From Web Page Storage to Living Web Archives

Thomas Risse

JISC, the DPC and the UK Web Archiving Consortium Workshop
Agenda

• Web Crawling today & Open Issues
• LiWA – Living Web Archives Project
• Selected working areas of LiWA
  – Dynamic Pages
  – Handling of Spam
  – Temporal coherence of crawls
  – Archive Interpretability
• Conclusions and Expected Project Results
Current Web Archiving at a Glance

Quality Review

Selection
- Preparation
- Discovery
- Filtering

Capture
- Temporal Coherence of Crawls
- Noise Filtering

Access
- Deep Web
- Long-term Interpretability

Archiving
- Index
- Storage
- Multimedia Content
- Link Extraction
- Fetching

User
LiWA – Living Web Archives (EU-IST 216267)

Next generation Web Archiving technology for:

– High Quality Web Archives
– Long-term Archive usability

➔ From Web page storage to “Living Web Archives”

Started Feb. 2008 (3 Years)
Some LiWA Objectives in more Detail

Quality Review
- Preparation
- Discovery
- Filtering

Access
- Deep Web
- Archiving
  - Index
  - Storage

Capture
- Noise Filtering
- Dynamic Pages
  - JavaScript, Flash
- Link Extraction
  - Fetching

Temporal Coherence of Crawls
- Multimedia Content

Archivist
- User
Some LiWA Objectives in more Detail

- Quality Review
- Archiving
- Access
- Long-term Interpretability
- Temporal Coherence of Crawls
- Noise Filtering
- Dynamic Pages (JavaScript, Flash)
- Multimedia Content

Flowchart:
- Selection
  - Preparation
  - Discovery
  - Filtering
- Capture
  - Link Extraction
  - Fetching
- Archiving
  - Index
  - Storage
- User

Deep Web
Link Extraction of Dynamic Pages

- Quality Review
- Access
- Selection
- Preparation
- Discovery
- Filtering
- Noise Filtering
- Capture
- Link Extraction
- Fetching
- Multimedia Content
- Archiving
- Index
- Storage
- Dynamic Pages (JavaScript, Flash)
- Temporal Coherence of Crawls
- User
- Multimedia
- Deep Web
- Long-term Interpretability

LiWA
Living Web Archives
20/07/2009
Follow the links....

11.7 glob -- Unix style pathname pattern expansion

The glob module finds all the filenames matching a specified pattern according to the rules used by the Unix shell. No tilde expansion is done, but *, ?, and character ranges expressed with [] will be correctly matched. This is done by using the os.path.expand() and os.path.expanduser() functions in concert, and not by actually invoking a subshell. (Portfile and shell variable expansion, use os.path.expanduser() and os.path.expandvars().)

glob(pattern)

Return a possibly-empty list of filenames that match pattern, which must be a string containing a path specification. pattern can be either absolute (like Windows c:\temp\MyFiles) or relative (like ~/Desktop\*, and can contain shell-style wildcards. Broken symlinks are included in the results (as in the shell).

glob(pathname)

Return an iterator which yields the same values as (glob) without actually storing them all simultaneously. New in version 2.5.

For example, consider a directory containing only the following files: 1.gif, 2.png, and card.gif. glob(‘??.gif’) will produce the following results. Notice how any leading components of the path are preserved:

```python
>>> import glob
>>> list(glob.glob(‘/home/John\*’))
[‘/home/John’, ‘/home/John’]
>>> list(glob.glob(‘*.gif’))
[‘card.gif’, ‘1.gif’, ‘2.gif’]
```

See Also:

* Module fnmatch:
  Shell-style filename (not path) expansion.
Hmm...
Link Extraction of Dynamic Pages

• The Problem:
  – links don’t exist as raw text lying around
  – user interaction and code assemble them

• Current Approach:
  – “guessing” by assembling any fragments that look like links into URLs and trying them out
  – Can be very noisy - lots of wrong URL’s
Link Extraction of Dynamic Pages

• Approach
  – “pressing” the links and see what comes out
  – Execute code in a Javascript engine
  – Extract links from resulting DOM tree
  – Implementation based on WebKit
Noise Filtering

- Quality Review
- Access
- Selection
- Capture
- Index
- Storage
- Multimedia Content

Temporal Coherence of Crawls
Link Extraction
Fetching
Multimedia Content

User

Long-term Interpretability

LiWA
Living Web Archives
20/07/2009
Web spam: for (or against) search engines

Compute the out degree
On the Feasibility of Low-rank Approximation for Personalized PageRank

File Format: PDF/Adobe Acrobat - View as HTML transition matrix of the Web graph for computing personalized PageRank. ... out-degree. Hence the base of links ...

http://www.ilab.sztaki.hu/~stamas/publications/benczur05low_rank_ppr.pdf Cached - Similar pages

scho...
Web Spam: indexing vs. archiving

• Primary target: search engines to manipulate ranking
• As side effect, we also archive spam
• But very costly if not fought against:
  – traps crawler
  – 10+% sites
  – near 20% HTML pages
What can we do?

Ideal solution
• Automatic identification of spam pages

Requires
• Right selection of features to identify spam
• Development of new features e.g. creation and disappearance of new sites, pages
• Good training sets

Problem: Spam is constantly changing
→ Features need to be adapted
→ Updated training sets are necessary
   Training set need to be prepared manually
Temporal Coherence of Crawls

- Quality Review
- Access
- Selection
  - Preparation
  - Discovery
  - Filtering
- Capture
  - Link Extraction
  - Fetching
- Archiving
  - Index
  - Storage
- User
- Noise Filtering
- Dynamic Pages: JavaScript, Flash
- Multimedia Content
- LiWA: Living Web Archives

Long-term Interpretability

Deep Web
Temporal Coherence

• Capturing Web sites as “authentic” as possible
• Make a site snapshot at once is not possible
• Crawlers need to be polite to web sites
  – Slow crawling, maybe with delays
  – Pages are changing during site crawl

→ When Do we have a coherent crawl?
Coherence by Example

\[ t_{\text{coherence}} = [t_2, t_3) \]
Coherence by Example

\[ t_{\text{coherence}} \in \emptyset \]

\[ p_1 \quad \bullet \quad \quad \quad \bullet \quad p_2 \]

\[ p_3 \quad \bullet \quad \quad \quad \bullet \quad p_4 \]

\[ t_1 = t_s \quad t_2 \quad t_3 \quad t_4 = t_e \]
Coherence Analysis Technology

Temporal Coherence Report

Overview

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total requests</td>
<td>6674</td>
</tr>
<tr>
<td>Revisited pages</td>
<td>859/96</td>
</tr>
<tr>
<td>Crawl Duration</td>
<td>2h 52m</td>
</tr>
<tr>
<td>Revisit Duration</td>
<td>1h 38m</td>
</tr>
</tbody>
</table>

Changes

<table>
<thead>
<tr>
<th>Changes Details</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages with changed links</td>
<td>16</td>
</tr>
<tr>
<td>Revisited pages</td>
<td>859/96</td>
</tr>
<tr>
<td>Revisit Duration</td>
<td>1h 38m</td>
</tr>
</tbody>
</table>

Details

- Pages with changed links:

Creation of automatically generated reports

Visualization of coherence defects
Motivation

Archives store content over long time ranges
- Content is created latest in the year of archiving
- Content typically creators use the language of that time
Overview Process

Step 1: Word Sense Discrimination

Find the meaning of a term in one collection at one period in time

Step 2: Tracking Evolution

Detect evolution of terms
Data Sets for Evaluation (1/2)

Data Set Requirements
- Large corpus
- Fully digitized
- Long time range – Increase probability of terminology evolution
- Not too domain specific (like the Mesh corpus)
- Homogeneous language
- Time annotated

Using Web Archives
- Large digital corpus
- At most 10 Years old
- Inhomogeneous with all the ”noise” of the web
- Not suitable for initial evaluations
Data Sets for Evaluation (2/2)

- **Newspaper Archives**
  - Fully digitized corpora
  - Controlled language
  - Clear time annotations
- **Süddeutsche Zeitung (ger.)**
  - Spans from year 1994 - 2006
  - ~ 1.3 Million articles
- **London Times Archive (engl.)**
  - Spans from year 1785-1985
  - ~ 20 Million articles

**Strategy**

- Year 2: Initial evaluations on well known corpora
- Year 3: Apply technology to web archives, .gov.uk crawls provided by EA
Term. Evol - Conclusions and Future Work

**Terminology Extraction**
- Extraction
- Stop word removal
- Lemmatization
- Correcting OCR errors

**Word Sense Discrimination**
- Other types of clustering
- Metrics for evaluation

**Detecting Evolution**
- Methods for comparing clusters and detecting evolution
- Methods for evaluation

Nina Tahmasebi
Conclusions and Expected Project Results

Improving Web Archiving Technology
- Rich Media Capturing
- Spam Processing
- Archive Coherence
- More general scope: Improving Archive Interpretability

Selected results will be integrated in
- Heritrix Crawler (our test-bed)
- Hanzo Archives Crawler

Evaluation in two test cases:
- Streaming Media WebArchive by Sound & Vision
- „WebArchivists Workbench“ by European Archive and Nat. Lib. Czech Republic
Thank you!

More information on
http://www.liwa-project.eu/
The LiWA Consortium

Archiving Companies
European Archive
Hanzo Archives

Archiving Users
Stichting Nederlands Instituut voor Beeld en Geluid,
National Library of the Czech Republic,
Moravian Library

Technical & Scientific
Leibniz University
L3S Research Center,
Max-Planck-Institut für Informatik,
Computer and Automation Research Institute
Hungarian Academy of Sciences