WhiteListRobots

White List for Robots.txt

Nutch now has a white list for robots.txt capability that can be used to selectively on a per host and/or IP basis turn on/off robots.txt parsing. Read on to find out how to use it.

List hostnames and/or IP addresses in Nutch conf

In the Nutch configuration directory (conf/), edit nutch-default.xml (and/or nutch-site.xml) and add the following information:

```
<property>
  <name>http.robot.rules.whitelist</name>
  <value></value>
  <description>Comma separated list of hostnames or IP addresses to ignore robot rules parsing for.
  </description>
</property>
```

For example, try this, to whitelist the host, baron.pagemewhen.com:

```
<property>
  <name>http.robot.rules.whitelist</name>
  <value>baron.pagemewhen.com</value>
  <description>Comma separated list of hostnames or IP addresses to ignore robot rules parsing for.
  </description>
</property>
```

Testing the configuration

Create a sample URLs file to test your whitelist. For example, create a file, call it "url" (without the quotes) and store each URL on a line:

```
http://baron.pagemewhen.com/~chris/foo1.txt
http://baron.pagemewhen.com/~chris/
```

Create a sample robots.txt file, e.g., "robots.txt" (without the quotes):

```
User-agent: *
Disallow: /
```

Build the Nutch runtime and execute RobotRulesParser

Now, build the Nutch runtime, e.g., by running {{`}}ant runtime{{`}}. From your nutch SVN or git checkout top-level directory, run this command:

```
java -cp build/apache-nutch-1.11-SNAPSHOT.job:build/apache-nutch-1.11-SNAPSHOT.jar:runtime/local/lib/hadoop-
core-1.2.0.jar:runtime/local/lib/crawler-commons-0.6.jar:runtime/local/lib/slf4j-log4j12-1.7.5.jar:runtime/local
/lib/slf4j-api-1.7.9.jar:runtime/local/lib/log4j-1.2.17.jar:runtime/local/lib/guava-16.0.1.jar:runtime/local/lib
/commons-logging-1.1.3.jar:runtime/local/lib/commons-cli-1.2.jar org.apache.nutch.protocol.RobotRulesParser
robots.txt url Nutch-crawler
```

You should see the following output:

```
Whitelisted hosts: [baron.pagemewhen.com]
Whitelisted hosts: [baron.pagemewhen.com]
Whitelisted hosts: [baron.pagemewhen.com]
whitelisted:        http://baron.pagemewhen.com/~chris/foo1.txt
whitelisted:        http://baron.pagemewhen.com/~chris/
```