

一条鱼 + 一条鱼

http://zhaohaolin.iteye.com



zhaohaolin

浏览: 90456 次

性别:

来自: 杭州

我现在离线

详细资料 留言簿

搜索本博客

最近访客 [>>更多访客](#)

- eggh
- renfuijiang
- Dping
- huapuyu

- 博客分类
- 全部博客 (749)
  - 硬件 (6)
  - 软件 (18)
  - 软件工程 (33)
  - JAVA (207)
  - C/C++/C# (76)
  - JavaScript (8)
  - PHP (1)
  - Ruby (3)
  - MySQL (11)
  - 数据库 (18)
  - 心情记事 (11)
  - 团队管理 (18)
  - Hadoop (1)

2011-05-09

◀ [Heritrix3.0教程 使用入门\(二\) 开始抓取](#) | [heritrix配置篇](#) ▶

[Heritrix3.0教程 使用入门\(一\) 下载安装与运行](#)

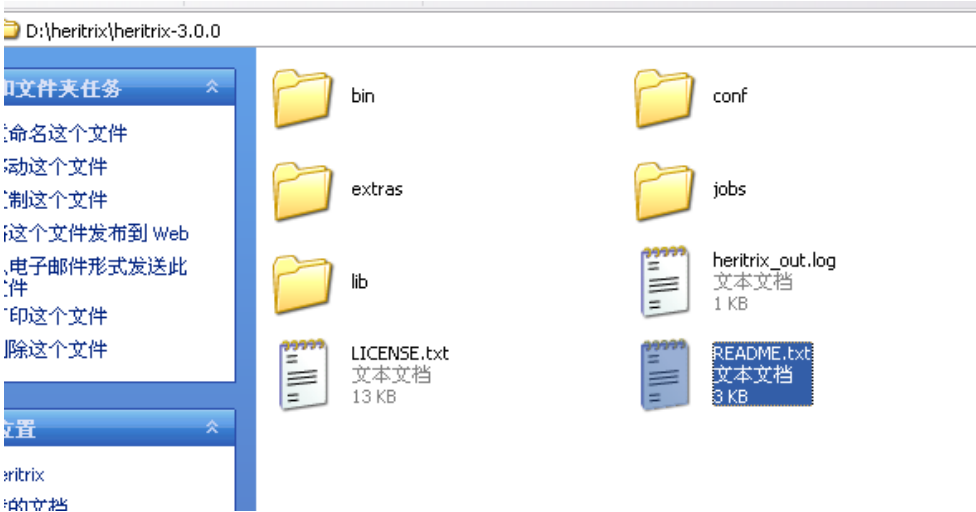
博客分类: [Heritrix](#)

[IE](#) [浏览器](#) [Web](#) [.net](#)

本博客属原创文章,转载请注明出处:<http://www.yun5u.com/articles/heritrix3-1.html>

Heritrix3.0.0在2009年底发布,但资料甚少.我这里就先抛砖引用,以前也分析过Heritrix1.4.3,但只是源码,不系统.这里就系统的介绍Heritrix的使用,源码分析和借鉴.先介绍Heritrix的下载与使用吧.

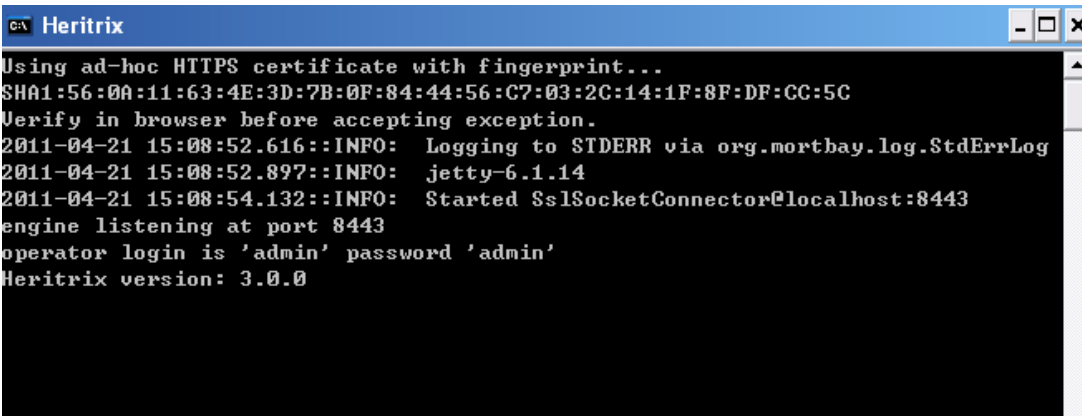
1. 下载,下载地址:<http://sourceforge.net/projects/archive-crawler/files/heritrix3/>. 下载后的截图为



这里大家可以看下README.TXT文件.这里面有对Heritrix基本的介绍.

2. 下面开始使用Heritrix3.0.0

进入CMD(开始->运行),进入Heritrix3.0.0所在目录,我这里是D:/heritrix/heritrix3.0.0/bin,这里大家截图也可以看到.输入以下命令:heritrix -a admin:admin,这里会跳出一个cmd,截图如下:




就表示你已经启动Heririx成功,然后在浏览器里输入,https://localhost:8443(注意,是https,不是http). 由于Heritrix3.0.0已通过https登录,用户名跟密码就是以上输入的admin:admin.所以不同于早期版本,我这里用的是火狐浏览器,界面可能如下

- [spring \(22\)](#)
- [mybatis\(ibatis\) \(7\)](#)
- [tomcat \(16\)](#)
- [velocity \(0\)](#)
- [系统架构 \(5\)](#)
- [JMX \(8\)](#)
- [proxool \(1\)](#)
- [开发工具 \(14\)](#)
- [python \(10\)](#)
- [JVM \(27\)](#)
- [servlet \(5\)](#)
- [JMS \(26\)](#)
- [ant \(2\)](#)
- [设计模式 \(5\)](#)
- [智力题 \(2\)](#)
- [面试题收集 \(1\)](#)
- [孙子兵法 \(16\)](#)
- [测试 \(1\)](#)
- [数据结构 \(5\)](#)
- [算法 \(19\)](#)
- [Android \(11\)](#)
- [汽车驾驶 \(1\)](#)
- [lucene \(1\)](#)
- [memcache \(10\)](#)
- [技术架构 \(5\)](#)
- [OTP-Erlang \(7\)](#)
- [memcached \(15\)](#)
- [redis \(20\)](#)
- [浏览器插件 \(3\)](#)
- [sqlite \(3\)](#)
- [Heritrix \(9\)](#)
- [Java线程 \(1\)](#)
- [scala \(0\)](#)
- [Mina \(2\)](#)
- [汇编 \(2\)](#)
- [Netty \(14\)](#)

我的相册





### 此连接是不受信任的

您想使用 Firefox 来安全连接至 **localhost:8443**，但是我们无法确认此连接为安全的。

通常，当您尝试安全连接时，站点会出示受信任的标识，以证明您访问的是正确的地址。然而，现在无法验证此网站的标识。

#### 怎么办？

如果您过去连接到此网站并且没有发现问题，那么此错误信息表示可能有人想冒充该网站，所以您应该停止浏览。

立即离开！

▶ 技术细节

▶ 我已充分了解可能的风险

点此进入

ie等可能不一样.然后点击我已充分了解可能的风险,点添加例外,再输入用户名跟密码,也就是刚才的admin,admin后,便可以进入Heritrix3.0.0 web界面了.大概如下:

### Heritrix Engine 3.0.0

Memory: 2722 KiB used; 5056 KiB current heap; 260160 KiB max heap

Jobs Directory: [D:\heritrix\heritrix-3.0.0\bin\jobs](#)

Job Directories (0) 

rescan

Add Job Directory

Create new job directory with recommended starting configuration

Path: D:\heritrix\heritrix-3.0.0\bin\jobs/ 

create

Specify a path to a pre-existing job directory

Path: 

add

You may also compose or copy a valid job directory into the main jobs directory via outside means, then use the 'rescan' button above to make it appear in this interface. Or, use the 'copy' functionality at the bottom of any existing job's detail page.

出现以上界面,就表示你可以使用Heritrix去抓取数据了,但这里还需配置Job,也就是抓取任务.

这里先大概介绍下界面:

1. Memory 表示内存使用情况
2. Jobs Directory :表示抓取job目录,默认是Heritrix\_home/jobs
3. rescan按钮表示扫描jobs目录,目录有改动,也就是抓取任务有增加或者删除,这里则都会显示
4. create按钮表示创建一个Job
5. add按钮表示添加一个已经存在的job,这里是输入这个job所在的路径

好了,这里基本上可以下载并使用Heririx了.下一篇则介绍如何配置CrawlJob,也就是抓取任务去抓取数据.

[室内设计培训,免费推荐就业](#)  
自力学院室内设计培训 专业/权威/速成 被评为国内A级室内设计培训示范学院  
[www.zili.cn](#)

Google 提供的广告

分享到:



◀ [Heritrix3.0教程 使用入门\(二\) 开始抓取](#) | [heritrix配置篇](#) ▶

23:11 | [评论 / 浏览 \(0 / 219\)](#) | 分类:[编程语言](#) | [相关推荐](#)

▶ MORE

评论

发表评论



B7211164594159

[共 11 张](#)

我的留言簿 [>>更多留言](#)

- [zhaohaolin](#) 写道步青龙 写道 你好 你好  
-- by [步青龙](#)
- 声明，本站所有资料除非特别标明是原创外，其它文件均来自网友的转载，如有侵犯你的版权 ...  
-- by [zhaohaolin](#)
- 你好  
-- by [步青龙](#)

其他分类

- [我的收藏](#) (245)
- [我的代码](#) (0)
- [我的论坛主题帖](#) (0)
- [我的所有论坛帖](#) (5)
- [我的精华良好帖](#) (0)

最近加入群组

- [系统架构与架构应用](#)
- [Hadoop](#)
- [高级语言虚拟机](#)
- [Hibernate](#)
- [JBPM @net](#)

存档

- [2011-10](#) (19)
- [2011-09](#) (34)
- [2011-08](#) (37)
- [更多存档...](#)

评论排行榜

- [持续集成之路——Maven（续）](#)
- [jedis线程池的代码【转】](#)
- [c++线程池](#)
- [驯服爬虫 Heritrix](#)
- [Java深度历险（二）——Java类的加载、链接...](#)



[您还没有登录,请您登录后再发表评论](#)



---

声明：ITeye文章版权属于作者，受法律保护。没有作者书面许可不得转载。若作者同意转载，必须以超链接形式标明文章原始出处和作者。  
© 2003-2011 ITeye.com. All rights reserved. [ 京ICP证110151号 京公网安备110105010620 ]