浙江大学计算机学院

Java 程序设计课程报告

2022—2023 学年秋冬学期

题目	简易问答机器人
学号	
学生姓名	
所在专业	计算机科学与技术
所在班级	

目　　录

1　引言

本次开发的程序是一个简易问答机器人。这个题目需要用到网络爬虫、索引建立、Java IO 等知识，有助于加深对 Java 程序设计语言和 Java 外部库的使用方式的理解。

1．1　设计目的

按照设计要求，我所设计的简易问答机器人是一个有类似于搜索引擎功能的控制台程序。具体功能如下：

(1) 使用爬虫程序爬取新浪爱问知识人网站的“理工学科”板块（链接），提取每个问题的标题和答案，并建立 Lucene 索引。

(2) 用户可以通过输入关键词或正则表达式按照标题查询信息，得到匹配的问题列表。

(3) 用户可查看匹配的问题和查看问题下的最佳答案。

1．2　设计说明

本程序采用 Java 程序设计语言，在 NetBeans IDE v8.0.2 下编辑、编译与调试。除引用的外部库外，全部程序编写均由我一人完成。

本程序使用的外部库是：

(1) soup v1.15.3，用以解析 html 文件。

(2) Lucene core & queries & queryparser v4.10.0，用来进行索引建立、更新以及查询。

(3) IKAnalyzer2012_FF，中文分词器。

2 总体设计

2．1 功能模块设计

本程序需实现的主要功能有：

(1) 用户自定义爬取页数（最多 100 页）并建立 Lucene 索引；

(2) 通过 Lucene 索引，按关键词查询索引内符合要求的问题列表；

(3) 显示用户查询到的问题的答案。

程序的总体功能较少，如图 1 所示：

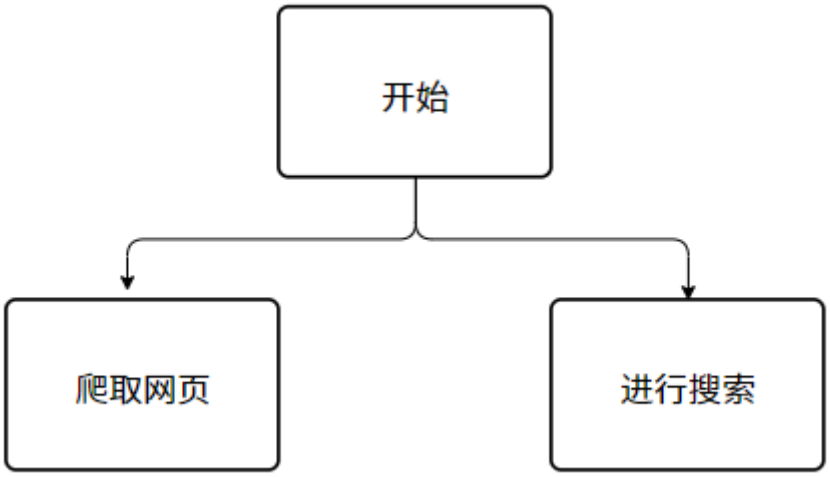

图 1 总体功能图

2. 2 流程图设计

由于程序是单线程的控制台程序，程序总体流程图仍可参考 Main 类说明，此处忽略。

3　详细设计

3．1　Main 类设计

在本程序设计中，Main 类是整个程序的入口类。类中只包含一个 main()方法，作为程序入口，其运行流程图如下所示：

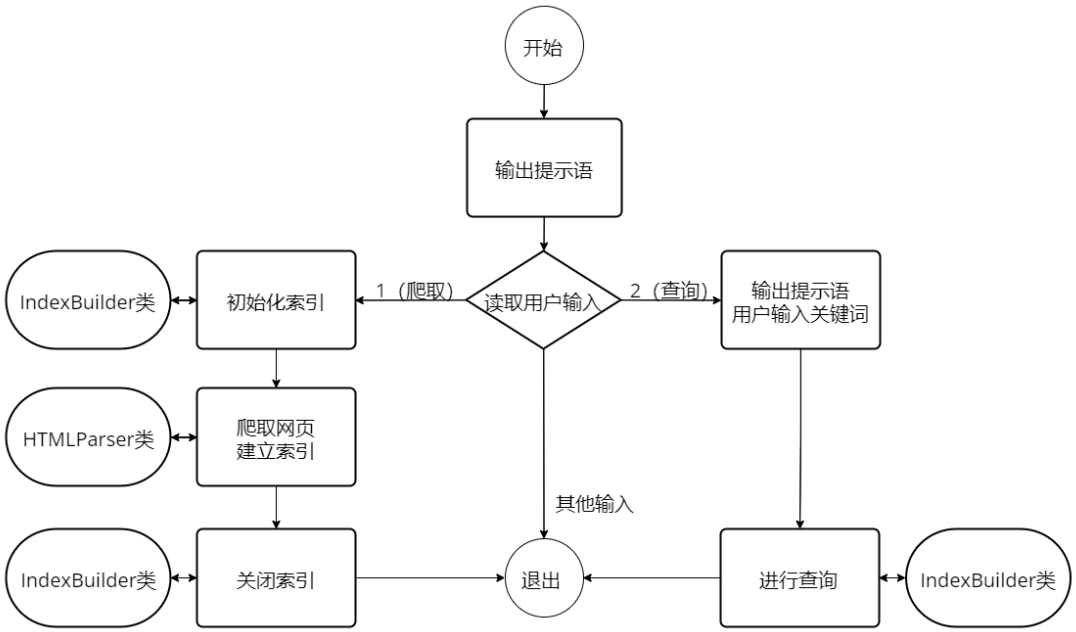

图 2 Main 类中 main()方法的流程图

输入输出模式：鉴于程序是控制台应用，我直接使用了 Scanner 类，并在构造函数中采用 GBK 编码来得到输入数据。

异常处理模式：鉴于程序不太需要异常处理相关机制，我只在 catch 段中输出异常信息及相关程序栈信息，以方便调试。

3．2　HTMLParser 类设计

根据上述的流程图可以看出，HTMLParser 类是用来爬取网页、建立问题索引的类，其源代码存在 HTMLParser.java 中。HTMLParser 类有以下成员方法和变量：

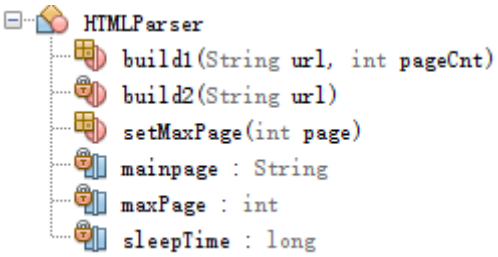

图 3 HTMLParser 类的成员

以下是该类中成员数据和方法的详细说明：

（1）成员变量

① mainpage 是私有的 String 类对象，存放一个常量字符串"https://iask.sina.com.cn"，用来计算出每个要爬取的问题对应的页面链接（一般是 mainpage + url，其中 url 是方法的一个参数）。

② maxPage 是私有的 int 类对象，表示需要爬取的问题页数（最多 100 页）。这个页数可以自行修改。

③ sleepTime 是私有的 long 类对象，用来确定每次爬取页面之前 Thread.sleep()的时间（预设定为 1000，即 1s）。这个时间可以由程序员修改，但太短可能会被 403 forbidden。实际测试中，设定为 1s 爬取一页也会被 forbidden，最后我是换了一次网络才爬取完毕的。

（2）方法

① setMaxPage(int)方法单纯用以依照参数设置 maxPage 变量。

② build1(String, int)方法按照参数所示的 url（需要是一个问题分类下的列表页）爬取数据（列表中的问题和对应链接、下一页的链接），并记录已爬取的页数。build1()的运行流程如图 4 所示。

③ build2(String)方法按照参数所示的 url（需要是某个问题的详细页面）爬取数据（问题标题及答案）并通过 IndexBuilder 类存入索引。Build2()的运行流程如图 5 所示。

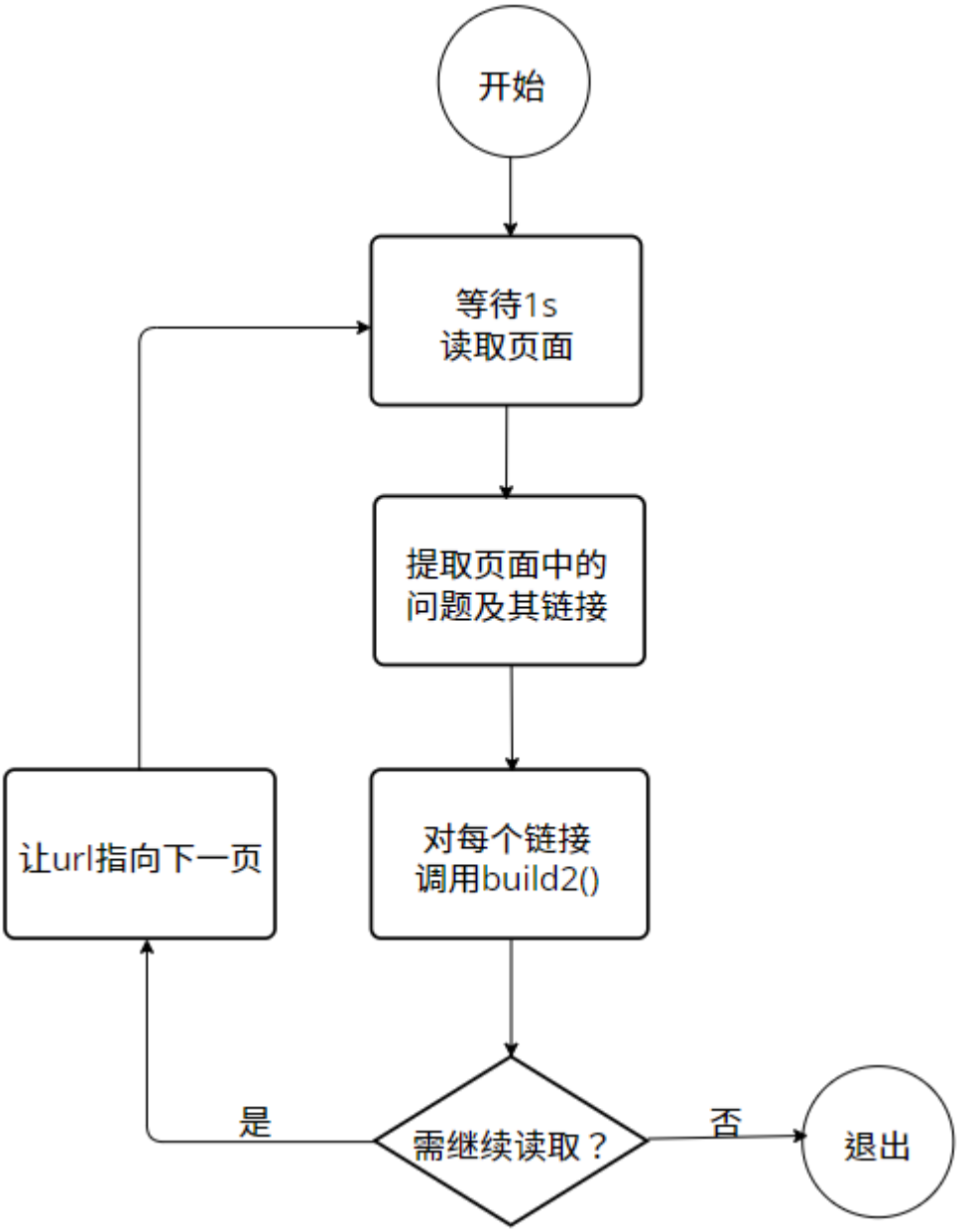

图 4 build1()流程图

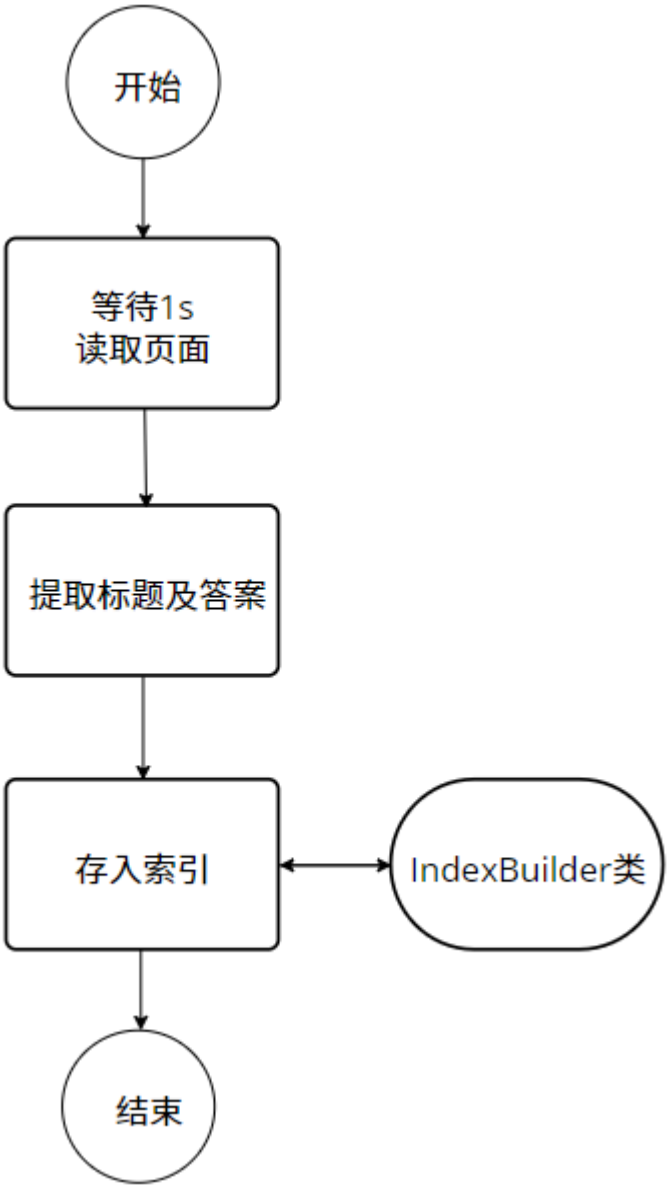

图 5 build2()流程图

异常处理模式：setMaxPage()不抛出异常。build1()和 build2()抛出 InterruptedException 异常,但通过 try-catch 块自行处理 IOException 异常(即访问连接超时)。处理方式是进行一次尾递归（即重新访问网页）。

3．3　IndexBuilder 类设计

根据上述的流程图可以看出，IndexBuilder 类是用来进行程序与索引间交互所使用的类，其源代码存在 IndexBuilder.java 中。IndexBuilder 类有以下成员方法和变量：

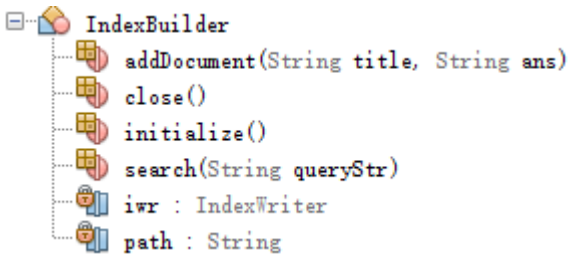

图 6 IndexBuilder 类的成员

以下是该类中成员数据和方法的详细说明：

（1）成员变量

① path 是私有的 String 类对象，存放一个常量字符串"d:/iaskIndex"。这是 Lucene 索引所存放的目录。

② iwr 是私有的 IndexWriter 类对象，用以对索引进行写入操作（其详细实现是在 Lucene 外部库中）。

（2）方法

① initialize()方法在指定的目录使用 IKAnalyzer 分词器创建一个新的 IndexWriter 对象，并让 iwr 指向它。

② addDocument(String, String)方法利用传入的 title 和 answer 参数建立一个文档并通过 iwr 加入到索引项目中。

③ close()方法关闭 iwr，完成索引创建。

④ search(String)方法完成一次以 queryStr 为关键词列表的查找，并输出前 20 个匹配数据的 title 和 answer 域。

该类的代码大部分是模仿示例程序编写的。

异常处理模式：把异常全部抛出。

4 测试与运行

4．1 程序测试

经大量测试，实验要求的最基本功能实现已经完备，可以期待后续完善一些其他功能。

4．2 程序运行

以下实验结果不在控制台界面运行，而是采用 NetBeans 的调试器，其输入输出情况与控制台类似。

进入程序，会先收到提示语句：

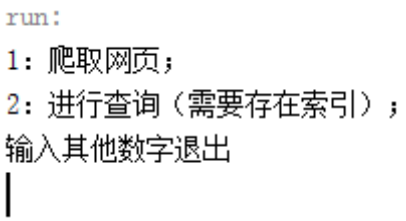

图 7 程序入口提示

如果选择输入 1，则会收到以下提示：

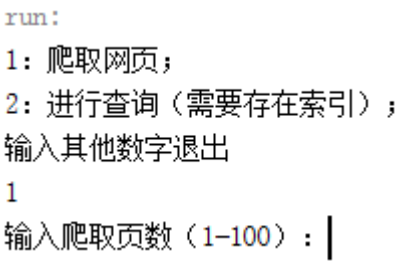

图 8 爬取入口提示

输入合适页数后程序即开始爬取并输出提示信息。爬取完指定页数后程序自动终止。

```
输入爬取页数（1-100）：100
读取页数1
https://iask.sina.com.cn/b/9t0DYVygF5.html读取完成
https://iask.sina.com.cn/b/2wNJYuW9OvS.html读取完成
https://iask.sina.com.cn/b/2xbvCyJuLuU.html读取完成
https://iask.sina.com.cn/b/9goMZVYnaM.html读取完成
https://iask.sina.com.cn/b/pgjgP2PWBSo.html读取完成
https://iask.sina.com.cn/b/1qFjDAGYK6c.html读取完成
https://iask.sina.com.cn/b/phoRWgffqms.html读取完成
https://iask.sina.com.cn/b/phvGVHlg276.html读取完成
```

图 9 爬取情况提示（成功）

```
https://iask.sina.com.cn/b/9klm1lfAe8.html读取失败；重试……
https://iask.sina.com.cn/b/9klm1lfAe8.html读取完成
https://iask.sina.com.cn/b/2au2ZKFiMtsA.html读取失败；重试……
https://iask.sina.com.cn/b/2au2ZKFiMtsA.html读取完成
```

图 10 爬取情况提示（失败）

在索引建立完成后，如果选择进行查询，会得到以下提示信息：

```
run:
1：爬取网页；
2：进行查询（需要存在索引）；
输入其他数字退出
2
输入查询关键词（以空格隔开）：
```

图 10 查询入口提示

输入关键词即可查询到最优匹配的 20 条提问与其对应的最优回答，之后程序终止。

```
输入查询关键词（以空格隔开）：优势 我
路灯杆防撞墩有哪些我们不了解的优势？
我们所安装的电杆防撞墩质地坚韧：该产品采PolyethylineLLDPE主要原料，质轻坚韧，搬运简便，耐震、耐冲击，给我们电杆带来了较好的保护；

我看重庆自成的是数字式电子汽车衡,有什么优势吗?
数字的话,挺有优势的,高度精准,完全是智能控制,比其他形式的要方便许多。
```

图 11 示例查询结果

5．总结

通过这个程序的编写，我对 Java 程序设计语言中的 IO 控制、网络支持、外部库使用、异常处理机制等更加熟悉了。

本次程序编写运用了很多外部库，但同时也有很多需要程序员自行处理的部分，如连接超时的处理等。通过对这些结果的处理，我认识到，编写一个完整的程序需要对程序语言有着深刻的了解。

总体说来，本次程序编写的结果还是成功的。

参考文献

[1] Java 中 scanner.next()键盘输入中文乱码以及转码乱码的问题 - 代码先锋网(codeleading.com)

[2] (26 条消息) java 爬虫入门 jsoup 入门（简单示例，五分钟）_蛋黄卷阿龙的博客-CSDN博客