

Installation and Deployment Guide



1. Installation (Stand Alone)

Check the requisite softwares:

1. JDK 1.6, Apache Ant,
2. Oracle Berkeley DB 4.8 (A little tricky to install on linux, download from <http://www.oracle.com/technology/software/products/berkeley-db/db/index.html>), and follow the instructions
3. Cygwin (if Windows)
4. Download the latest Joycrawler from <http://code.google.com/p/joycrawler/>
5. Unzip it to any directory, and "cd" to that directory.

2. Run Examples (Stand Alone)

Each example job will be configured in two files, the seeds-*.txt in top dir contains all the crawl seeds for this job and configuration file joycrawler-*.xml which configures all the variables located in conf/. The * will be replaced by Example name, which you should use in the following command line.

To download the webpages and index them

```
$ ant index -Dconf=EXAMPLE -Dnative.path=YOUR_BERKELEY_DB_INSTALLATION_DIR
```

Note: for windows, you can simply specify native.path variable to ./lib/native.

```
C:\Users\LIUSONG\Desktop\joycrawler-0.20.0>ant index -Dconf=hadoopcn -Dnative.path=
./lib/native
```

To start your search server on local machine

```
$ ant server -Dnative.path=YOUR_BERKELEY_DB_INSTALLATION_DIR -Ddbfolder=DB_FOLDER
```

```
C:\Users\LIUSONG\Desktop\joycrawler-0.20.0>ant server -Dnative.path=
./lib/native
-Ddbfolder=...\db_hadoopcn
Buildfile: C:\Users\LIUSONG\Desktop\joycrawler-0.20.0\build.xml

compile:
  [javac] C:\Users\LIUSONG\Desktop\joycrawler-0.20.0\build.xml:129: warning: '
includeantruntime' was not set, defaulting to build.sysclasspath=last; set to fa
lse for repeatable builds

server:
  [java] Listening from 1987...
```

DB_FODLER is described in org.joy.index.DBFolder property in configuration file.

The initialization will finish if it prompts "Listening to 1987"

To start search, launch another terminal, cd to the installation directory

\$ ant searcher -Dhosts=localhost

Then you can type the search string "keywords":"pageNo(from 0)"

for example, *map reduce:0*

```
C:\Users\LIUSONG\Desktop\joycrawler-0.20.0>ant searcher -Dhosts=localhost
Buildfile: C:\Users\LIUSONG\Desktop\joycrawler-0.20.0\build.xml

compile:
  [javac] C:\Users\LIUSONG\Desktop\joycrawler-0.20.0\build.xml:129: warning: '
includeantruntime' was not set, defaulting to build.sysclasspath=last; set to fa
lse for repeatable builds

searcher:
map reducer:0
  [java] http://hadoop.apache.org/common/docs/r0.20.1/cn/mapred_tutorial.html
        Hadoop Map Reduce教程 0.003571061242837459 Hadoop Map Reduce教程Apa
che >Hado
```

3. Custmize your search

Try to modify the configuration file and seeds file on your demand, and name it as example's, then repeat the steps in section 2.

For example, if you want to crawl and index "apache.org"

First, create a file named "seeds-apache.txt" contains following line:

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

Second, create a configuration file contains following content named "joycrawler-apache.xml" in conf/ folder

```
<configuration>
<property>
  <name>org.joy.crawler.regEx</name>
  <value>http://.*.apache.org/.*</value>
  <description>
    the regular expression that all the URL we crawled must match.
    To add multiple value for this property, please use correct regular expression.
    http://(. *youth.suda.edu.cn.*|. *www.suda.edu.cn.*), eg.
  </description>
</property>
</configuration>
```

Then adopt the codes provided in section 2 to perform your index and search.