



## 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

3

© 2012, Jamie Callan

## 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

4

© 2012, Jamie Callan

## What is the Lemur Project? Datasets



### ClueWeb09

+ various derived data

### ClueWeb12

### ClueWeb12++

5

© 2012, Jamie Callan

## What is the Lemur Project? Services



### Search services

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

### Attribute Lookup Service

6

© 2012, Jamie Callan

## What is the Lemur Project? Services





### ClueWeb09 Category A English

Our computational resources are limited, so we require programs to abide by our [usage policy](#).  
If you need to run a set of queries, please consider the [Batch Query Service](#), which is much more efficient.

©2009 The Lemur Project. Powered by the [Lemur Toolkit](#). Search engine last updated May 12 2011, Version 3.1. Contact [admin@lemurproject.org](mailto:admin@lemurproject.org) for more information.

7

© 2012, Jamie Callan

## 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  
 trec\_eval format
4. Select a file of queries to upload:
5. Upload your file:

8


© 2012, Jamie Callan

## 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 cmu dot edu).

ClueWeb09 Trec ID:

Check to use only cached images  
*(if checked, missing images will not be requested from the original page)*

9

© 2012, Jamie Callan

## What is the Lemur Project? Services



### Search services

- Interactive search, script-based search
- Batch query service

### 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 priors](#) (values)
- Use document ids (keys) to lookup [Waterloo spam scores](#) (values)
- Use urls (keys) to lookup document ids (values)

2. Select a file of keys to upload:

3. Upload your file:

10

© 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**

11

© 2012, Jamie Callan

## 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

12

© 2012, Jamie Callan

## 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)

13

© 2012, Jamie Callan

## 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

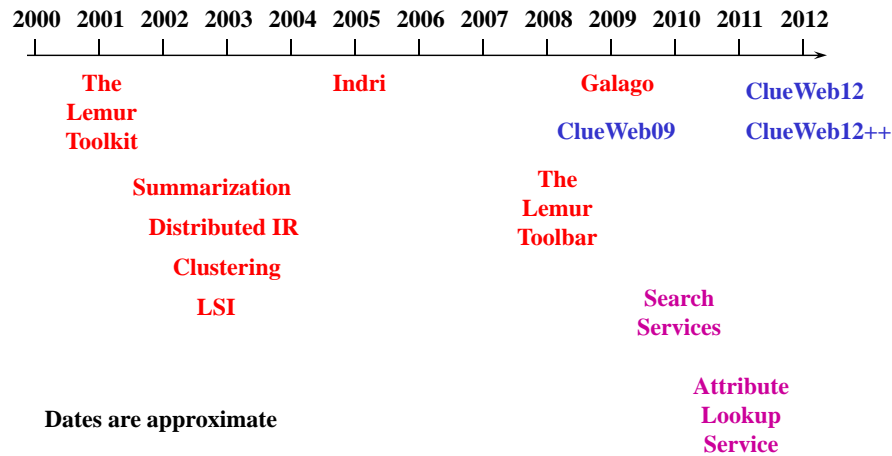
### How do we do it?

- There is no magic recipe
- A sequence of 3-4 year grants, each with its own personality
  - New science
  - Community research infrastructure
  - ...

14

© 2012, Jamie Callan

## Evolution of the Lemur Project



Dates are approximate

15

© 2012, Jamie Callan

## 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**

16

© 2012, Jamie Callan



## 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 useable disk
- 1 Gbs network (supposedly – we never came close to that)

**Timespan:**

- January – February, 2009

17

© 2012, Jamie Callan

## 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)**

18

© 2012, Jamie Callan

## 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**

19

© 2012, Jamie Callan

## ClueWeb09 Distribution

### 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**

20

© 2012, Jamie Callan

## 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**

21

© 2012, Jamie Callan

## 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

22

© 2012, Jamie Callan

## **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

23

© 2012, Jamie Callan

## **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**

24

© 2012, Jamie Callan

## ClueWeb12: Images

### The crawler saved files that allow pages to be rendered

- Some user-studies require fully-rendered pages
- css, xml, javascript, ...
- jpg, gif, ...

25

© 2012, Jamie Callan

## ClueWeb12: Crawling Architecture



### Crawler

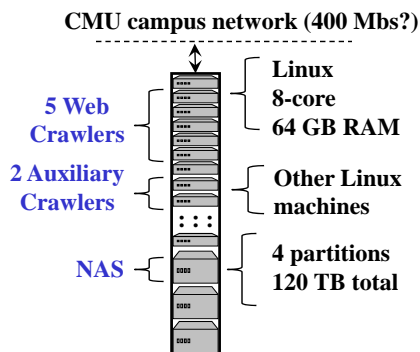
- The Internet Archive's Heritrix crawler

### Hardware

- 7 nodes on our boston cluster

### Timespan:

- February 10 – April 10, 2012



26

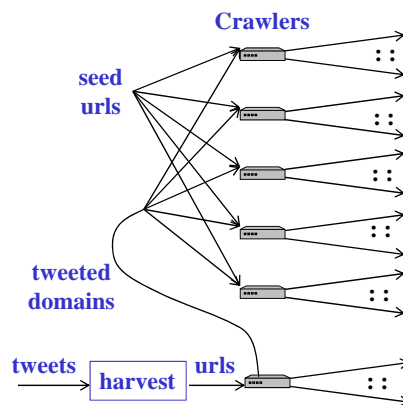
© 2012, Jamie Callan

## 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



27

© 2012, Jamie Callan

## ClueWeb12: Wikipedia



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

28

© 2012, Jamie Callan

## ClueWeb12: Wikitravel



wikitravel was crawled completely

We downloaded and included a full copy of wikitravel

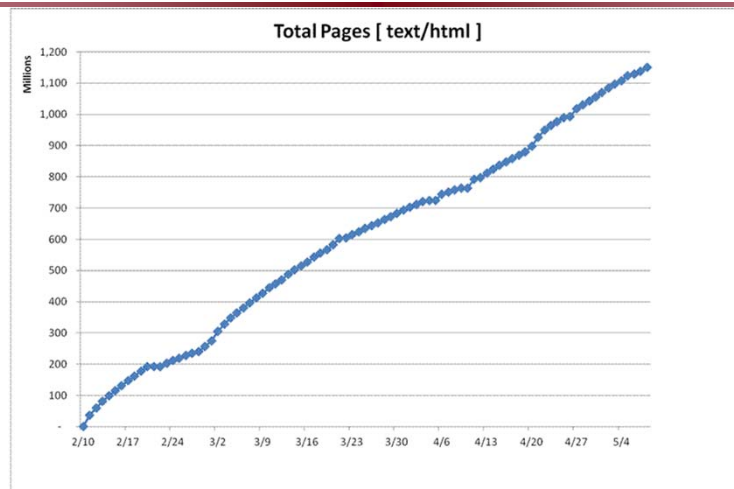
- XML format, so not quite compatible with crawled content
- A useful resource for a variety of research purposes

Requested by Charlie Clarke, for TREC purposes

29

© 2012, Jamie Callan

## ClueWeb12: Crawling Speed



30

© 2012, Jamie Callan

## 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!

31

© 2012, Jamie Callan

## 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

Mime Types	URLs	TB
text/html	1,236,987,268	37.07
image/jpeg	490,309,571	22.60
image/gif	138,440,291	1.56
image/png	89,206,456	2.29
text/xml	60,025,026	0.60
application/pdf	47,743,025	30.69
text/dns	23,668,730	0.00
text/css	21,000,186	0.21
text/plain	19,186,553	0.80
application/x-javascript	12,681,242	0.29
application/javascript	8,632,328	0.19
unknown	7,550,045	0.03
application/atom+xml	5,562,101	0.14
application/rss+xml	5,243,933	0.08
application/octet-stream	4,660,004	1.37
text/javascript	3,421,667	0.08
application/xml	3,351,094	0.04
application/msword	3,018,277	0.61
image/x-icon	2,977,166	0.01
application/x-shockwave	1,912,243	0.50
image/jpg	1,616,069	0.04
image/pjpeg	944,658	0.05
text/calendar	857,466	0.00

32

© 2012, Jamie Callan



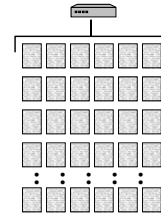
## ClueWeb12: Dataset Organization

### 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)

33

© 2012, Jamie Callan

## ClueWeb12: Dataset Organization

### 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

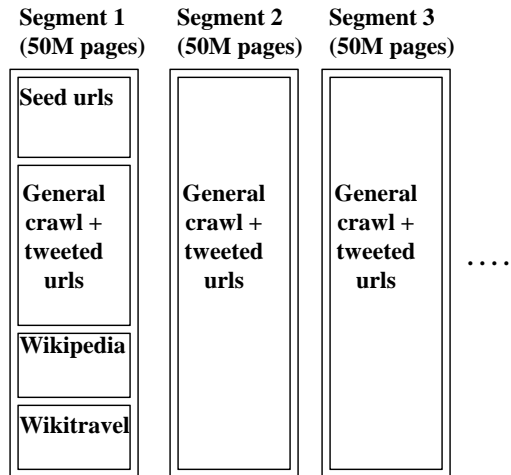
### Organize files into a TREC-style directory hierarchy

- 100 files per subdirectory

34

© 2012, Jamie Callan

## ClueWeb12: Dataset Organization



35

© 2012, Jamie Callan

## 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](http://code.google.com)
  - We used just the first 2,400 characters of each document
    - » Improved speed
    - » Didn't hurt average accuracy in our tests

36

© 2012, Jamie Callan

## ClueWeb12: Pornography Filtering

### 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

37

© 2012, Jamie Callan

## ClueWeb12: Spam Filtering

Part of our original plans, but not done

38

© 2012, Jamie Callan

## 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”

39

© 2012, Jamie Callan

## 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

40

© 2012, Jamie Callan

## 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?**

41

© 2012, Jamie Callan

---

**Thanks!**