

信息检索系统的设计与实现

姓名：李 阳

学号：2017111276

日期：2017.12.26

目录

基于 Lucene 的小型信息检索系统的设计与实现	1
一、系统设计要求	3
二、系统设计	4
三、系统实现	5
1. 后台业务逻辑	5
2. 程序界面	6
四、使用说明	6

一、系统设计要求

使用 Lucene 的 API 接口，设计并实现一个小型的信息检索系统，用户界面如图 1 所示：

文档路径:

索引路径:

建立索引

查询关键字:

查询

查询结果如下:

首页 上一页 下一页 尾页

图 1 检索系统界面示例

具体要求：

1. 支持的文档类型有：txt、doc、pdf、html、ppt、xls 和 xml；
2. 支持中英文文档内容；
3. 支持分页显示，每页显示数目可动态配置；
4. 需要上交可运行的程序和源代码，以及程序的设计和使用说明文档；
5. 验收时有统一的文档测试集。

二、系统设计

1. Lucene 是一套用于全文检索和搜寻的开源式程序库，由 Apache 软件基金会支持和提供，Lucene 提供了一个简单却又强大的应用程式接口，能够做全文的索引和搜寻。在 Java 环境里 Lucene 是一个成熟的免费开源工具，就其本身而言，Lucene 是最受欢迎的免费 Java 信息检索程序库，所以在使用 Lucene 时，在 Java 环境里是最方便的，本程序将采用 Java 来完成编写。
2. 从系统设计要求分析出，本系统要想实现两个核心功能：
 - 1) 根据用户提供的数据源创建索引
 - 2) 根据用户输入的关键词查找完成信息检索。
3. 用户在电脑中选择自己想要构建索引的数据集，然后讲索引文件写入系统的某个文件夹，所以本系统不需要数据库，只需要完成文件的读写操作就可以。

整个系统的设计图如图 2 所示：

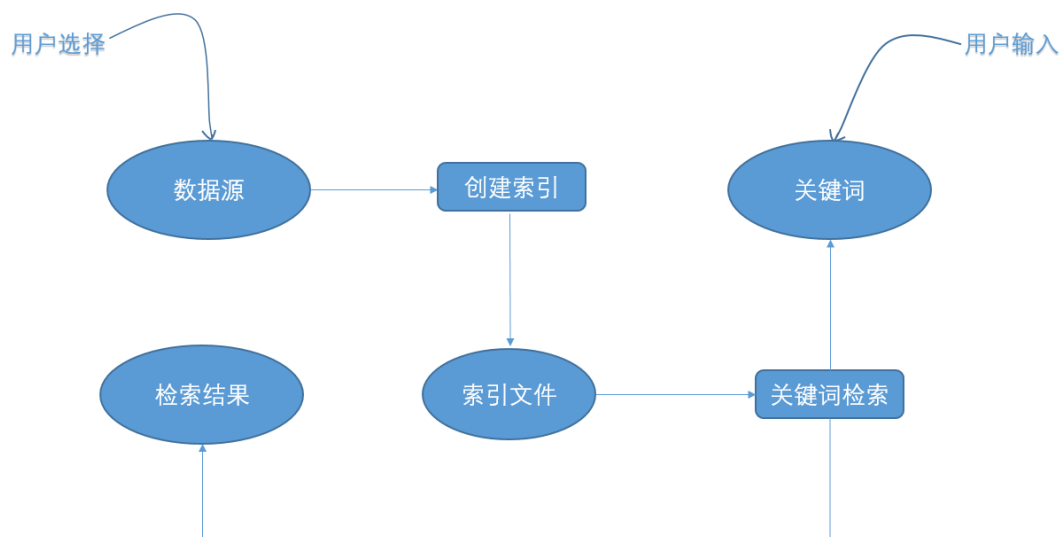


图 2 系统设计

4. 使用 MVC 模式实现整个程序设计，将界面和后台业务逻辑分离，提高代码的可读性和重构性。具体设计类图如图 3 所示：

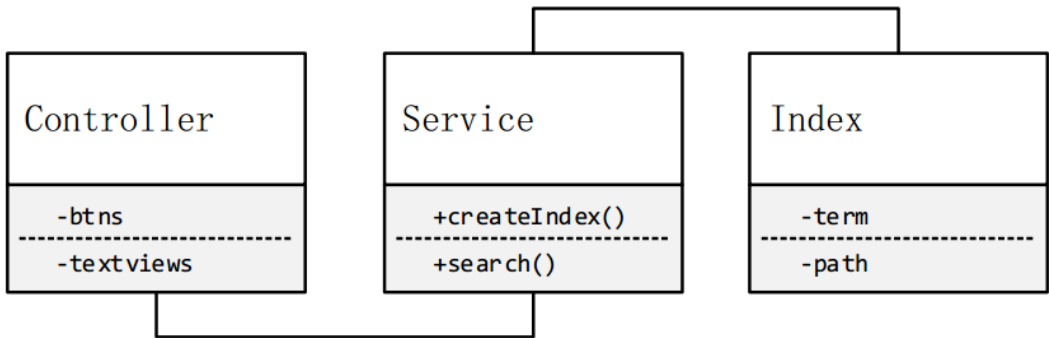


图 3 ULM 类图

三、系统实现

1. 后台业务逻辑

使用 Java 在安装了 Java 8 的环境下实现, 文件读以及路径的读写使用 java.nio.*。索引创建使用 Lucene 的核心功能类 IndexWriter，信息检索使用的核心类 IndexSearcher。因为本系统是全文检索，所以针对不同的文档类型，需要使用不同的读取工具包来实现对不同的文档类型的内容读取, 引入的工具包如下图 4 所示:












 commons-logging-1.2.jar	2017/9/8 23:08	Executable Jar File	61 KB
 jxl.jar	2011/10/28 8:35	Executable Jar File	689 KB
 lucene-analyzers-common-7.1.0.jar	2017/10/13 16:13	Executable Jar File	1,584 KB
 lucene-core-7.1.0.jar	2017/10/13 16:12	Executable Jar File	2,715 KB
 lucene-highlighter-7.1.0.jar	2017/10/13 16:13	Executable Jar File	194 KB
 lucene-queryparser-7.1.0.jar	2017/10/13 16:13	Executable Jar File	376 KB
 pdfbox-app-2.0.8.jar	2017/12/14 16:37	Executable Jar File	7,841 KB
 poi-3.17.jar	2017/9/8 23:17	Executable Jar File	2,638 KB
 poi-ooxml-3.17.jar	2017/9/8 23:17	Executable Jar File	1,445 KB
 poi-ooxml-schemas-3.17.jar	2017/9/8 23:17	Executable Jar File	5,786 KB
 poi-scratchpad-3.17.jar	2017/9/8 23:17	Executable Jar File	1,358 KB

图 4 系统引用的 jar 包

2. 程序界面

使用 JavaFX 框架实现界面设计，可以在 fxml 文件中书写界面设计代码，也可以通过可视化界面拖拽，本文使用 fxml 文件书写，界面略简陋。界面提供用户选择数据源文件夹位置和索引存储位置，和创建索引按钮，最终用户输入关键词，点击检索按钮，在 Table 中展示检索结果。整体的实现界面如图 5 所示：

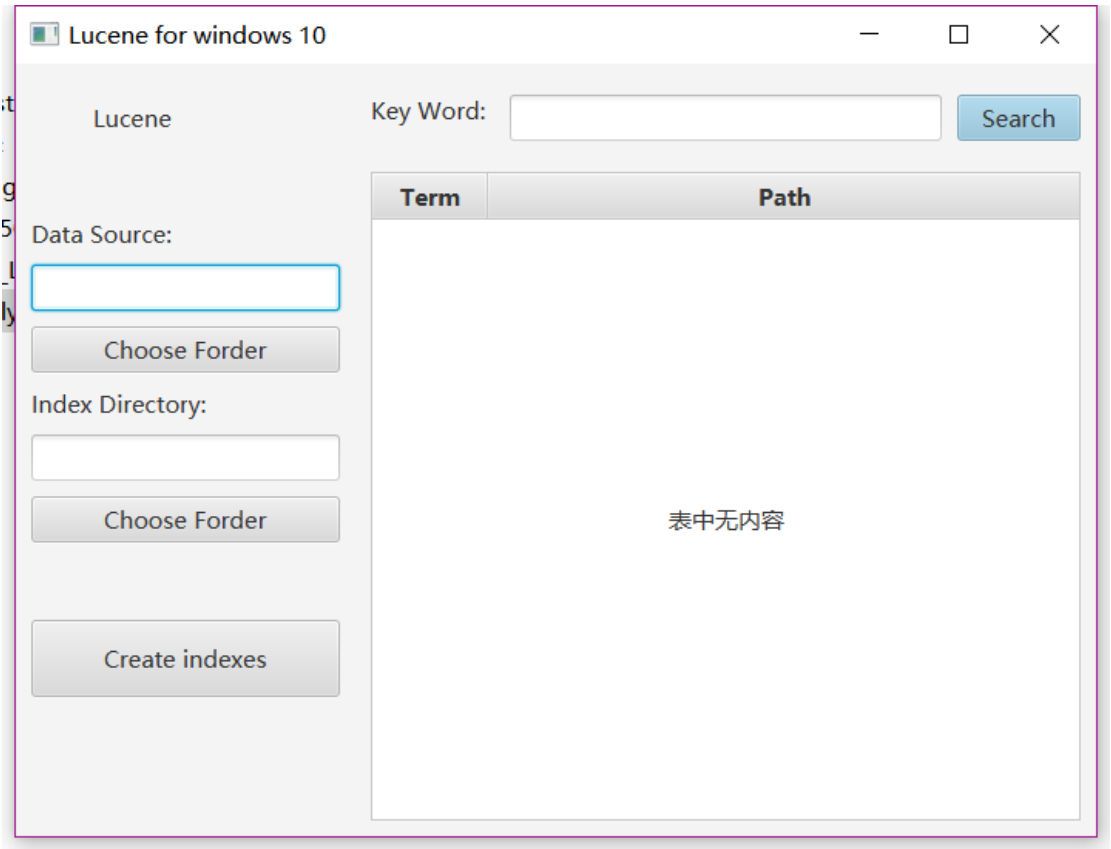
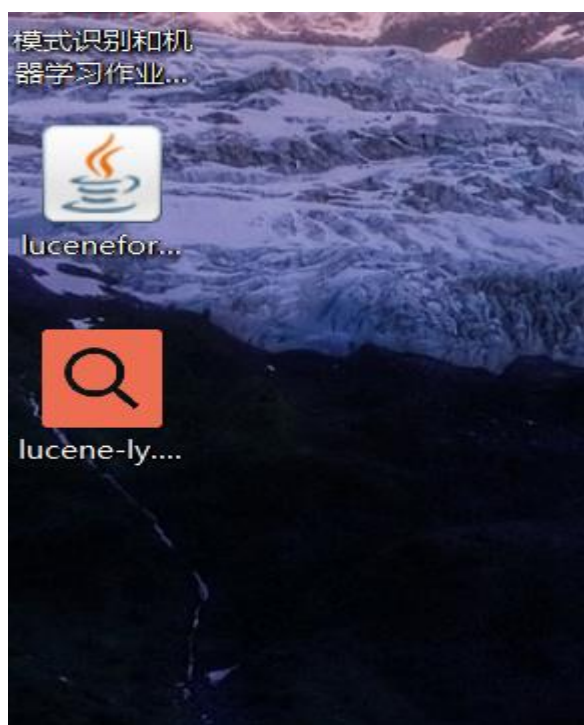


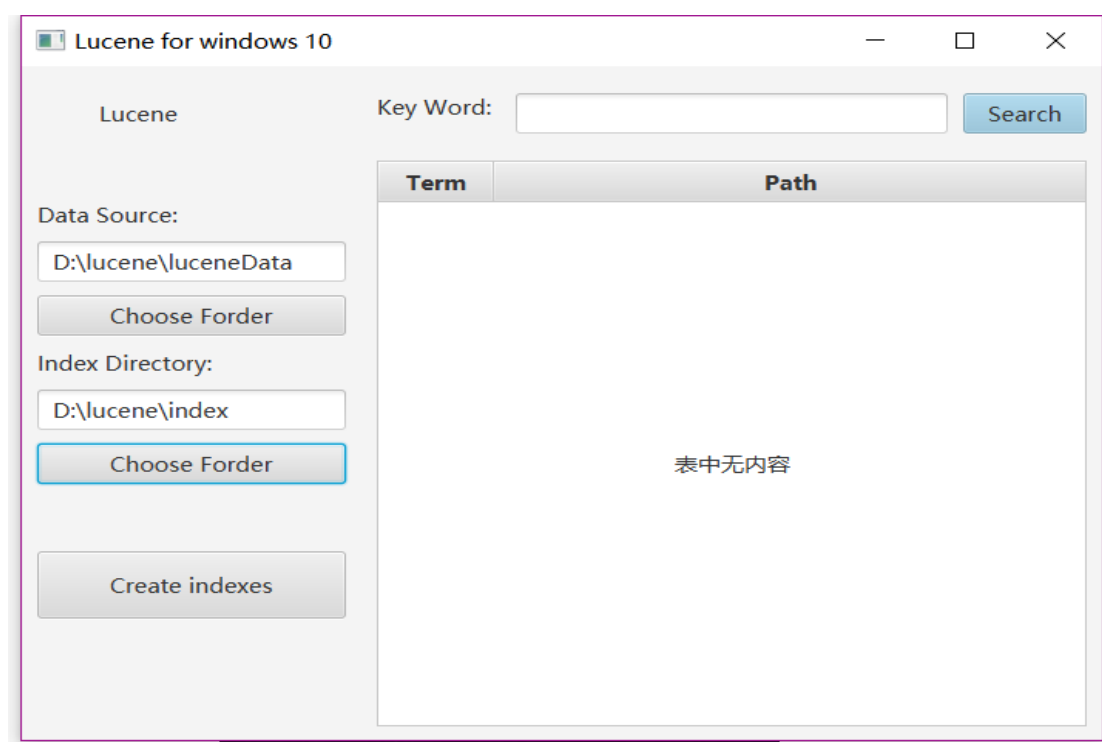
图 5 程序界面设计

四、使用说明

1. 双击打开桌面的 Lucene-Ly.exe 可运行



2. 点击 Choose Folder 分别选择数据源所在的文件夹目录和创建好的索引的存放目录（索引存放目录最好新建文件夹或者保证该文件夹为空）。



3. 点击 Create indexes 就开始创建索引，索引创建完成之后会有创建成功提示框。

