

ZIGGY

THE EXPERIMENTAL WEB CRAWLER

presents

ADVENTURES IN WEB CRAWLING WITH ERLANG AND SOLR

by Alan G. Labouseur - alan@Labouseur.com

Warning: You must be over two years of age to view the contents of this document. Any reproduction, retransmission, redistribution, reeducation, memorization, retardation, reverberation, retribution or other re-generalization either explicit or implicit must not be attempted without the express written consent of Ian Fleming, Ted Codd, and Stevie Ray Vaughan.

The reader of this document may be subject to alien abduction(s) and/or visitation(s). Any such events shall be at the users own expense and the creator of this document shall not be held liable for any damages that may be incurred from said abduction or visitation. If any part of this agreement shall at a later date be determined to be void, invalid, non-existent, incoherent, or otherwise not applicable, the rest of this agreement shall remain in effect for a period of no less than 10 years beyond the proof of $P \neq NP$.

Introduction

This document describes Ziggy, the experimental web crawler, written by Alan G. Labouseur.

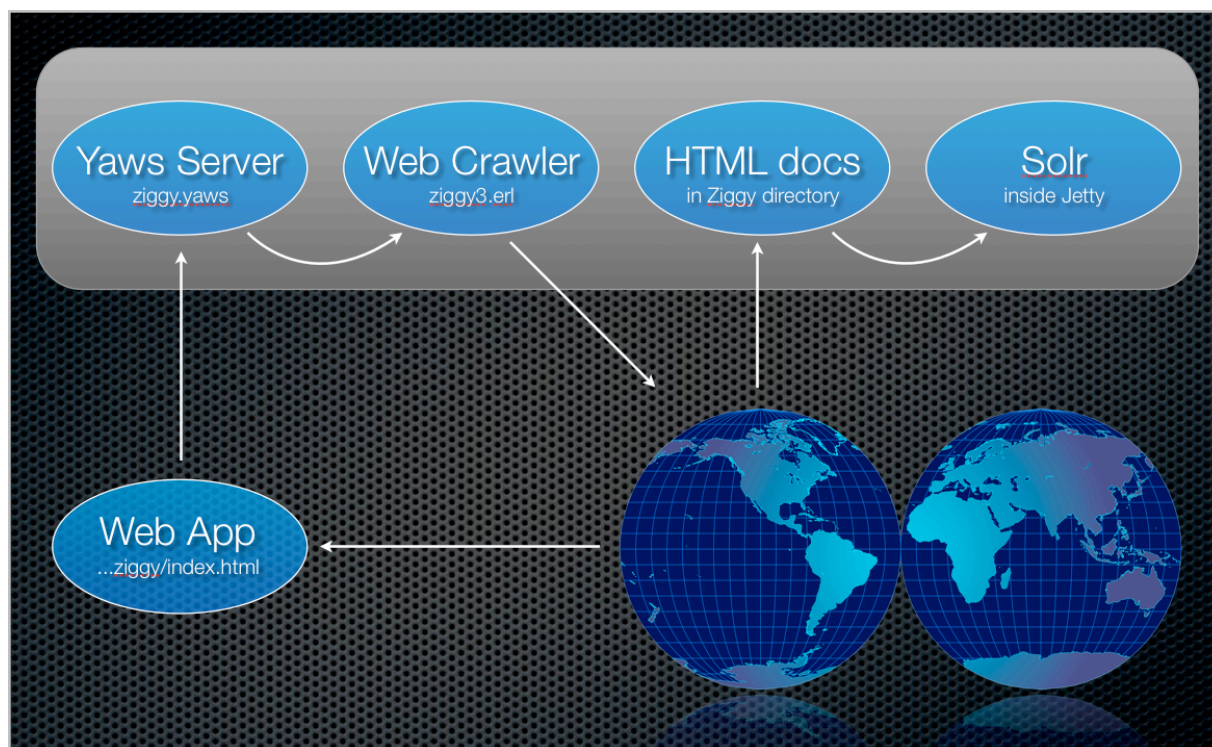
Goals

Our goals, in general, were “to create a (1) web crawler and (2) demonstrate it with an indexer.” Specifically, to “write a simple single-threaded web crawler that would find new pages on the web and pass text found in them as documents to be indexed by Lucene. Starting from a single input URL the crawler should download a page, then wait at least 5 seconds before downloading the next page. [The] program should find other pages to crawl by parsing link tags found in previously crawled documents.”

Ziggy accomplishes the goals of crawling and indexing in two parts:

1. A web crawler written in Erlang with a web-based interface written in HTML5, CSS3, and JavaScript, using Yaws (Yet Another Web Server) as the server, also written in Erlang.
2. Indexing, viewing, and analysis with a web-based interface to Lucene called Solr (which itself uses Jetty) customized by Alan and integrated into the Ziggy web interface.

Architecture



Demo:1 - Crawl the Web

The main Ziggy interface is available online at <http://www.labouseur.com/projects/ziggy/> . This is accessible anywhere in the world you have uncensored internet connectivity. As of this writing, however, Ziggy uses Yaws and Solr on your local computer only, so you'll need those installed before you can do much of anything other than admire the retro-Apple][look of the web app. See the configuration section for details on how to get all of the Ziggy components, Yaws, Solr, and helpful resources for them.

```
Z I G G Y ][
. . . initializing YAWS . . .

] Begin crawling at 

] Stop crawling after  sites. 

Erlang R14B03 (erts-5.8.4) [source] [not 64-bit] [smp:8:8] [rq:8] [async-threads:0] [hipe]
[kernel-poll:true] [Alan]
Result: "Crawling compelte."
Crawl History:
["http://www.labouseur.com/projects/ziggy/test/rush.html",
"http://www.labouseur.com/projects/ziggy/test/asimov/",
"http://www.labouseur.com/projects/ziggy/test/bond/",
"http://www.labouseur.com/projects/ziggy/test/b5.html",
"http://www.labouseur.com/projects/ziggy/test/trek/",
"http://www.labouseur.com/projects/ziggy/test/"]

] Remember to Index
] View data in Solr 
```

```
-rwxr-xr-x@ 1 Alan Alan 377 Nov 24 10:21 kill.sh
-rwxr-xr-x@ 1 Alan Alan 1283 Nov 24 10:21 post.sh
-rwxr-xr-x@ 1 Alan Alan 100 Nov 25 15:24 start-solr.sh
-rwxr-xr-x@ 1 Alan Alan 74 Nov 26 09:14 start-yaws.sh
-rw-r--r-- 1 Alan Alan 1924 Nov 26 11:07 ziggy-2011-11-26-1.html
-rw-r--r-- 1 Alan Alan 561 Nov 26 11:07 ziggy-2011-11-26-2.html
-rw-r--r-- 1 Alan Alan 666 Nov 26 11:07 ziggy-2011-11-26-3.html
-rw-r--r-- 1 Alan Alan 528 Nov 26 11:07 ziggy-2011-11-26-4.html
-rw-r--r-- 1 Alan Alan 450 Nov 26 11:07 ziggy-2011-11-26-5.html
-rw-r--r-- 1 Alan Alan 543 Nov 26 11:07 ziggy-2011-11-26-6.html
-rw-r--r--@ 1 Alan Alan 48052 Nov 26 10:57 ziggy.yaws alias
-rw-r--r-- 1 Alan Alan 4936 Nov 26 11:07 ziggy3.beam
-rw-r--r--@ 1 Alan Alan 14945 Nov 26 10:56 ziggy3.erl
```

Demo:2 - Indexing

Index with Lucene by posting the new HTML files to Solr, with `post.sh`.

```
./post.sh
Posting file ziggy-2011-11-26-1.html to http://localhost:8983/solr/update
<?xml version="1.0" encoding="UTF-8"?>
<response>
<lst name="responseHeader"><int name="status">0</int><int name="QTime">8</int></lst>
</response>
.
.
Posting file ziggy-2011-11-26-6.html to http://localhost:8983/solr/update
<?xml version="1.0" encoding="UTF-8"?>
<response>
<lst name="responseHeader"><int name="status">0</int><int name="QTime">3</int></lst>
</response>

Committing.
<?xml version="1.0" encoding="UTF-8"?>
<response>
<lst name="responseHeader"><int name="status">0</int><int name="QTime">22</int></lst>
</response>
Ziggy ]]
```

Demo:3 - Results in Solr

All Documents

The screenshot shows the Apache Solr web interface. At the top, there is a search bar with the text "Find:" and a "[Go]" button. Below the search bar, it indicates "6 results found in 0 as Page 1 of 1".

On the left side, there are three panels: "Field Facets", "Query Facets", and "Range Facets". The "Field Facets" panel shows a list of facets with counts:

- Test (6)
- Ziggy (6)
- Alan (5)
- Crawler (5)
- Experimental (5)
- g (5)
- Labouseur (5)
- Page (5)
- The (5)
- Web (5)
- Asimov (1)
- B5 (1)
- Bond (1)
- Document (1)
- Rush (1)
- Star (1)
- Trek (1)


The main content area displays a list of search results. Each result includes a title, score, links, author, and keywords. The results are:

- Ziggy test document**
Score: 1.0
Links: rect "http://www.labouseur.com/projects/ziggy/test/trek" rect "http://www.labouseur.com/projects/ziggy/test/bond" rect "http://www.labouseur.com/projects/ziggy/test/asimov" rect "http://www.labouseur.com/projects/ziggy/test/rush.html" rect "http://www.labouseur.com/projects/ziggy/test/b5.html"
Author:
Keywords:
- Star Trek test page - Ziggy, the experimental web crawler - Alan G. Labouseur**
Score: 1.0
Links: rect "http://www.labouseur.com/projects/ziggy/test/b5.html"
Author:
Keywords:
- B5 test page - Ziggy, the experimental web crawler - Alan G. Labouseur**
Score: 1.0
Links: rect "http://www.labouseur.com/projects/ziggy/test/trek"
Author:
Keywords:
- Bond test page - Ziggy, the experimental web crawler - Alan G. Labouseur**
Score: 1.0
Links:
Author:
Keywords:
- Asimov test page - Ziggy, the experimental web crawler - Alan G. Labouseur**
Score: 1.0
Links:
Author:
Keywords:
- Rush test page - Ziggy, the experimental web crawler - Alan G. Labouseur**
Score: 1.0
Links:
Author:
Keywords:

At the bottom of the page, it says "6 results found, Page 1 of 1". There is also a footer with options like "enable_debug enable_annotation XML" and documentation links.

Documents matching “ripped off”

Lucene Solr Browser for Ziggy



Find: [Go]

2 results found in 52 ms Page 1 of 1

Field Facets

title

- [Alan](#) (2)
- [Crawler](#) (2)
- [Experimental](#) (2)
- [G](#) (2)
- [Labouseur](#) (2)
- [Page](#) (2)
- [Test](#) (2)
- [The](#) (2)
- [Web](#) (2)
- [Ziggy](#) (2)
- [B5](#) (1)
- [Star](#) (1)
- [Trek](#) (1)
- [Asimov](#) (0)
- [Bond](#) (0)

Star Trek test page ~ Ziggy, the experimental web crawler ~ Alan G. Labouseur

Score: 0.29930896

Links: rect \"http://www.labouseur.com/projects/ziggy/test/b5.html\"

Author:

Keywords:

B5 test page ~ Ziggy, the experimental web crawler ~ Alan G. Labouseur

Score: 0.29930896

Links: rect \"http://www.labouseur.com/projects/ziggy/test/trek\"

Author:

Keywords:

Under the covers of the first result

```
Star Trek test page ~ Ziggy, the experimental web crawler ~ Alan G. Labouseur
Score: 0.29930896
Links: rect \"http://www.labouseur.com/projects/ziggy/test/b5.html\"
Author:
Keywords:
toggle explain
0.29930896 = (MATCH) sum of:
  0.14965448 = (MATCH) weight(text:ripped in 1), product of:
    0.70710677 = queryWeight(text:ripped), product of:
      1.6931472 = idf(docFreq=2, maxDocs=6)
      0.41762865 = queryNorm
    0.2116434 = (MATCH) fieldWeight(text:ripped in 1), product of:
      1.0 = tf(termFreq(text:ripped)=1)
      1.6931472 = idf(docFreq=2, maxDocs=6)
      0.125 = fieldNorm(field=text, doc=1)
  0.14965448 = (MATCH) weight(text:off in 1), product of:
    0.70710677 = queryWeight(text:off), product of:
      1.6931472 = idf(docFreq=2, maxDocs=6)
      0.41762865 = queryNorm
    0.2116434 = (MATCH) fieldWeight(text:off in 1), product of:
      1.0 = tf(termFreq(text:off)=1)
      1.6931472 = idf(docFreq=2, maxDocs=6)
      0.125 = fieldNorm(field=text, doc=1)
```

Solr Admin (Ziggy [])

10.0.0.7:8983

cwd=/usr/bin/solr/example SolrHome=solr/./

HTTP caching is OFF

Schema Browser I See [RAW SCHEMA.XML](#)

| |
|----------------|
| HOME |
| FIELDS |
| BODY |
| TEXT_REV |
| TEXT |
| KEYWORDS |
| SUBJECT |
| LINKS |
| ID |
| AUTHOR |
| TITLE |
| CATEGORY |
| LAST_MODIFIED |
| DESCRIPTION |
| CONTENT_TYPE |
| COMMENTS |
| DYNAMIC FIELDS |
| FIELD TYPES |

Field: links

Field Type: **STRING**

Properties: Indexed, Stored, Multivalued, Omit Norms, undefined, Sort Missing Last

Schema: Indexed, Stored, Multivalued, Omit Norms, undefined, Sort Missing Last

Index: Indexed, Stored, Omit Norms, Lazy

Index Analyzer: org.apache.solr.schema.FieldType\$DefaultAnalyzer

Query Analyzer: org.apache.solr.schema.FieldType\$DefaultAnalyzer

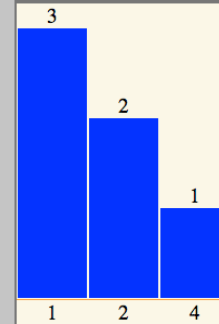
Docs: 3

Distinct: 6

Top Terms

| term | frequency |
|--|-----------|
| rect | 3 |
| "http://www.labouseur.com/projects/ziggy/test/trek" | 2 |
| "http://www.labouseur.com/projects/ziggy/test/b5.html" | 2 |
| "http://www.labouseur.com/projects/ziggy/test/bond" | 1 |
| "http://www.labouseur.com/projects/ziggy/test/rush.html" | 1 |
| "http://www.labouseur.com/projects/ziggy/test/asimov" | 1 |

Histogram



Configuration

Erlang

- Main site: <http://www.erlang.org/>
- Nicer documentation: <http://erldocs.com/>
- The best Q&A on Erlang: <http://stackoverflow.com/questions/tagged/erlang>
- Resources from The Blunt Professor: <http://www.labouseur.com/courses/erlang/>

Yaws

- Main site: <http://yaws.hyber.org/>
- The best Q&A on Yaws: <http://stackoverflow.com/questions/tagged/yaws>

Lucene

- Main site: <http://lucene.apache.org/java/docs/index.html>
- A cool book on Lucene: <http://www.manning.com/hatcher3/>
- The best Q&A on Lucene: <http://stackoverflow.com/questions/tagged/lucene>
- About Lucene's scoring: <http://www.lucentutorial.com/advanced-topics/scoring.html>

Solr

- Main site: <http://lucene.apache.org/solr/> (Jetty is included in the Solr download)
- A decent Solr tutorial: <http://lucene.apache.org/solr/tutorial.html>
- Getting started with Solr video: <http://www.lucidimagination.com/devzone/videos-podcasts/how-to/getting-started-solr-14-tutorial>
- The best Q&A on Solr: <http://stackoverflow.com/questions/tagged/solr>

Java

- In the unlikely event you do not have Java: <http://www.oracle.com/technetwork/java/javase/downloads/index.html>

Ziggy

- Web app: <http://www.labouseur.com/projects/ziggy/index.html>
- Server-side code: <http://www.labouseur.com/projects/ziggy/ziggy.yaws.txt>
- Web crawler code: <http://www.labouseur.com/projects/ziggy/ziggy3.erl.txt>
- Various CSS and HTML customizations to Solr embedded in Solr config.

Final Thoughts

While Ziggy does not play guitar he does crawl the web, and occasionally exhibits a repulsive need to be something more than human.

All your base are belong to Ziggy.