Nutch robot

Table of contents

If you're reading this, chances are you've seen a Nutch-based robot visiting your site while looking through your server logs. Our software obeys robots.txt files and robot META tags in HTML. These are the standard mechanisms for webmasters to tell web robots which portions of a site a robot is welcome to access.

1. Sysadmins/robots.txt

We're a software project, not a service, so please understand that a misbehaving crawler appearing with our Agent string is not run by us. Our software may be run by anyone. However, we'd still like to hear about any bad behavior. If possible, please include the name of the domain and some representative log entries. We can be reached at nutch-agent@lucene.apache.org.

Our software obeys the robots.txt exclusion standard, described at http://www.robotstxt.org/wc/exclusion.html#robotstxt. Different installations of the Nutch software may specify different agent names, but all should respond to the agent name "Nutch". Thus to ban all Nutch-based crawlers from your site, place the following in your robots.txt file:

```
User-agent: Nutch
Disallow: /
```

2. Webmasters/Robots META

If you do not have permission to edit the /robots.txt file on your server, you can still tell robots not to index your pages or follow your links. The standard mechanism for this is the robots META tag, as described at http://www.robotstxt.org/wc/meta-user.html.

3. Contact us

If your site has problems or questions about the Nutch crawler, please send an email to the Nutch agent mailing list.