Efficient Summarization of URLs
Using CRC32 for Implementing URL Switching

Zornitza Genova
Kenneth J. Christensen
Department of Computer Science and Engineering
University of South Florida
Tampa, Florida 33620
{zgenova, christen}@csee.usf.edu

This material is based upon work funded by the National Science Foundation under grant no. 9875177

Topics

- Introduction
- Problems
- Digesting of URL lists
- Evaluation
- Summary and future work
**Introduction continued**

- Content is replicated throughout the Internet

![Diagram]

- Same file found in three places
- Request is sent to the origin server
- Should it be redirected? How? Where?

**Introduction continued**

- URL routers must share URL lists to build routing tables

![Diagram]

- Sharing of URL lists
- Origin server
- Caches
- Temporary server
**Problems**

- URL lists...
  - Contain URLs of all distributed files
  - Are very large (~ 5 to 120 Mbytes for a single Web site)
  - each URL router keeps several lists in memory!

  **Problem #1**: Sharing consumes network bandwidth and prohibits frequent updates  
  **Problem #2**: Variable length URLs make routing table look-ups slow

**Digesting of URL lists**

- Reduce size of URL list by digesting
  - In the literature...
    - MD5 plus Bloom filter [2]

  - Our idea...
    - Use CRC32 for each URL in list
    - Goals: 1) Less CPU resources than MD5-Bloom  
      2) Faster look-up than MD5-Bloom
    - Issue: Does not solve the problem of false hits (collisions) due to non-unique CRC32
Evaluation of digesting

- **Experiment:** CPU time and digest size
  - Input: URL lists from HTTP traces
  - Examine: CPU time to generate digest; size; (hashing) collisions
  - Compare: different compression methods

<table>
<thead>
<tr>
<th>Method</th>
<th>CPU (sec)</th>
<th>Size (MB)</th>
<th>Coll. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD5-Bloom (8)</td>
<td>89.13</td>
<td>9.74</td>
<td>0.03</td>
</tr>
<tr>
<td>CRC32</td>
<td>16.22</td>
<td>9.74</td>
<td>0.03</td>
</tr>
<tr>
<td>LZ compression</td>
<td>89.13</td>
<td>9.74</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Evaluation of digesting continued

- **Experiment:** Collisions as a function of URL length
  - Input: URL lists from HTTP traces
  - Examine: Effect of URL string length on collisions
  - Measures: URL string length, collisions
Summary and future work

- URL routers are needed to automatically distribute load
  - Between content sources “Internet wide”

- URL routing tables are large
  - We use CRC32 signatures to digest URL lists

- Our method is better
  - Less CPU resource than MD5-Bloom
  - Same or better collision rate than MD5-Bloom

- Future work
  - Performance evaluation of routing table look-up mechanisms

References


