小型信息检索系统设计与实现

冯温迪
2016111435
网络技术研究院
Edit with LaTeX.

2016 年 12 月 21 日

目录

1 系统要求

使用开源软件 Apache Lucene 提供的 API 接口，设计并实现一个小型信息检索系统，用户界面如图 1所示。

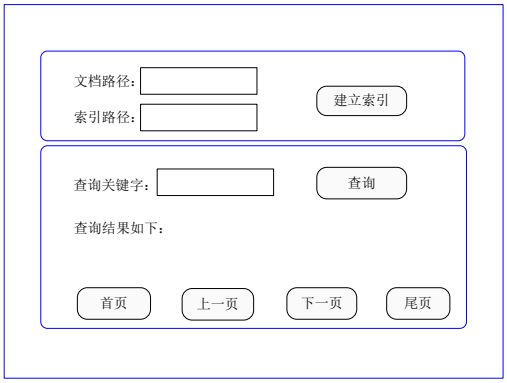


图 1: 要求用户界面

具体要求：

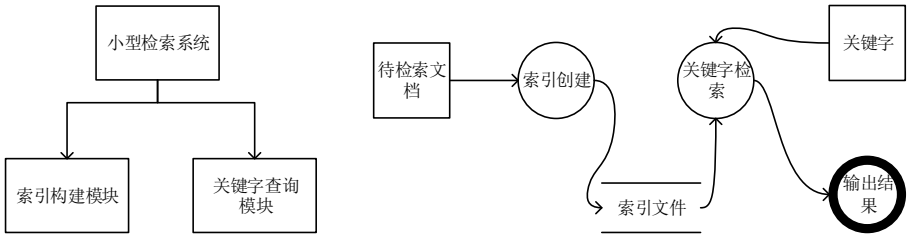
1. 支持的文档类型有：txt、doc、pdf、html、ppt、xls 和 xml；
2. 支持中英文文档内容；
3. 需要上交可运行的程序和源代码，以及程序的设计和使用说明文档；
4. 验收时有统一的文档测试集。

2 系统概要设计

系统要使用 Apache Lucene 所提供的 API 实现文件检索功能。Apache Lucene 为我们提供了一个全文检索引擎的架构，提供了完整的查询引擎和索引引擎，部分文本分析引擎（英文与德文两种西方语言）。Lucene 的目的是为软件开发人员提供一个简单易用的工具包，以方便的在目标系统中实现全文检索的功能，或者是以此为基础建立起完整的全文检索引擎 [1]。

通过对需求的分析，我们可以了解到整个系统主要有两个核心功能——构建索引以及关键词检索。系统结构图如图 2a所示。

系统的数据流向也比较明显，系统没有使用数据库，数据通过用户指定本地磁盘文件夹获取，输入经过索引构建模块生成索引文件，然后在查询模块，生成的索引文件和查询关键字作为输入，输出关键词及该关键字所在的文件的路径。系统数据流图如图 2b所示。



(a) 系统结构图　　(b) 系统数据流

图 2: 系统概要设计图表

3　系统详细设计

我们现在将第 2节中介绍的系统的两个模块分解，进行详细的设计。

由于功能比较简单，对于核心功能只用两个类组合即可。因为我们最后实现的形式是桌面应用程序（Desktop Application）所以，我们打算使用 MVC 的设计模式。这样可以使得界面与后台业务逻辑分开，提高代码的可读性。C 即 Ccontroller 用于控制界面的逻辑结构，响应事件；M 即 Model 是数据模型，这里就是 Index 类，V 即 View 就是程序的界面。那么我们的 UML 类图如图 3所示。

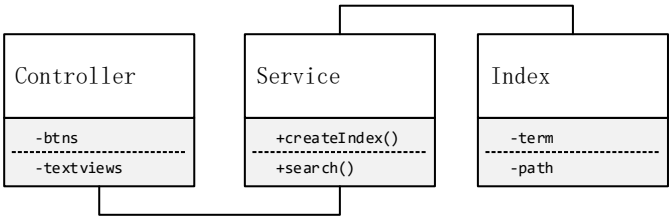


图 3: UML 类图

4　系统实现

本系统采用 Java 语言实现，一点是出于 Apache Lucene 的 API 接口是 Java 接口，另一点出于 Java 有跨平台的特点，即一次开发，随处运行（Develop once, run anywhere ）由于需要实现图形界面，这里我们使用 JavaFX 框架。JavaFX 框架有着简单、漂亮的特点，而且在编辑图形界面的时候，允许开发者使用 SceneBuilder 通过鼠标拖拽的方式“画”界面。

程序在创建索引的时候，首先从用户选定的文件夹读取每个文件，对每个文件，我们先判断文件的类型，然后用相关类型的文件解析器，读入文件的文本信息，然后通过 Apache Lucene 提供的

IndexWriter 类的对象，调用 addDocument() 方法将该文件添加到索引。这时，Lucene 会自动帮我们创建索引文件。

程序在关键字搜索的时候，系统使用的核心 API 是 IndexSearcher。通过该类的对象调用 search 方法，返回查询关键字所在文档以及命中次数（HIT）。

系统最后的运行结果如图 4所示。

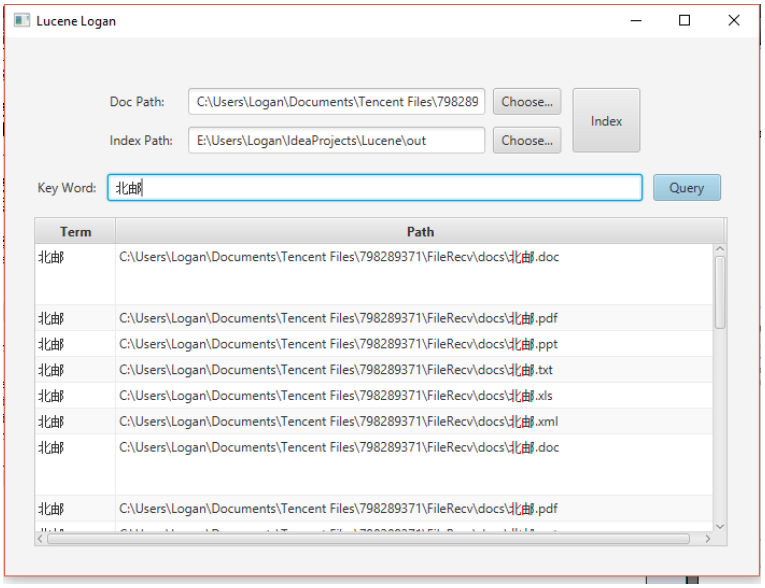


图 4: 系统运行结果

5 系统测试

本小型检索系统的测试系统选用了与开发系统相同的 Windows 平台，因为是 Java 程序，所以程序是跨平台的，用户可以使用任何安装了 Java8 的机器。

系统测试文档是课程所给的 docs 文档集合，包括 21 个文件，其中有 pdf、word、txt、html、xml、ppt、xml 类型的文件各 3 个。

系统测试用例为如表 1所示。经过测试，对于所有的测试用例，系统都能正确检出。

表 1: 系统测试用例

关键词	预计检出文档数
bupt	11
北邮	11
bupt 北邮	21
lcc	4

参考文献

[1] Michael McCandless, Erik Hatcher, and Otis Gospodnetic. *Lucene in Action: Covers Apache Lucene 3.0*. Manning Publications Co., 2010.