**Lucene**是一套用于[全文检索](https://baike.baidu.com/item/%E5%85%A8%E6%96%87%E6%A3%80%E7%B4%A2)开源库

**什么是全文检索**

全文检索是指计算机索引程序通过扫描文章中的每一个词，对每一个词建立一个索引，指明该词在文章中出现的次数和位置，当用户查询时，检索程序就根据事先建立的索引进行查找，并将查找的结果反馈给用户的检索方式。这个过程类似于通过字典中的检索字表查字的过程。

**对比一下Lucene的检索方式相比以前的检索方式有哪些优点**

为了解决数据库压力和速度的问题，我们的数据库就变成了索引库，我们使用Lucene的API的来操作服务器上的索引库。这样完全和数据库进行了隔离。

**数据查询方法**

**顺序扫描法**, 所谓顺序扫描，例如要找内容包含一个字符串的文件，就是一个文档一个文档的看，对于每一个文档，从头看到尾，如果此文档包含此字符串，则此文档为我们要找的文件，接着看下一个文件，直到扫描完所有的文件。

**倒叙索引**,提取资源中关键信息， 建立索引 （目录）

搜索时，根据关键字（目录），找到资源的位置

**Lucene与搜索引擎的区别**

Lucene是一套用[**Java**](http://lib.csdn.net/base/java)或其它语言写的全文检索的工具包

搜索引擎是一个全文检索系统，它是一个单独运行的软件系统

**lucene的工作方式**

lucene提供的服务实际包含两部分：一入一出。所谓入是写入，即将你提供的源（本质是字符串）写入索引或者将其从索引中删除；所谓出是读出，即向用户提供全文搜索服务，让用户可以通过关键词定位源。

**写入流程**  
  源字符串首先经过analyzer处理，包括：分词，分成一个个单词；去除stopword（可选）。将源中需要的信息加入Document的各个Field中，并把需要索引的Field索引起来，把需要存储的Field存储起来。 将索引写入存储器，存储器可以是内存或磁盘。  
**读出流程**  
 用户提供搜索关键词，经过analyzer处理。对处理后的关键词搜索索引找出对应的Document。用户根据需要从找到的Document中提取需要的Field。

**Lucene的执行过程**

1、创建一个索引文件目录，然后把需要检索的信息 用Field 对应匹配的 封装成一个Document文档对象，将这个对象放入索引文件目录中，这里既可以将索引存放到磁盘中，也可以放入内存中，如果放入内存，那么程序关闭索引就没有了，所以一般都是将索引放入磁盘中；

2、如果发现信息有问题需要删除，那么索引文件也要删除，否则检索的时候还是会查询得到，这个时候需要根据索引id去删除对应的索引；

3、如果发现信息被更新了，那么索引文件也要更新，这个时候需要先将旧的索引删除然后添加新的索引；

4、最后是全文搜索，先创建索引读取对象，然后封装Query查询对象，调用search()方法得到检索结果。

## API核心

### Document

Lucene创建索引时的原始文档。

### IndexableField

* LongField：long类型，会切词
* StringField：string类型，通过StringField构建的字段不会被切词（分词）。
* TextField：Text类型，会切词。
* new XXXField(name：字段的名称,value：字段的值,store：是否存储到索引库中)

### Directory

用来指定索引库存放的位置。   
- Directory，指的是文件磁盘的索引路径   
- RAMDirectory，指的是内存中的索引路径

### Analyzer

在对Docuemnt中的内容进行索引之前，需要使用分词器进行分词 ，分词的目的是为了搜索。分词的主要过程就是先分词后过滤。

* 分词：采集到的数据会存储到document对象的Field域中，分词就是将Document中Field的value值切分成一个一个的词。
* 过滤：包括去除标点符号过滤、去除停用词过滤（的、是、a、an、the等）、大写转小写、词的形还原（复数形式转成单数形参、过去式转成现在式）等。
* 停用词是为节省存储空间和提高搜索效率，搜索引擎在索引页面或处理搜索请求时会自动忽略某些字或词，这些字或词即被称为Stop Words(停用词)。比如语气助词、副词、介词、连接词等，通常自身并无明确的意义，只有将其放入一个完整的句子中才有一定作用，如常见的“的”、“在”、“是”、“啊”等。
* 搜索使用的分析器要和索引使用的分析器一致

#### 中文分词器

* StandardAnalyzer：   
  单字分词：就是按照中文一个字一个字地进行分词。如：“我爱中国”，   
  效果：“我”、“爱”、“中”、“国”。
* CJKAnalyzer   
  二分法分词：按两个字进行切分。如：“我是中国人”，效果：“我是”、“是中”、“中国”“国人”。

### QueryParser(查询解析)

使用场景:需要对查询关键词进行分词时候,使用查询解析器.

### MultiFieldQueryParser

作用：指定多个字段检索并分词

### TermQuery

作用：根据词条进行检索，不会切词了，因为词条已是最小单位。

### WildcardQuery

作用：模糊检索

### FuzzyQuery

作用：相似度检索   
FuzzyQuery(Term term)：默认支持模糊字数为2；   
FuzzyQuery(Term term, int maxEdits)：maxEdits：模糊字数，[0,2]之间，若为0，相当于TermQuery。   
FuzzyQuery(Term term, int maxEdits, int prefixLength)：prefixLength，指定要有多个前缀字母必须完全匹配。

### NumericRangeQuery

作用：数字范围搜索（演示：略），最后两个参数的含义是：minInclusive，是否最小包含，maxInclusive，是否最大包含。

### MatchAllDocsQuery

作用：查询所有

### BooleanQuery

作用：组合查询   
1、MUST和MUST表示“与”的关系，即“交集”，相当与AND。   
2、MUST和MUST\_NOT前者包含后者不包含。   
3、MUST\_NOT和MUST\_NOT没意义   
4、SHOULD与MUST表示MUST，SHOULD失去意义；   
5、SHOULD与MUST\_NOT相当于MUST与MUST\_NOT。   
6、SHOULD与SHOULD表示“或”的关系，相当与OR。

Lucene常用功能介绍

1. lucene索引创建
2. lucene索引检索
3. lucene分词器
4. Query创建
5. IndexSearcher常用搜索

\*\*\*\*\*\*\*\*\*\*\*\*\*Lucene索引创建-创建索引关键类\*\*\*\*\*\*\*\*\*\*\*\*\*

IndexWriter

IndexWrite indexWrite=new IndexWrite(directory,indexWriteConfig);

// directory字典

// indexWriteConfig配置信息

Directory存储路径

Directory directory=FSDirectory.Open(“D://index/testindex”);

Analyzer分词器

Analyzer analyzer=new StandardAnalyzer(Version.Lucene\_30);

Document文档

Document doc=new Document();

Filed域

Doc.add(new TextField(“fieldName”,”测试标题”,Store.YES));

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*end\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*