The ClueWeb09 Dataset

Jamie Callan\textsuperscript{1}, Mark Hoy\textsuperscript{1}, Changkuk Yoo\textsuperscript{2}, & Le Zhao\textsuperscript{1}

\textsuperscript{1}Language Technologies Institute  \hspace{1em} \textsuperscript{2}Daum Communications Corp
School of Computer Science  \hspace{1em} Republic of South Korea
Carnegie Mellon University
Why Build a New Web Dataset?

There are not many web datasets available for research:

- wt10g: 1.7 million pages from 1997
- gov2: 25 million pages from 2004
- uk-2006: A partial crawl of the .uk domain
  - Available from Yahoo! Research (?)

wt10g and gov2 are the most widely available
... but are not very representative of the web
Why Build a New Web Dataset?

The NSF / Google / IBM CluE cluster was available
• Many machines
• Lightly loaded in late 2008 / early 2009
• Willing to temporarily provide a fast network connection

Craswell & Fetterly’s breadth-first crawl (SIGIR ’08)
• 700 million pages (later extended to 1 billion)
• Inspirational
How We Built It:
An Initial Plan + Community Input

Initial plan: A **best-first** crawl of 700 million – 1 billion pages
• Approximate ‘Tier 1’ of a commercial search engine
• Complement the Craswell & Fetterly crawl

A white paper was circulated and revised several times

A broad community commented
• Colleagues in the research community
• Google, Microsoft, Yahoo
• NIST
How We Built It:
Key Ideas That Shaped the Dataset

Scope

• Be big enough to be credible
  – 500M to 1B web pages

• Unfiltered content
  – Give researchers the real web
    » Spam, pornography, …

• Avoid temporal skew
  – Complete the crawl quickly
How We Built It:
Key Ideas That Shaped the Dataset

Languages
• 50% English
  – Provide high coverage of one language
• 50% the next 9 most important languages on the web
  – Chinese, Japanese, Korean
  – Spanish, French, German, Portuguese, Italian
  – Arabic
How We Built It:
Key Ideas That Shaped the Dataset

Include the full English wikipedia

• A last minute addition

  … thanks to the Wikimedia Foundation for enabling this
How We Built It:
The Crawler

We used a modified version of the Nutch crawler
- Open source, written in Java
- Runs under Hadoop
- Crawl ordered by OPIC (an approximation of PageRank)

Major modifications
- Added language id
- Improved OPIC propagation for redirected links
- Many modifications to improve crawler speed
- Modifications to improve crawler reliability
How We Built It: Basic Crawler Architecture

• **Get N urls**
  – Initially from the seed file
  – Later selected by OPIC from the web graph

• **Send urls to multi-node / multi-threaded download processes**
  – Download urls, trying to be nice, spread the load, etc
  – Each process ran for about 2 hours

• **Process downloaded pages**
  – Extract urls, language id, update web graph, …

• **Repeat**
How We Built It: Crawl Seeds

There were two types of seed URLs

• urls from an earlier 200 million page crawl
  – URLs that had high OPIC scores

• urls returned by commercial search engines
  – Submit query, add top N results to the seed file
  – Search engines: Google, Yahoo, MSN, Baidu (Chinese)
How We Built It: Crawl Seeds From Commercial Search Engines

Queries were generated in a variety of ways

• Selected from the AOL query log
  1,050 most frequent queries + 1,050 random queries
  Translated to other languages by Google Translate

• Generated from DMOZ category names
  2,000 queries from largest categories (up to depth 3)
  – E.g., “Northern Mariana Islands”, “Snowbiking”
  Translated to other languages by Google Translate

• Provided by Yahoo: 1,000 most frequent queries × 9 languages

• Provided by Sogou: 1,000 most frequent queries (Chinese)
The Crawl

© 2009, Jamie Callan
## Language Distribution

<table>
<thead>
<tr>
<th>Rank</th>
<th>Language</th>
<th>Internet Users (%)</th>
<th>Crawl Goal (%)</th>
<th>Crawl Actual (%)</th>
<th>Crawl Actual (Million Pages)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>29.40%</td>
<td>50.00%</td>
<td>48.41%</td>
<td>503.9</td>
</tr>
<tr>
<td>2</td>
<td>Chinese</td>
<td>18.90%</td>
<td>17.00%</td>
<td>17.05%</td>
<td>177.5</td>
</tr>
<tr>
<td>3</td>
<td>Spanish</td>
<td>8.50%</td>
<td>7.70%</td>
<td>7.62%</td>
<td>79.3</td>
</tr>
<tr>
<td>4</td>
<td>Japanese</td>
<td>6.40%</td>
<td>5.80%</td>
<td>6.47%</td>
<td>67.3</td>
</tr>
<tr>
<td>5</td>
<td>French</td>
<td>4.70%</td>
<td>4.20%</td>
<td>4.89%</td>
<td>50.9</td>
</tr>
<tr>
<td>6</td>
<td>German</td>
<td>4.20%</td>
<td>3.80%</td>
<td>4.79%</td>
<td>49.8</td>
</tr>
<tr>
<td>7</td>
<td>Arabic</td>
<td>4.10%</td>
<td>3.70%</td>
<td>2.80%</td>
<td>29.2</td>
</tr>
<tr>
<td>8</td>
<td>Portuguese</td>
<td>4.00%</td>
<td>3.60%</td>
<td>3.61%</td>
<td>37.6</td>
</tr>
<tr>
<td>9</td>
<td>Korean</td>
<td>2.40%</td>
<td>2.10%</td>
<td>1.74%</td>
<td>18.1</td>
</tr>
<tr>
<td>10</td>
<td>Italian</td>
<td>2.40%</td>
<td>2.10%</td>
<td>2.62%</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>Rest Others</td>
<td>15.10%</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
</tr>
</tbody>
</table>
A Blunder

• Information about url redirection was discarded
  – A major problem for people who care about web graphs

• During the summer, we recreated redirect information for the Category B subset of the data
  – Available on the wiki
Summary of the ClueWeb09 Dataset (Category A)

• **Size (count):** 1.04 billion web pages
• **Size (TB):** 25 Terabytes (uncompressed)
• **Crawl period:** January & February, 2009
• **Crawl order:** OPIC (an approximation of PageRank)

• **7,944,351,835 outlinks**
  – 4,780,950,903 unique urls
The Category B Subset

The Category B subset was defined to make it easier for groups not yet ready to scale up to 1 billion documents.

**Size:** 50 million documents
- About 2x the gov2 dataset
- 454,075,638 outlinks
  - 428,136,613 unique urls

There were no strong opinions about how to define the subset
- So … we picked something convenient
The Category B Subset

What does it consist of?

• English crawl seeds: 2.5 million
• Crawled pages: 41.8 million
• English wikipedia: 6.0 million

This might be an unusual subset of the web … or not

• Highly ranked pages for reasonable (?) queries
• Pages closely linked to those pages
• English wikipedia
ClueWeb09-Image Dataset

Some research requires text + graphics data

- E.g., user studies
ClueWeb09-Image Dataset

Some research requires text + graphics data
• E.g., user studies

After the text crawl was complete, we crawled the image data
• Size (count): 870 million images
• Size (TB): 23 Terabytes (mostly uncompressable)
• Crawl period: May – July, 2009

Currently being transferred back to CMU
Dataset Related Services

Carnegie Mellon hosts a variety of dataset-related services

- The ClueWeb09 wiki
  - Language id, web graph, redirects, working with warc, ...
- Derived data (e.g., PageRank data)
- Indri search engine for Category A (English) and Category B
- Page rendering service
What We Wish We Had Done Differently

In order of importance…

1. Save redirect information
   – Deleted accidentally due to miscommunication

2. Complete the crawl in 30 days, instead of 60 days
   – An original goal, not achieved

3. Include wikipedias for each of the 10 languages
   – Wikipedia was a (very) late dataset requirement

4. Gather text + images, rather than text followed by images
   – We had the software, but not the bandwidth or disk
What Next?

We hope that there will be more large web datasets
• It was an interesting experience
  – We learned a lot, we would do it again
• The research community needs more good web datasets

Should the next big web dataset use the same approach?
• The IR community should debate what it wants next
  – Redo ClueWeb09 one year later?
  – Weekly crawls of important / fast changing sites?
  – …
We Couldn’t Have Done It Without A Whole Lot of Help

Nick Craswell
Dennis Fetterly
Jim French
Don Metzler
Ian Soboroff
… and many others