The Lemur Project
And its ClueWeb12 Dataset

Jamie Callan
Language Technologies Institute
Carnegie Mellon University
callan@cs.cmu.edu

Outline

• Introduction to the Lemur Project
  – “A dozen years of service to the IR community”
• ClueWeb09
• The making of ClueWeb12
• What might lie ahead
What is the Lemur Project?

The Lemur Project is a collaboration between Croft & Callan
- Plus students and staff
  (not as many as you might expect)

The Lemur Project creates community research infrastructure
- Open-source software
- Widely-available datasets
- Services that support research and education

What is the Lemur Project? Software

The Lemur Toolkit
- Search engines, clustering, LSI, summarization, distributed IR, …
- Mostly obsolete and discontinued

Indri and Galago

The Lemur Toolbar
What is the Lemur Project?

Datasets

- ClueWeb09
  + various derived data
- ClueWeb12
- ClueWeb12++

What is the Lemur Project?

Services

Search services
- Interactive search, script-based search
- Batch query service
- Page rendering service

Attribute Lookup Service
What is the Lemur Project?

Services

Search services
• Interactive search, script-based search

ClueWeb09 Batch Service

1. Select the index to query:
   • ClueWeb09 Category A English
   • ClueWeb09 Category B

2. Select the maximum number of results per query:
   • 100
   • 1000

3. Select output format:
   • Indri default format
   • treex_eval format

4. Select a file of queries to upload:

5. Upload your file:
What is the Lemur Project?

Services

Search services
• Interactive search, script-based search

ClueWeb09 Rendering Service

The ClueWeb09 Rendering Service provides fully-rendered pages (text and images) from the ClueWeb09 dataset to organizations that have a license to use the data.

A password is required to use this service. Organizations which have a license to use the data may request access by contacting Jamie Callan (callan at cs dot csu dot edu).

ClueWeb09 Trec ID: cluesweb09-mw0111-50-37515

Submit

ClueWeb09 Attribute Lookup Service

1. Select a type of lookup:
   - Use document ids (keys) to lookup urls (values)
   - Use document ids (keys) to lookup CMU PageRank priority (values)
   - Use document ids (keys) to lookup WordNet synset scores (values)
   - Use urls (keys) to lookup document ids (values)

2. Select a file of keys to upload:

3. Upload your file:

© 2012, Jamie Callan
Why Do We Do It?

I view the Lemur Project as a set of opportunities
• An opportunity to work with a friend
• An opportunity to work on state-of-the-art projects
• An opportunity to attract great students
• An opportunity to find out how good our work really is
• An opportunity to help others

Community research infrastructure has a long lifespan
• Stable software and datasets for our own research

Most of my motivation is selfish, not altruistic

What is Involved? (50,000 Foot Perspective)

Faculty (2)
• Raise money, set goals and requirements, keep everyone on track

Students (multiple, varies)
• Research, prototypes, innovation
• Students affiliate with the project for awhile, and then drift away

Programmers (2)
• Software development, maintenance, documentation, support
• Programmers provide stability and consistency
What is Involved?  
(50,000 Foot Perspective)

Everything starts with a grant proposal
• A set of tasks
• An assignment of tasks to UMass, CMU, or both
• Resources (money for students and hardware)

The two groups work somewhat independently
• Weekly meetings within a group
• Monthly meetings between the two groups
• Email communication whenever necessary

Everyone is trusted to do their part (and mostly they do)

Funding

Funding for long-term projects is a challenge
• Funding sources are wary of funding a project for a long time
  – It reduces funding for other projects
• Funding sources (and their funding sources) prefer new work

How do we do it?
• There is no magic recipe
• A sequence of 3-4 year grants, each with its own personality
  – New science
  – Community research infrastructure
  – …
Evolution of the Lemur Project

The Lemur Toolkit
- Summarization
- Distributed IR
- Clustering
- LSI

Indri

Galago

ClueWeb12

ClueWeb09

ClueWeb12++

Dates are approximate

ClueWeb09

Why?
- Jamie was using Yahoo’s M45 to do crawling for another project
  - 200M web pages, but not a good general-purpose dataset
- NSF’s Cluster Exploratory (CLUE) program was willing to provide resources for a more realistic web dataset
  - Funding to support labor and a small amount of storage
  - Dedicated use of the Google/IBM cluster for several months
- CMU university administration was willing to allow the distribution of a dataset for research purposes

Interest + experience + hardware + institutional support
ClueWeb09

Crawler: A heavily modified version of Nutch
• OPIC, url queueing, load balancing, language filtering, ….

Hardware:
• 100 node Hadoop cluster (retired search engine hardware)
• 33 TB of usable disk
• 1 Gbs network (supposedly – we never came close to that)

Timespan:
• January – February, 2009

ClueWeb09 Postprocessing

1. Group by language
2. Segment each language into chunks of 50M documents
3. Sort each segment by url (to improve compression)
4. Segment into files of about 1 GB each
5. Organize into a directory hierarchy (100 files per directory)
6. Split into 4 segments of 1.5 TB each (for shipping)
ClueWeb09 Distribution

Distribution process
• A license signed by CMU and the other organization
  – A small percentage (try to) negotiate modifications
• Process payment (usually)
• Order and copy disks
• Ship disks

More than 250 copies licensed around the world

Why not distribute via the network?
• We would love to do this!
• But … using 4-8 TB of CMU network bandwidth each week (more in peak periods) would not be nice

Why not distribute via Amazon?
• It is expensive
  – 4 TB of storage (US$ 466/month) and I/O (US$ 490/copy)
  – 4 TB of network bandwidth at the recipient institution

We expect to support this at some point
ClueWeb12

Why?
• Datasets with different properties produce better research
• We might be able to generate a better dataset
  – Less porn, less spam
  – A valid web graph
  – Capture images and tweeted urls
• Amortize learning costs over two datasets
  – This was naïve 😞
• We still had access to big computer clusters
  – We were wrong about this 😞

Experience + institutional support + over-confidence

ClueWeb12: Seeds

There were 2,820,500 seeds

Group 1:
• Select 10M ClueWeb09 urls with highest PageRank scores
• Discard urls that are not in the best 10% of Waterloo spam scores

Group 2:
• 262 most popular urls from English-speaking countries (Alexa)

Group 3:
• 5,950 travel sites provided by Charlie Clarke
ClueWeb12: Blacklist

Ignore sites in the following URLBlacklist.com categories
• Pornography
• Malware, phishing, spyware, virusinfected
• Filehosting, filesharing

Ignore sites that opted out
• About a dozen sites

ClueWeb12: Other Ignored Files

The crawler ignored urls that didn’t appear to be text
• Flash, audio, video
• Compressed files

The crawler truncated files longer than 10MB in size
ClueWeb12: Images

The crawler saved files that allow pages to be rendered
• Some user-studies require fully-rendered pages
• css, xml, javascript, …
• jpg, gif, …

ClueWeb12: Crawling Architecture

Crawler
• The Internet Archive’s Heritrix crawler

Hardware
• 7 nodes on our boston cluster

Timespan:
• February 10 – April 10, 2012
ClueWeb12: Twitter urls

A Twitter feed was monitored during the crawl

Harvest urls from English tweets
• Crawl the url
• Add the domain to the web crawl


Wikipedia was crawled in the same way as every other site
• No special seeds, no special treatment

We downloaded and included a full copy of wikipedia
• XML format, so not quite compatible with crawled content
• A useful resource for a variety of research purposes
wikitravel was crawled completely

We downloaded and included a full copy of wikitravel
• XML format, so not quite compatible with crawled content
• A useful resource for a variety of research purposes

Requested by Charlie Clarke, for TREC purposes
ClueWeb12: Crawling Speed

**Goal:** 1.0 billion pages (+ images) in 8 weeks (20.0M/day)
- ClueWeb09 collected 1B pages (without images) in 8 weeks

**Reality:** 1.2 billion pages (+ images) in 13 weeks (13.3 M/day)
- 1.16 TB / day
- We were the largest user of CMU bandwidth 5-6 for months
  - Averaging at least 30% (?) of total campus network capacity
  - The university was very (very) nice about it
  » Thank you CMU!

ClueWeb12: Crawl Statistics

104 TB of data downloaded
- **Text/html:** 37 TB, 1.2 B URIs
- **Other:** 67 TB, 1.0 B URIs
  - Over 7,800 mime types (!)

Collecting just text/html would reduce effort by 64%
- Researchers say that they need more than just the html
- We will see if this is true
ClueWeb12: Dataset Organization

A crawler instance writes to a set of warc.gz files
- Multiple files written simultaneously, to improve I/O

A warc file contains different types of information
- http response headers
- Web pages
- Css, javascript, xml, images, …

Most IR people don’t want the raw crawler output
- Too many different types of information
- Poor compression ratio (i.e., more disks to ship)

Select 6 warc files written by instance i at about the same time
- Discard everything except web pages
- Merge the 6 files
- Sort by domain name
- Segment, to produce files of about 1 GB, uncompressed
- Compress
  - We get about 6x compression

Organize files into a TREC-style directory hierarchy
- 100 files per subdirectory
ClueWeb12: Dataset Organization

- Seed urls
- General crawl + tweeted urls
- Wikipedia
- Wikitravel

Segment 1 (50M pages)
Segment 2 (50M pages)
Segment 3 (50M pages)

ClueWeb12: English Filtering

Non-English pages were removed during post-processing
- ClueWeb09 used TextCat for language id
  - a heuristic classifier that uses 300 frequent n-grams
- ClueWeb12 uses langdetect for language id
  - Naïve Bayes with character n-grams
  - Open-source, published on code.google.com
  - We used just the first 2,400 characters of each document
    » Improved speed
    » Didn’t hurt average accuracy in our tests
Pornography was removed during post-processing
- Crawl 1M porn pages from blacklist sites (positive instances)
- Crawl 2.5M pages starting from 18K good seeds (negative instances)
- Select the best 750 features (chi-square)
- Use Galago to generate the feature vector for each document
  - kstem terms, tf feature values
- Train C4.5 and Naïve Bayes text classifiers
  - 300,000 instances, 10x cross-validation
  - C4.5: 95% accuracy, even distribution of errors
  - NB: 85% accuracy, skewed distribution of errors
- Secondary filter based on density of “adult” words

ClueWeb12: Spam Filtering

Part of our original plans, but not done
ClueWeb12:
Filtering Statistics

Preliminary (possibly wrong) statistics
• 51% of URIs downloaded were text, 49% graphics or other
• 8% (?) of text URIs were non-English
• ???% of text URIs were “adult”

ClueWeb12:
Availability

Distribution planned for September 1

Licensing terms similar to ClueWeb09
• Free license
• A fee to cover distribution costs
  – Disks, labor, shipping
  – The fee declines as disk prices decline and capacity increases
What Next for the Lemur Project?

**Software**
- Possibly greater emphasis on Galago
- Possibly greater emphasis on Indri support for scientific search

**Datasets**
- ClueWeb12++: ClueWeb12 + social media data

**Services**
- Faster ClueWeb09 and ClueWeb12 search

What do you need?

Thanks!