# Lucene框架

## Lucene简介

### 什么是lucene？

Lucene是一个apache公司下的一个开源的全文检索引擎工具包。简单说，lucene可以实现全文检索功能，lucene是一个工具包。（jar包---不能单独使用）

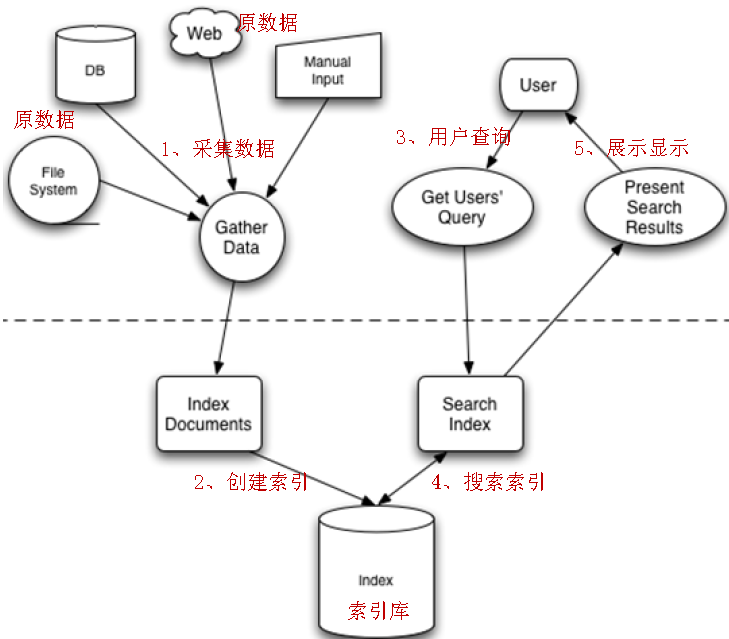
全文检索就是把要搜索的目标文档中的词提取出来，然后根据词建立索引，最终根据搜索索引找到目标文档的目的。

所以全文检索的过程：建立索引、搜索索引两个过程。

### Lucene的应用场景：搜索引擎和站内搜索（常用）

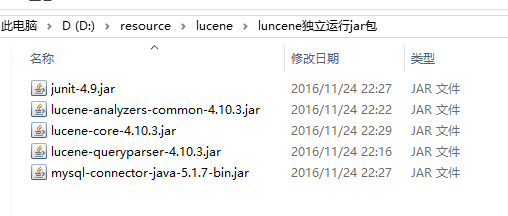
需求：使用lucene实现站内搜索，搜索站内图书信息。数据在关系型数据库中。

### Lucene架构图



## 如何使用

第一步：导入jar包



第二步：建立索引（IndexWriter）[*Ps:之前我们需要搭建一个采集数据的dao*](#_采集数据)

建议索引，其实就是一个采集数据的过程，需要使用[爬虫项目](#_常见爬虫项目)

由于数据资源多种多样，首先进行数据标准化处理，使用document对象封装记录。然后，使用IndexWriter的addocument()方法建立索引。

*Field属性：是否分词、是否索引、是否存储*

[*Ps:可以通过luke工具查看建立的索引*](#_使用luke工具查看索引数据)

第三步：搜索索引（IndexSearcher）



索引

## Ps：附件部分

### 常见爬虫项目

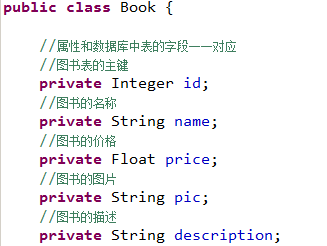
|  |  |  |
| --- | --- | --- |
| 名称 | 下载网址 | 介绍 |
| Solr | <http://lucene.apache.org/solr> | solr是apache的一个子项目，支持从关系数据库、xml文档中提取原始数据。 |
| Nutch | <http://lucene.apache.org/nutch> | Nutch是apache的一个子项目，包括大规模爬虫工具，能够抓取和分辨web网站数据。 |
| jsoup | http://jsoup.org/ | jsoup 是一款Java 的HTML解析器，可直接解析某个URL地址、HTML文本内容。它提供了一套非常省力的API，可通过DOM，CSS以及类似于jQuery的操作方法来取出和操作数据。 |
| heritrix | http://sourceforge.net/projects/archive-crawler/files/ | Heritrix 是一个由 java 开发的、开源的网络爬虫，用户可以使用它来从网上抓取想要的资源。其最出色之处在于它良好的可扩展性，方便用户实现自己的抓取逻辑。 |
|  |  |  |

### 采集数据

针对本例子的需求，从关系数据库中采集 book表的数据。

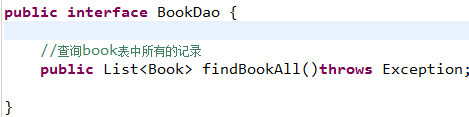
创建dao接口读取book表中的数据。

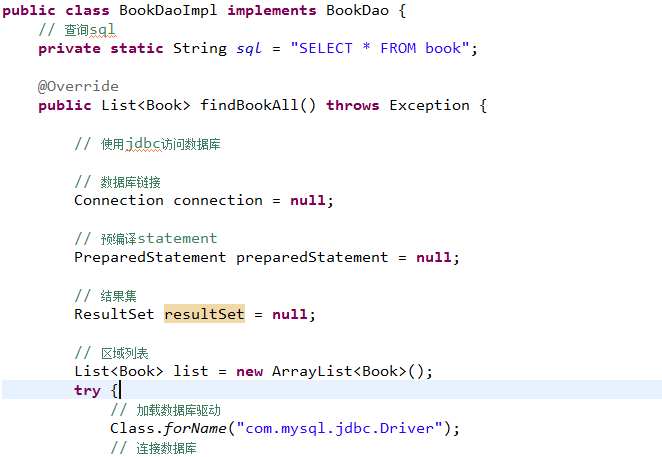
#### po实体类





#### Dao接口





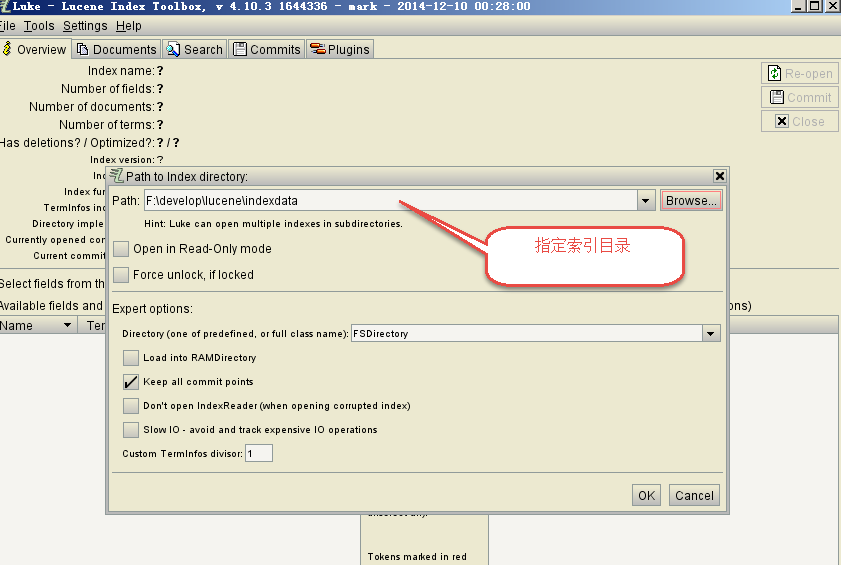
### 使用luke工具查看索引数据

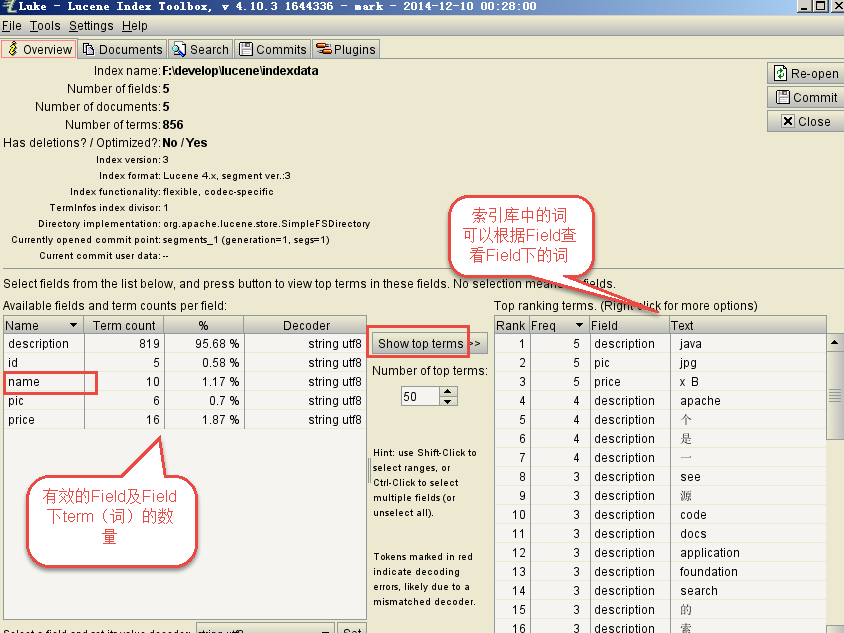
Luke作为Lucene工具包中的一个工具（http://www.getopt.org/luke/），用于查询、修改lucene的索引文件。

打开Luke方法：

cd luke所在目录







### Field分类:



## 自我总结

