**基金爬虫文档**

**2015年10月20日**

## 介绍

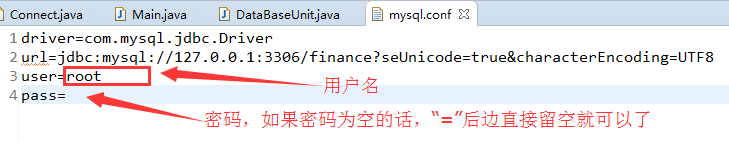
此爬虫属于垂直型爬虫，用于爬取《中国银河证券基金评价业务》的基金业绩评价日报表中的数据，数据的网址为： <http://fund.chinastock.com.cn/jjpj/index.jsp>

爬取的方式是根据用户的需要爬取某一天，或者某一个连续时间段内的基金的数据，爬取后的数据储存在Mysql数据库；

爬取的数据项包括：基金代码、基金分类、基金名称、机构、基金成立日期、基金价格、日期；

## 数据库配置

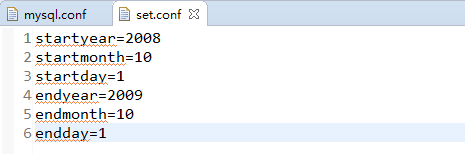
数据库的配置是在mysql.conf文件中配置的，用户在使用的时候，修改用户名、密码即可，其余的一般情况下不需要做任何修改，详情见下图：



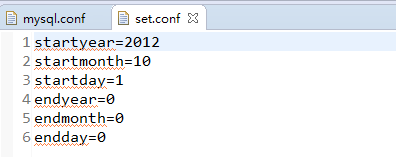
## 爬取范围配置

此网站基金的数据是有限的，我们爬取数据的时候是爬取指定日期或日期范围的数据，该配置项的配置文件为set.conf

爬取连续时间段的数据：如果用户需要爬取某一连续时间段的数据的时候，可使用此方法，startyear、startmonth、startday后边配置的是爬虫爬取的起始的年、月、日，endyear、endmonth、endday后边配置的是爬虫爬取的终止的年、月、日。如果要爬取2008年10月1日——2009年10月1日的数据的话，按照以下配置即可；



爬取指定日期的数据：如果用户需要爬取指定日期的数据的话，配置要爬取的日期到startyear、startmonth、startday即可，endyear、endmonth、endday则需要配置为0，即：如果要爬取2012年10月1日的数据，配置如下：



## 使用说明

此程序用java语言编写，可以运行在安装有java环境的机器上即可；

* JDK 1.6及其以上版本
* Mysql 5.0及其以上版本

下面介绍在linux下程序的运行：

1. 通过已生成的sql文件，建立数据库，数据表，存储过程等；
2. 使用源码打包为jar文件；
3. 上传打包后的项目到linux服务器；
4. 使用“nohup java -jar fund.jar &”命令运行程序（fund.jar为我打包后的程序名称）；

## 日志

程序运行的日志文件再log目录下，日志的文件名为“log”+爬虫运行的开始时间