中国移动采购和招标网数据爬取程序——开发说明

**1.程序需求描述**

爬取中国移动采购和招标网的数据，并存储到hbase上。

<https://b2b.10086.cn/b2b/main/listVendorNotice.html?noticeType=2#this>

1. 程序开发环境

1>. maven3.3.9项目，使用eclipse开发。

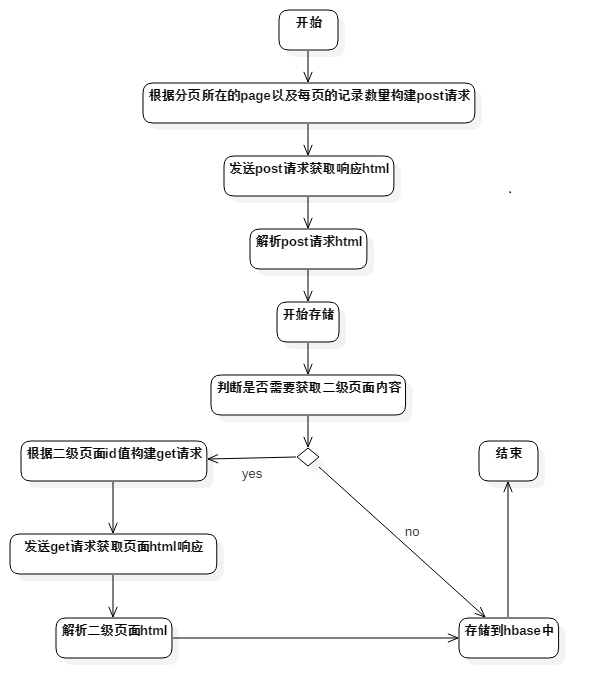
2>. hbase版本为——1.2.0-cdh5.9.0

zookeeper版本——3.4.5-cdh5.9.0

jdk 1.7以上

log4j相关版本

1. 程序执行流程



描述：

对于不同的noticetype（公告类型）,程序分开执行，在程序执行开始传入noticetype类型开启关于该类型公告的爬取过程。

POST请求是采用每一个分页一个线程的方式进行请求的，然后开启解析该分页上面的html线程，获取每一个record，如果需要获取record对应的二级页面内容，则调用get请求，获取相应的html内容（get部分未使用线程）,最后整合进行存储。

1. 程序主要功能：
2. . 模拟POST请求响应
3. . 模拟GET请求响应
4. . 解析响应html文档
5. . 存储到hbase上
6. 源码地址：

<http://192.168.9.111/gitbucket/chenxiang/ViewNoticeContent>