

Nutch, Open-Source Web Search

Doug Cutting
<doug@nutch.org>

Lucene is...

- A mature Apache open-source project;
- Java library for text indexing and search;
 - Not an application;
- A large community of contributors;
- The search technology behind a lot of web sites & applications (ZOË, JIRA, Lookout, Furl, etc.)
- <http://jakarta.apache.org/lucene/>
- A book out this summer!

Nutch is...

- A young open-source project;
- Web search application software;
- A few part-time paid developers;
- A growing number of contributors;
 - paid and un-paid.
- Behind a growing number of sites.

Nutch isn't...

- A business;
 - But is a non-profit legal entity to own copyright;
 - No employees.
- A search site;
 - But want to power lots of search sites;
 - From domain-specific, to whole-web.
- A research project.
 - But want to be platform for research.

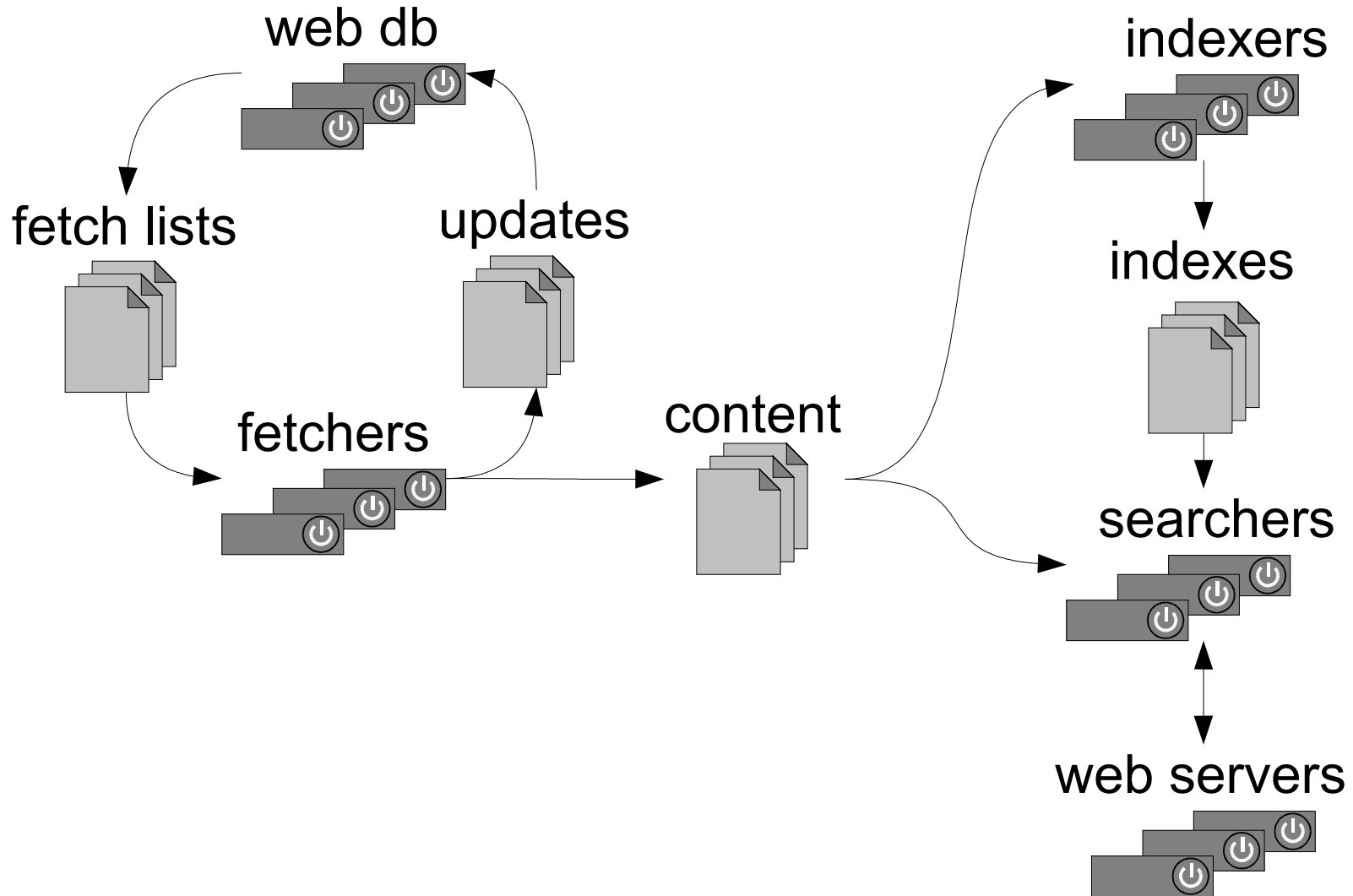
Nutch's Civil Goals

- Increase transparency of web search.
 - search is essential to internet navigation
 - yet algorithms are secret
- A free, open-source implementation should help.

Nutch Technical Goals

- Scale to entire web
 - pages on millions of different servers
 - billions of pages
 - complete crawl takes weeks
 - very noisy
- Support high traffic
 - thousands of searches per second
- State-of-the-art search quality

Nutch Architecture



Web Database

- Page Database
 - Used for fetch scheduling.
- Link Database
 - Represents full link graph.
 - Stores anchor text associated with each link.
 - Used for:
 - Link analysis;
 - Anchor text indexing.
- This is not an RDBMS application!

Scalability

- To meet scalability goals:
 - multiple simultaneous fetches
(100+ pages/second / CPU, ~10M / day)
 - parallel, distributed db update
(100M pages @ 100 pages/second / CPU)
 - distributed search
(2-20M pages, 1-40 searches/second / CPU)

But intranets are different!

Part 1: Scale

- Fetch, DB & search can all run on one box.
- Complete crawl takes only hours.
- Handful of servers on LAN—easy to overload!
- Lessons:
 - need to throttle fetcher
 - need much simple operation—single command
 - can crawl deeper

But intranets are different!

Part 2: Control

- cleaner content
- knowledge about structure of sites (cgi's, etc)
- lessons:
 - can index more dynamic content (cgi's, etc.)
 - can customize crawler better to site

But intranets are different!

Part 3: Quality

- only ~1M pages
- lesson:
 - not great for link analysis
 - but plenty for anchor text

Intranet How To

Step 1: Install

- Nutch requires only Java & JSP.
- Download & unpack.
- No admin GUI (yet)
 - command line
 - config files

Intranet How To

Step 2: Configure

- Specify root URLs.
- Specify URL filters.
 - a separate config file, containing regexps
 - each either includes or excludes URLs
 - first matching pattern determines fate of each URL
- Optionally, add a config file specifying:
 - delay between fetches
 - num fetcher threads
 - levels to crawl

URL Filter Example

```
# skip image and other suffixes
-\.(gif|jpg|pdf|doc|sit|rtf|exe)$

# skip URLs w/ certain characters
-[?*!@=]

# accept hosts in nutch.org
+^http://([a-z0-9]*\.)*nutch.org/

# skip everything else
-.
```

Intranet How To

Step 3: Test Run

- Crawl just a few levels deep, ~5
- Examine output log for:
 - warnings
 - exclude some file types?
 - sites hit too hard (e.g., infinite sites)
 - exclude some hosts or paths
 - sites not hit?
 - add more root urls, or crawl deeper

Intranet How To

Step 4: Finish up

- customize the look and feel
 - by default, uses XSLT template
 - or can roll your own.
- perform a full crawl (depth = ~10)
- tell folks about it!

Advantages

- Free!
- Scalability & quality.
- Open source easier to:
 - Customize
 - e.g., ranking, operators, look & feel, bells & whistles
 - Debug
 - You've got the full source!
 - Extend
 - Non-HTTP, non-HTML content, metadata, etc.

Demonstrations

- <http://labs.yahoo.com/demo/nutch/>
- <http://www.mozdex.com/search.html>
- <http://www.objectssearch.com/en/search.html>
- <http://devjr.cws.oregonstate.edu:8080/>
- <http://www.nutch.org/>

Preliminary Evaluation at OSU: Nutch versus a Google Appliance

- For OSU's top-25 queries:
 - 9 queries nutch and google were both perfect: 10/10
 - 2 queries nutch was slightly better
 - 2 queries google was slightly better than nutch
 - 1 query google was much better: 10 to 6
 - 1 query google was much better: 10 to 6
 - 1 query both scored 5
 - Google Appliance had a slight overall advantage.

Check it out!

<http://www.nutch.org/>



doug@nutch.org