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

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

24/07/2009
Some LiWA Objectives in more Detail

- Selection
  - Preparation
  - Discovery
  - Filtering

- Quality Review

- Access

- Deep Web
  - Archiving
    - Index
    - Storage
  - Link Extraction
  - Fetching

- Capture
  - Temporal Coherence of Crawls
  - Noise Filtering

- Dynamic Pages
  - JavaScript, Flash
- Multimedia Content

- User

- Long-term Interpretability

24/07/2009
Follow the links....
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

1. Selection
2. Quality Review
3. Access
4. Long-term Interpretability
5. Index
6. Storage
7. Multimedia Content
8. Fetching
9. Link Extraction
10. Capture
11. Temporal Coherence of Crawls
12. Archivist

Preparation
Discovery
Filtering
Capture
Extraction
Quality
Review

User

LiWA Living Web Archive
24/07/2009
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

schools for pharmacy phh mortgage cendant songs ring tones community credit union houston philadelphia penn s ... settlement hawaii insurance commissioner debt coverage ratios auto loan refinance classic video games online what health insurance long beach schools financial credit union insurance umbrella policy disaster unemployment insurance ... mag mutual insurance company debit & credit chevron gas credit card money affiliate car loan application paradis casino photos progressive insurance claims office halloween bingo sheet bunion world poker open pharmacy mass ... services credit union mortgage rates outlook cover insurance arts administration degree credit counseling government lose weight casino star odds against 7 even party poker ipo

Compute the out degree compute tnhse out degree compute the out degree compute the 0ut degree compute the ... detree comute the out degree compute the out degree compute t5he out degree comute the out degree compute the out degree compute the oujt degree compute the out degree compute the out degree compute the out degree compute the ou...
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

2004 .de crawl courtesy: T. Suel
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

Selection
- Preparation
- Discovery
- Filtering

Noise Filtering

Capture
- Link Extraction
- Fetching

Deep Web

Access

Quality Review

Archiving
- Index
- Storage

Long-term Interpretability

Dynamic Pages
JavaScript, Flash

Multimedia Content

Temporal Coherence of Crawls

User

Link Extraction

24/07/2009
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

\[ t_{coherence} = [t_2, t_3) \]

- \( p_1 \)
- \( t_1 = t_5 \)
- \( t_2 \)
- \( t_3 \)
- \( t_4 = t_e \)

- \( p_2 \)
- \( p_3 \)
- \( p_4 \)
Coherence by Example

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

\[ p_1 \]

\[ p_2 \]

\[ p_3 \]

\[ p_4 \]

\[ t_1 = t_5 \quad t_2 \quad t_3 \quad t_4 = t_e \]
Coherence Analysis Technology

Temporal Coherence Report

Overview

<table>
<thead>
<tr>
<th>Source</th>
<th><a href="http://example.com/old-page.html">http://example.com/old-page.html</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total count</td>
<td>65016</td>
</tr>
<tr>
<td>Revisited pages</td>
<td>95046</td>
</tr>
<tr>
<td>Grad Duration</td>
<td>12.5 sec</td>
</tr>
<tr>
<td>Revisit Duration</td>
<td>11.7 sec</td>
</tr>
</tbody>
</table>

Changes

| Pages with changed links | 16 |
| Revisited pages | 95043 |
| Revisit Duration | 11.7 sec |

Details

Pages with changed links:
- http://example.com/old-page.com/old-page.html
- http://example.com/old-page.com/old-page.html
- http://example.com/old-page.com/old-page.html

Creation of automatically generated reports

Visualization of coherence defects
Long-term Interpretability

Selection
- Preparation
- Discovery
- Filtering

Noise Filtering
- Link Extraction
- Fetching

Capture

User

Deep Web

Quality Review
- Preparation
- Discovery
- Filtering

Access

Archival

Temporal Coherence of Crawls

Link Extraction

Dynamic Pages
- JavaScript
- Flash

Archiving
- Index
- Storage

Multimedia Content

Long-term Interpretability

24/07/2009
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

St. Piter Burh 1703
St. Petersburg 1703-1914
Petrograd 1914-1924
Leningrad 1924-1991
St. Petersburg 1991-present
Overview Process

Step 1: Word Sense Discrimination

Step 2: Tracking Evolution
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**
Find methods that are time independent for

- 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
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/