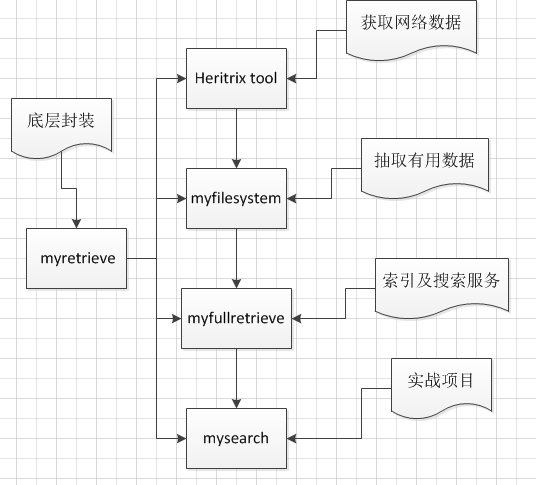
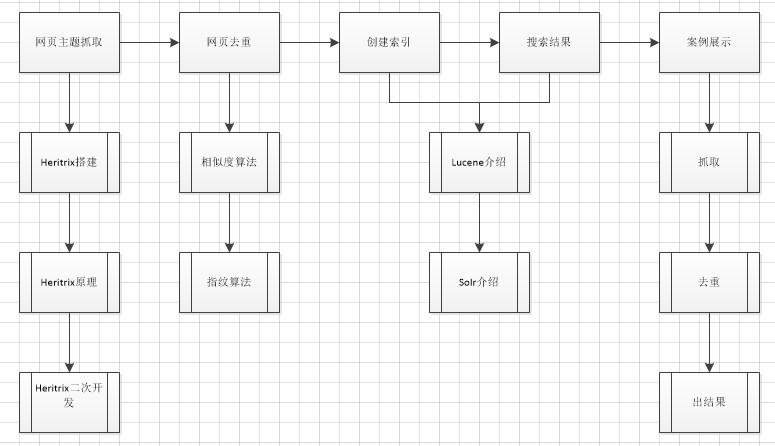
垂直搜索引擎是针对某一行业的专业搜索引擎，是搜索引擎的细化和延伸，是对网页库内的某类专门的信息进行一次整合，定向分字段抽取出需要的数据进行处理后再以某种形式返回给用户。





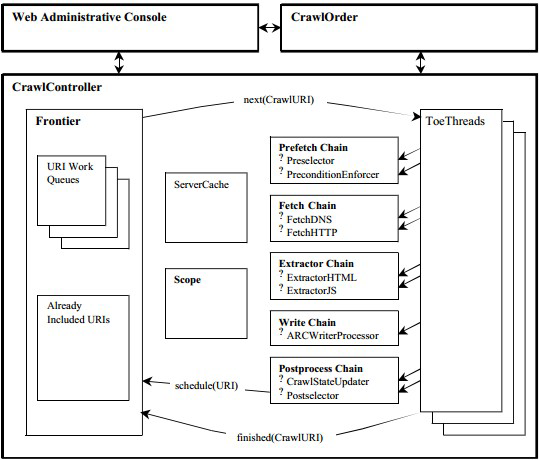
# 搭建heritrix

## 什么是网络爬虫

网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列，直到满足系统的一定停止条件。

网络爬虫能做什么

## .Heritrix原理



## Heritrix搭建

1）将 heritrix-14-src\src\java 下的 com、org 和 st 三个文件夹拷贝进 heritrix 工程的 src 下。

2）将 heritrix-14-src\src\resources\org\archive\util 下的文件 tlds-alpha-by-domain.txt 拷贝到 heritrix\src\org\archive\util 中。

3）将 heritrix-14-src\src 下 conf 文件夹拷贝至 Heritrix 工程根目录。

4）将 heritrix-14-src\src 中的 webapps 文件夹拷贝至 Heritrix 工程根目录。

# 如何进行主题抓取

## 什么是主题抓取

主题网络爬虫是指尽可能快地爬行、采集尽可能多的与预先定义好的主题相关的网页。主题网络爬虫可以通过对整个 Web 按 主题分块采 集, 并将 不 同 块的 采 集 结果 整 合 到一 起, 以提 高 整 个Web 的采集覆盖率和页面利用率。

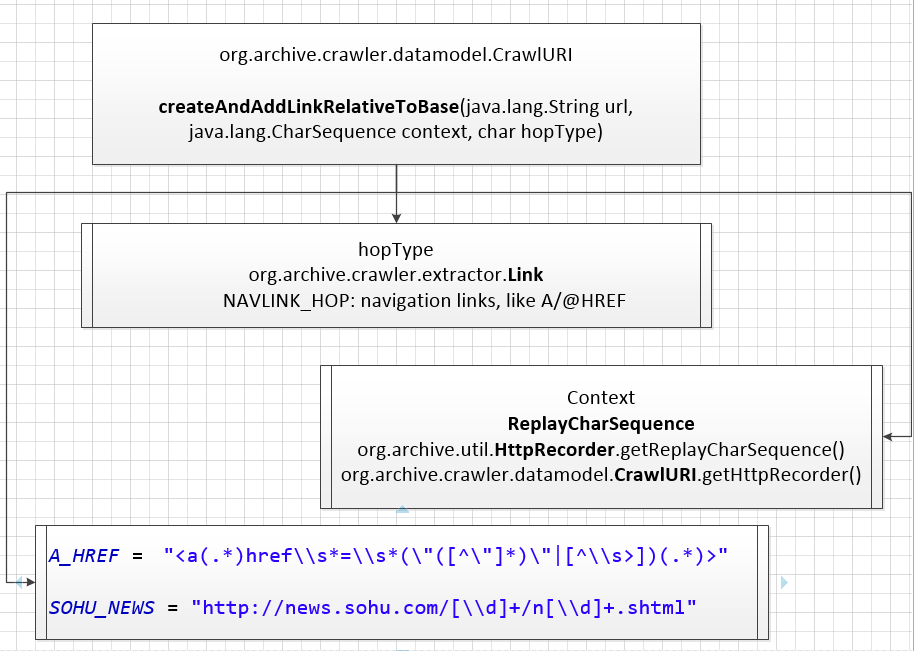
## 主题抓取的意义

随着网络上海量信息的爆炸式增长, 通用搜索引擎面临着索引规模、更新速度和个性化需 求等多 方面的 挑战。面 对这些挑战, 适 应特 定 主 题和 个 性 化搜 索 的 主题 网 络 爬 虫( focused crawler or topical crawler) 应运 而 生。

## 主题抓取的策略

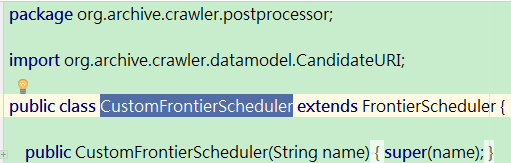
一、基于链接

二、基于内容 \*\*

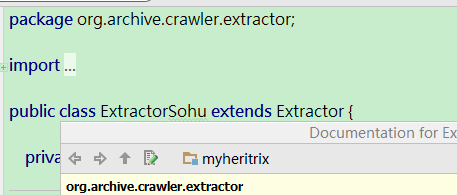


## 如何用heritrix进行主题抓取

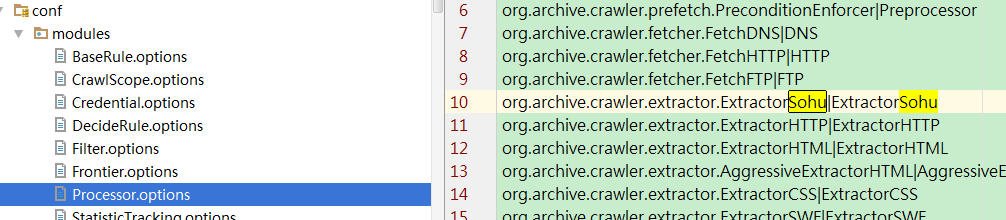
1. 扩展FrontierScheduler



二、扩展Extractor



以上两个类添加到配置文件中：



# heritrix优化





ELFHashQueueAssignmentPolicy

然后在hetritrix.properties 中替换

## ELFHash算法

在默认的情况下，Heritrix使用HostnameQueueAssignmentPolicy来产生key值，而这个策略是用hostname作为key值的，因此一个域名下的所有链接都会被放到同一个线程中去。如果对Heritrix分配URI时的策略进行改进，利用ELFHash“可执行链接格式” ( Executable and Linking Format，即ELF )算法把url尽量平均分部到各个队列中去，就能够用较多的线程同时抓取一个域名下的网页，速度将得到大大的提高。

public static long ELFHash(String str) {

long hash = 0;

long x = 0;

for (int i = 0; i < str.length(); i++) {

hash = (hash << 4) + str.charAt(i);

if ((x = hash & 0xF0000000L) != 0) {

hash ^= (x >> 24);

hash &= ~x;

}

}

return (hash & 0x7FFFFFFF);

}

## 关于robot.txt

org.archive.crawler.prefetch.PreconditionEnforcer

**private boolean considerRobotsPreconditions(CrawlURI curi){**

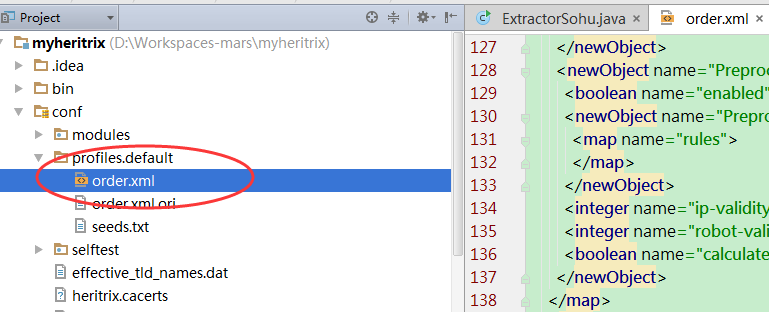
**return false;**

**}**

不去访问robots.txt 文件。

将heritrix打包成工具

用jobs 中生成的order.xml 替换默订的order.xml 文件



# 解析html页面

java正则表达式

基于模板获取网页内容

利用htmlparser解析html

# 中文分词介绍

Lucene自带的分词

ICTCLAS

IK

## 领域词识别

**六、网页去重**

网页去重的意义

网页去重的主要方法

什么是tf\*idf

基于指纹算法的网页去重

**七、Lucene6快速索引与搜索**

       如何用lucene创建索引

      如何用lucene搜索结果

      Lucene中intfield怎么搜索

       Lucene的结果高亮显示

**八、Lucene6索引的相关操作**

      创建索引

      修改索引

    删除索引

      索引优化

**九、Lucene6的query**

TermQuery

       BooleanQuery

各种query… …

**十、Lucene的Filter及自定义排序**

       Filter

      Lucene自带排序及指定权重

      Lucene自定义排序

**十一、Solr快速索引与搜索**

      什么是solr

       为什么工程中要使用solr

      Solr的原理

       如何在tomcat中运行solr

      利用solr进行索引与搜索

**十二、Solr的查询及Filter**

      solr的各种查询

      solr的Filter

      solr的排序

       solr的高亮

**十三、Solr的facet介绍**

  solr的某个域统计

   solr的范围统计

**十四、Solrcloud集群搭建**

   zookeeper简介

   solrcloud集群搭建

**十五、搜索服务的工具封装**

    工厂模式

    封装搜索服务\_lucene

    封装搜索服务\_solr

封装工具，支持配置

**十六、项目实战**

    项目需求分析及框架选择

    Struts 16介绍

    整合spring 0.1

    整合hibernate 1

    jquery-easyui 5

    6.heritrix 在工程中的运用

    7.搜索框架在工程中的运用

    8.Flexpaper模仿百度文库

    9.文件上传

    10.相关代码编写

    1搜索结果优化

    1项目总结