Managing Duplicates in a Web Archive

Daniel Gomes
André L. Santos
Mário J. Silva
University of Lisbon
Tomba Web Archive (tomba.tumba.pt)
Tomba architecture

- Periodical crawls of the Portuguese web
- Repository:
  - Distributed and extensible storage system
  - Support accesses while new contents are being loaded
Duplication within the archive

<table>
<thead>
<tr>
<th>Crawl</th>
<th>Date</th>
<th>#URLs</th>
<th>% duplicates within the crawl (horizontal)</th>
<th>% duplicates from the last crawl (vertical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>07-2002</td>
<td>1.6M</td>
<td>23%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>10-2002</td>
<td>1.2M</td>
<td>21%</td>
<td>7%</td>
</tr>
<tr>
<td>3</td>
<td>03-2003</td>
<td>3.5M</td>
<td>15%</td>
<td>10%</td>
</tr>
<tr>
<td>4</td>
<td>10-2003</td>
<td>3.3M</td>
<td>11%</td>
<td>19%</td>
</tr>
<tr>
<td>5</td>
<td>06-2004</td>
<td>4.4M</td>
<td>7%</td>
<td>18%</td>
</tr>
</tbody>
</table>

- On average 25% of the archived contents were duplicates: waste of storage space
- Disks are cheap but the web is enormous.
- Storage space must be spared
Horizontal duplication

- **Mirrors**
  - `tucows.ip.pt` == `www.tucows.com`

- **URL aliases**

- **Default error messages**
  - “This web page uses frames, but your browser doesn't support them”
Vertical duplication

- Web collections built incrementally
- Exact duplicates
  - The page remains unchanged
- Partial duplicates
  - The page suffers changes with time

URL_A

Crawl 1

URL_A

Crawl 2

Content

URL_A

Crawl 3

Content'

URL_A
Objective: save on storage space by eliminating duplicates

• Can not jeopardize the system’s performance
• Consider a distributed and extensible storage space
• Consider preservation issues
1. Avoid crawling duplicates: estimate frequency of change (Cho,03)

• Advantages
  – Saves bandwidth
  – Reduces duration of the crawl

• Disadvantages
  – Does not eliminate horizontal duplication
  – Requires historical data to create model
  – Assumes URL persistence
URLs are not persistent

\[ y = -0.1373 \ln(x) + 1.0683 \]

\[ R^2 = 0.928 \]

- Analysis of the persistence of URLs among 10 crawls of the Portuguese web gathered for 3 years
- Half of the URLs took less than 2 months to disappear
2. Delta storage mechanisms

- Store only the part that has changed

- Advantages
  - Eliminates partial duplicates
  - Available software (CVS)

- Disadvantages
  - Assumes URL persistence
  - Does not eliminate horizontal duplication
  - Documents are rebuild: preservation problem
3. Distributed file systems

• Advantages
  – Available software: NFS, AFS
  – Extensible, distributed

• Disadvantages
  – General usage, not designed to fulfill the requirements of web archives
    • POSIX interface: support permissions, caching of files
  – Difficult management: requires admin. privileges
  – Do not have a built-in mechanism to eliminate duplicates
    • Store everything.
    • Find and delete duplicates in batch
    • Disk IO is expensive
4. Compression

• Compression algorithms save space by eliminating duplication between contents

• Advantages
  – Many compression algorithms available (zip, rar, zlib)
  – Independent from URLs
    • Eliminates vertical duplicates
    • Eliminates horizontal duplicates

• Disadvantages
  – Duplicates are eliminated only within each compressed file
    • Scalability problem: files get too big
  – Corruption of 1 file jeopardizes several contents
Our solution: Elimination of exact duplicates based on content signatures

• Advantages
  – Independent from URLs
  – Eliminates vertical and horizontal duplication
  – Lightweight algorithm
  – Improve storage throughput
    • Duplicates are not written to disk

• Disadvantages
  – Does not eliminate partial duplicates
Validation: Webstore

1. Generate content signature
2. Find it among the volumes
3. If it is already stored, discard content
4. If it is a new content, store content
Write

- Elimination of dups. must not jeopardize performance
- NFS as baseline
  - Widely available
  - Reproducibility
- Time took to store 1000 pages
- Increasing number of clients storing on 1 volume
- Writes on average 50% faster than NFS
Read and Delete

• Reads on average 68% faster than NFS

• Deletes on average 60% faster than NFS
Conclusions

- There is vertical and horizontal duplication on the web
  - Duplicates are frequent within a web archive

- URLs transience is a problem
  - Estimation of frequency of change
  - Delta storage

- Webstore eliminates vertical and horizontal exact duplicates without jeopardizing performance
  - 57 million web contents in Webstore (1.5 TB).
“Future” Work

• Study URL and Content Persistence
http://webstore.sourceforge.net  (Webstore)
http://tomba.tumba.pt  (web archive)
http://xldb.fc.ul.pt  (research group)
dcg@di.fc.ul.pt

Thank you for your attention.