

* 航空航天领域新闻推荐功能使用说明文档

简要介绍：该程序基于需求——能够周期性地自动爬取预设的网站（如国内主要新闻门户、各大高校的官网等）的内容，从中筛选出内容与航空航天领域相关且包含了预设关键字（如复合材料、表面工程、重金属技术、结构工程、电子传感器、生物燃料、安全服务、新能源、仿真、推进力等）的网址。然后将这些符合要求的网址一方面能够自动通过邮件推送到邮箱中，另一方面也会讲网址保存到数据库中，以便以后能够通过日期查询。要注意的是，该程序目前虽然能单独运行，但是当初也是基于后端来设计和实现的，所以最后还是需要前端来协同完成整个程序。否则，像输入起始、终止日期来查询对应时间段内的网址功能是没法通过界面方式进行，只能通过修改程序完成。

程序组成：使用Heritrix 3.0 + Lucene 4.7.2 + Htmlparser 2.0 + MySQL,再基于需求定制功能。一、Heritrix 作为开源爬虫项目，能够指定特定网站，爬取里面的网页。二、通过Htmlparser解析所爬取网页的内容，以待下一步分析。三、最后通过Lucene来对网页内容进行文本分析，判断网页与预设的关键字内容是否匹配，最后筛选出符合要求的网页，推送到邮箱和保存到数据库中。四、最后能够定时从数据库中读取数据通过邮件推送到邮箱中。

程序使用：

1. 程序的主入口点为包com.airbus.delegate下的Delegation.java文件的main函数。前端也可以按自己逻辑来调用Delegation.java文件内的方法来完成相关功能。可以修改变量PERIOD（单位为毫秒）来控制推送间隔。调用其中的方法public void sendEmailByDate(String startDate,String endDate)，就能够发送startDate～endDate时间段内的符合要求的网址到接收邮箱中。前端可调用该方法来完成按日期推送功能。用法如下：Delegation d = new Delegation();

d.senEmailByDate(“2014-09-20”,”2014-09-30”);

1. 因为目前没有前端，所以要预设指定网站和关键字只能通过修改seed.txt和keywords.txt来完成。其中keywords.txt就在工程的根目录下，格式为每行一个关键字。而seed.txt路径为的./jobs/CrawlHTML/seed.txt，也就是在工程的jobs文件中的CrawlHTML文件内，格式为每行一个预设搜索网站的网址。
2. 关于设置发送和接收邮箱的操作，在包com.airbus.timer下的NFDFlightDataTimerTask的run()方法里面可以设置发送邮箱、密码，邮件主题，邮件内容还有接收人邮箱。具体可以看代码的注释。
3. 数据库方面使用了MySQL并已经安装了Navicat for mysql可以以图形化界面来管理所存储的数据。数据库连接操作都封装在com.airbus.jdbc包下的Constants.java和DataBaseUtil中，其中可以修改Constants.java来更改数据库的连接，注释有相关说明。
4. Delegation.java的checkInputDate()方法作用是模拟在输入栏输入日期返回相应的新闻，如果工程根目录下的dates.txt里面包含了起始日期和结束日期则会查找数据库并推送邮件。用户使用说明：dates.txt在工程的根目录下，第一行为起始日期，第二行为终止日期，日期格式为xxxx-xx-xx，分别为年月日。

最后：如果程序在运行中出现异常或其他问题，或想重新更改搜索日期推送邮件都可以重新启动。该工程的源代码已经托管到GitHub，链接为：https://github.com/berryjam/NewsRecommendation.git