开发文档

## 背景

传统的搜索引擎搜出来的东西很多都是没什么用的，还要在一大堆搜索的结果里找到自己想要的资料，这样很费事。而垂直搜索能够针对某一主题进行搜索，搜索出来的信息一般是与主题相关的，这样能够帮助用户更好的找到所要的资料信息。

## 需求分析：

对股票相关数据进行挖掘，提取出有价值的信息。建立索引库，进行搜索。

搜索主要分股票资讯搜索；个股搜索，对股票代码进行搜索；全文检索

## 功能模块：

一：抓取模块：

二：搜索模块：

.html文件

使用heritrix抓取

文本数据

解析数据

建立索引并存入数据库

索引库/数据库

Web页面：

输入关键字

检索索引库、数据库

关键字

结果页面

## 数据字典：

StockNews(股票资讯表)

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 说明 |
| id | Long | 唯一标识 |
| title | Varchar(30) | 文章标题 |
| author | Varchar(20) | 作者 |
| pubdate | Date | 发布日期 |
| url | Varchar(80) | 来源网址 |
| stockContent | Text | 股票文章内容 |
| stockFrom | Varchar(80)` | 来源 |
| cacheContent | Text | 缓存，用来存放网页源码文本 |

## 代码规范

1. 分页代码规范：

|  |  |
| --- | --- |
| pageCurrentNum | 当前页 |
| pageTotalNum | 总共页数 |
| pageNextNum | 下一页 |
| pagePreNum | 上一页 |
| pageTotalHits | 总条数（总共有多少记录） |
| pageItemNum | 每页显示几条数据 |

1. indexDao中搜索的方法中传入关键字searchkeyword ，int start ，int end；
2. LuceneUtils类需重新设计为一个单例 主要功能是初始化索引库目录和分词器

**public** **static** LuceneUtils getInstance()

{

**if**(*instance*==**null**)

{

*instance* = **new** LuceneUtils();

}

**return** *instance*;

}

1. 搜索类型：0表示股讯搜索

1表示个股搜索

2表示全文走私

## 功能实现

开发环境：

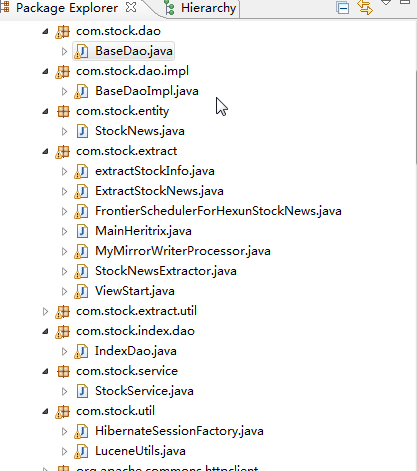
1. 抓取模块：使用heritrix1.4.4，扩展相关模块进行定向抓取数据存到本地磁盘
2. 搜索模块：使用struts2.3.x,hibernate3.6开源框架，jdk6.25,tomcat6.x,myeclipse8.5
3. 数据库sqlserver2005

功能：使用heritrix通过定制规则定向抓取网页，目前暂时抓取<http://stock.hexun.com>上的股讯信息，接下去抓取新浪博客揭幕者。并解析数据进行分词，建立索引库，相关数据存入数据库(暂时未存数据库)。股票搜索引擎可以根据股票代码搜索相关信息，全文检索信息，要实现像百度搜索结果页面一样，分页，权重、因子。

目前以实现功能：能够抓取<http://stock.hexun.com>下的股讯信息进行解析建立索引库，

web框架已基本搭建完成使用，能够实现全文检索，为实现分页。

下图是在heritrix中的结构



这里说明下com.stock.extract里的类

extractStockInfo类是对股票资讯网页的解析，暂时没用

ExtractStockNews类是对股票资讯网页的解析，解析过程全部放在同一个方法里，返回一个实体StockNews类。

FrontierSchedulerForHexunStockNews类继承heritrix的FrontierScheduler类，该类主要用类选择该网页中需要抓取的uri，并放入候选队列中。

MainHeritrix类是启动类，不是图像化启动。

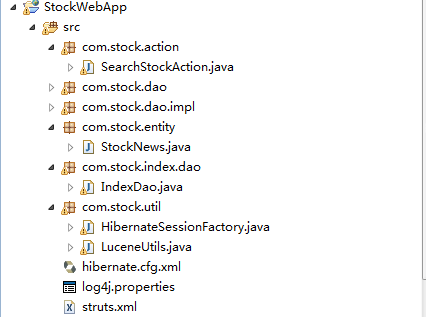
MyMirrorWriterProcessor类是用来将网页中的解析完所得的数据写入.txt中，并写到磁盘上

ViewStart类是图像化启动，暂时放着。计划是用该类来管理heritrix，而不是用web页面

com.stock.index.dao包中的IndexDao类是用来实现创建索引用的。

Com.stock.util包中的LuceneUtils用来初始化索引库目录，源文件目录，遍历文件。建立文档

Project StockWebApp



任务：抓取揭幕者信息：<http://blog.sina.com.cn/jiemuxiake>下的文章

正在实现的功能：分页，权重设计、排序

已实现功能：基本分页功能已经实现。

新增业务层：

### 搜索序列图



## 总结