The WebLab Project at Cornell University

Felix Weigel
Postdoctoral Associate
Dept. of Computer Science, Database Group
Cornell University
Could you tell ...

- When did the term “blog” emerge on the web?
- How do rumors spread across communities?
- How to model herd behavior in social networks?
- How do recall campaigns affect a company’s reputation?
- Who is usually cheaper, Amazon or Barnes & Noble?
- How does the web graph evolve over time?
- How to improve PageRank?
Mission of the WebLab Project

- The web is an extraordinarily rich source of information
- Web archives have saved snapshots of the web over many years

- Goals of the WebLab project:
  - Make this information accessible without demanding high technical or computing expertise
  - Provide an infrastructure for archived web data to enable all kinds of research about the web
The WebLab Project at Cornell

- Joint project of Cornell University and the Internet Archive
- Started in early 2006
- Participants at Cornell:
  - Computer Science Department
  - Information Science Program
  - Center for Advanced Computing (former Cornell Theory Center)
  - Collaboration with Cornell’s Institute for the Social Sciences
Challenges

Making web data available
• Crawl, Transfer, Storage, Clean-up, Indexing

Making web data accessible
• Interfaces for search, browsing, data extraction
• Infrastructure for collaborative data curation

Using web data
• Use cases from different disciplines
• Interfaces to analysis tools and applications

Making web data available
• Crawl, Transfer, Storage, Clean-up, Indexing
Availability

- Crawling is done by the Internet Archive
- Compressed raw data is transferred to Cornell
  - Crawled web pages in *ARC Files*
  - Metadata (URL, title, crawl timestamp, ...) in *DAT Files*
- Three types of storage:
  - Tape archive for backups
  - Relational database for metadata
  - ARC files in file system
- Main challenge: *Scalability*
  - Clean-up: e.g., remove duplicate URLs and dangling links
  - Indexing: Hyperlink structure; full text
  - Parallelization: Off-line and on-line tasks on computing clusters
Accessibility

- Research begins with ... Search!
- What’s in the archive?

Browsing Historical Web Pages with the Internet Archive’s Wayback Machine

Full-text Indexing with NutchWAX

Metadata Search Form

<table>
<thead>
<tr>
<th>Field</th>
<th>Include In Result</th>
<th>Negate Restriction</th>
<th>Restriction</th>
<th>Example of Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageID</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>LOCn01W3L1br8IY3I2sA==</td>
</tr>
<tr>
<td>URLID</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>449xx0H0Q7frlen1Y6Lr==</td>
</tr>
<tr>
<td>CrawID</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>1 2 3</td>
</tr>
<tr>
<td>HostID</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>1rQXLeTqZwejDg4N4LT7qA==</td>
</tr>
<tr>
<td>URL.Protocol</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>http://</td>
</tr>
<tr>
<td>URL.Host</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td><em>%edu</em> OR <em>%org</em></td>
</tr>
<tr>
<td>URL.Port</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>80</td>
</tr>
<tr>
<td>URL.Path</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>%plc%</td>
</tr>
<tr>
<td>URL.Extension</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>.htm</td>
</tr>
<tr>
<td>URL.QueryString</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>?no_article=52</td>
</tr>
<tr>
<td>ArchivalTime</td>
<td>☐</td>
<td>☐</td>
<td>(Begin)</td>
<td>02/14/2005 08:30:22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(End)</td>
<td>NB: MONTH/MONTH/YEAR</td>
</tr>
<tr>
<td>IPAddress</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIMEType</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>text/html</td>
</tr>
<tr>
<td>Language</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>en-us-ascii 0.54363 1.21</td>
</tr>
<tr>
<td>Title</td>
<td>☐</td>
<td>☐</td>
<td></td>
<td>%photo%</td>
</tr>
</tbody>
</table>
Web Data Extraction

• Researchers often don’t care about web pages, but specific substructures inside the pages
  – Blog postings
  – Online forums
  – Social tags/categories
  – News headlines
  – Tables of content
  – Bibliographies
  – Product details
  – Customer reviews
Web Data Collaboration Server

*Data extraction*
- Writing extraction code is a tedious task
- Create tools to make the data easily accessible in a structured format (e.g., tables in a database)

*Data sharing*
- Extracting the same data repeatedly is a waste of time and storage space
- Let users share their data and extraction rules

*Data curation*
- Web data is often incomplete and erroneous
Project Status

• Availability:
  – Infrastructure for (partly) automated crawl download and backup
  – Metadata database of two full crawls
    • 3 billion pages, 20 billion distinct URLs, 130 billion links, 11.5 TB on disk
  – Full-text indexing of two Amazon.com sub-crawls in progress
    • 30 million pages, 164 GB compressed

• Accessibility:
  – Prototype of the Web Data Collaboration Server in progress
  – First deployment to Social Scientists at Cornell later this year
  – No-code web data extraction and sharing will soon be operational

• Usability:
  – Research on exploration of extracted data
  – Novel Methods for “informal” database search in progress
Who’s Involved in the WebLab Project

Cornell
- William Arms
- Selcuk Aya
- Manuel Calimlim
- Pavel Dmitriev
- Johannes Gehrke
- Dan Huttenlocher
- Jon Kleinberg
- Christoph Koch
- Blazej Kot
- David Lifka
- Ruth Mitchell
- Biswanath Panda
- Mirek Riedewald
- Lucia Walle
- Felix Weigel

Internet Archive
- John Aizen
- John Berry
- Kris Carpenter
- Tracey Jacquith
- Brewster Kahle
- John Lee
- Gordon Mohr
- Michael Stack