasets: based on traces of requests to web servers from clients logs
  - Goal: Investigating trends in traces and generates distributions for the sizes of the retrieved files
  - Data retrieved belongs to stored files
  - Inapplicability: data retrieved is only from stored files
  - Goal: building a behavioral model based on collected web traces.
  - Inapplicability: goal is producing a long-lasting behavioral model not generating a current data template model
  - Empirical study—examines actual data retrieved rather than relying on traces in logs;
  - Datasets are not based on popularly accessed websites and are for existing stored files
  - Inapplicability:
    - Datasets do not apply well to today’s environment
    - Data retrieved belongs to stored files