# 浙江大学大学计算机科学与技术学院

Java 程序设计课程报告

2017-2018 学年秋冬学期

题目	Web 搜索引擎
学号	3160104875
学生姓名	杨樾人
所在专业	软件工程
所在班级	软工 1602

# 目录

1	引言	1
	1.1 设计目的	1
	1.2 设计说明	
2	总体设计	3
	2.1 功能模块设计	
	2.2 流程图设计	
3		
	3.1 爬取丁香园各个疾病的内容	5
	3.2 对每个疾病的内容建立索引	7
	3.3 用户通过输入查询关键词进行疾病的查询	S
4	测试与运行	11
	4.1 程序测试	11
	4.2 程序运行	11

### 1 引言

本次是用 Java 开发一个 Web 搜索引擎,需要学习 Web 爬虫、解析网页内容、对内容建立索引和查询等知识。通过本次作业,我可以对 Java 语言中的各项功能有更好的理解和使用,通过具体的程序来加深对 Java 语言的掌握,提高自己的编程水平,为以后的工作打下一定的基础。

### 1.1 设计目的

- 1. 写一个 Web 爬虫, 爬取问答类或课程类网站 的网页:
- 2. 解析网页内容,对内容进行结构化,并存储 到文件中:
- 3. 为内容建立索引;
- 4. 通过命令行进行内容检索,并展示内容列表

Web 搜索引擎是非常有实际意义的一个工程,不仅需要爬取网上的信息,还需要对这些信息建立索引,便于查询,本次我完成的搜索引擎功能如下:

- 1. 使用 Isoup 爬取丁香园的所有的疾病信息链接,约有一千一百个。
- 2. 对每个疾病信息链接再次进行爬取,对网页的内容进行解析,把有用的信息下载到本地文档之中。
  - 3. 使用 Lucene 对每个文档建立索引,并添加查询功能。
- 4. 用户通过在命令行输入字符串进行内容检索,程序会将检索到含有关键词的文档返回,并按相关度返回前20个最相关的文档。
  - 5. 用户可以通过输入文档前面的序号查看对应疾病的内容。

### 1.2 设计说明

本程序采用 Java 程序设计语言,在 Eclipse 平台下编辑、编译与调试。具体程序由我个人开发而成。工作时间轴如表 1 所示:

表1工作时间轴表

时间	完成的主要工作
12月1日	整个程序前期的需求分析和整体功能的架构 阅读 Lucene 和 Jsoup 的库文件,熟悉用法
12月7日	完成代码的编写工作,并对代码进行测试,查找 bug。
12月10日	完成报告的撰写工作

### 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. 爬取丁香园疾病的链接数据集
- 2. 爬取每个疾病链接的网页内容
- 3. 对爬取下来的每个疾病的内容建立索引
- 4. 用户通过输入查询关键词进行疾病的查询
- 5. 系统返回与用户查询关键词相关度最高的前 20 个疾病
- 6. 用户通过输入疾病前的代号,查看疾病的详细信息。

程序的总体功能如图 1 所示:

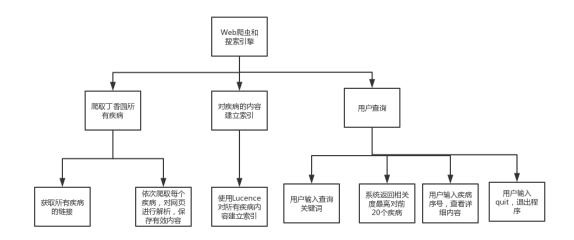


图 1 总体功能图

### 2.2 流程图设计

程序总体流程如图 2 所示:

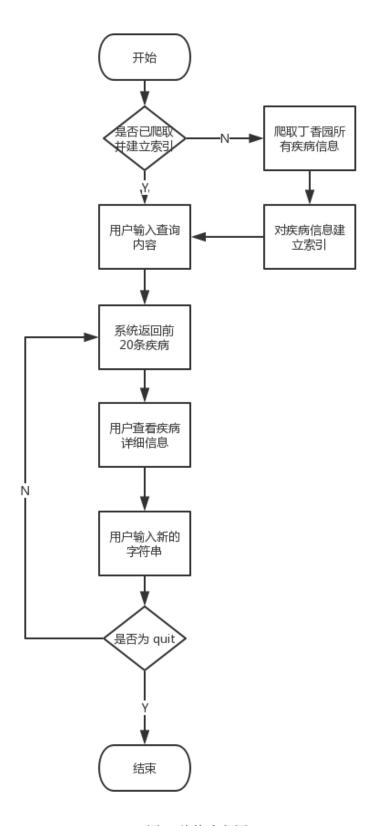


图 2 总体流程图

### 3 详细设计

本搜索引擎主要包含三部分,爬取网站内容,建立索引,用户查询,以下将做详细介绍。

### 3.1 爬取丁香园各个疾病的内容

打开丁香园疾病的汇总页面,每个疾病都有一个对应的链接,每种疾病的页面内容相同,有6个id,从1-6,分别对应简介、症状、病因、诊断、治疗、生活。每个id中,包含disease-detail-content-title和disease-detail-content-box两部分,disease-detail-content-content 是对title的详细展开介绍。

```
▶<div class="disease-detail-content-title">...</div
  ▶ <div class="disease-detail-content-box">...</div>
</div>
</div id="1">
 ▼<div id="2">
  ><div class="disease-detail-content-title">...</div
><div class="disease-detail-content-box">...</div>
</div>
▼<div id="3":
  ▶ <div class="disease-detail-content-title">...</div>
  ▶ <div class="disease-detail-content-box">...</div>
</div>
<div id="4">
  > <div class="disease-detail-content-title">...</div>
> <div class="disease-detail-content-box">...</div>
  </div>
▼ <div id="5">
  ▶<div class="disease-detail-content-title">...</div>
  ▶ <div class="disease-detail-content-box">...</div>
</div>
</div>
</div>
```

图 3 丁香园疾病网页

图 4 丁香园疾病网页详细信息

此部分只需要两个函数,主要是对疾病详细页面中有效内容的筛选。Spider 类中的 allUrl 存储页面中所有的疾病链接,currentUrl 是当前要爬取的页面,documentName 是存放的文件名。调用 Spider 类中的 getAllUrl 函数,返回所有的疾病链接,对疾病链接这个 String 数组进行遍历,对每个疾病链接调用 spiderUrl 函数,对该页面内容进行爬取,并存储到文档中。

UML 图如下:

### Spider

- + allUrl:String[]
- + currentUrl:String
- doucumentName:String
- + Spider():void
- getAllUrl(String): String[]
- spiderUrl(String):void

#### 图 5 爬虫 UML 图

以下是 UML 图中有关数据和方法的详细说明:

- (1) 成员变量
- a) allUrl 存储页面中所有的疾病链接,
- b) currentUrl 是当前要爬取的页面,
- c) documentName 是存放的文件名
  - (2) 方法
- a) Spider()方法是此类的入口函数,供主函数调用
- b) getAllUrl(String)是爬取丁香园疾病主页面的所有链接,并返回。
- c) spiderUrl(String)是爬取该页面的主要内容,并对页面进行解析,格式 化保存有价值的内容。

爬取页面的流程图如下:

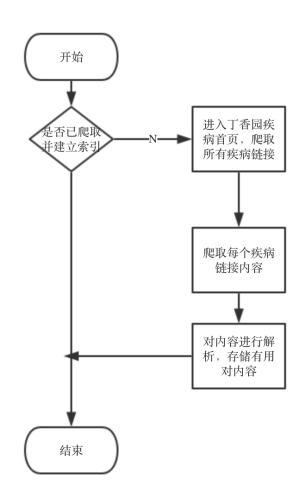


图 6 爬虫设计流程图

### 3.2 对每个疾病的内容建立索引

CreateIndex 通过调用 Lucene 包对每种疾病中的内容建立索引。该类使用时需要获取文档存储的位置。

Lucene 的索引包含多个 Document,每个 Document 可以包含多个 Field 对象。 我把每个疾病的内容分别放到一个 Document 中,然后添加到索引之中。Document 中存放了两个 Field,一个是疾病的名字,另一个是疾病的详细信息。

在 Query 时,只需要调用 Lucene 的 Query 函数,可以得到 Hits,里面包含了多个 doc,并按相关度对文档进行了排序。

## CreateIndex

- + filePath:String
- +getAllFile(String):String[]
- +createIndexForOneDocu(String):void
- +search(String):String[]

### 图 7 创建索引 UML 图

以下是 UML 图中有关数据和方法的详细说明:

- (1) 成员变量
- a) filePath 存储文档的位置。
  - (2) 方法
- b) getAllFile(String)获取文件夹内的所有疾病文件,返回文件名的字符 串数组。
- c) createIndexForOneDocu(String)是取出该文件的内容,并添加到Field 之中,然后加入到Document中,添加到Lucene的索引里。
- d) search(String)在索引中查找用户输入的字符串,并返回相关度最高的前 20 个文档的名字。

### 建立索引的流程图如下:



图 8 建立索引设计流程图

### 3.3 用户通过输入查询关键词进行疾病的查询

在建立好索引之后,主函数提供给用户搜索关键词的方法。

用户输入要查询的关键词,Query 类调用 CreateIndex 类中的 search 方法 对用户输入的关键词进行查找,返回包好关键字的相关度最高的前 20 个信息,显示在命令上,用户可以输入序号,查看该疾病的详细信息。

用户可以输入其他字符串,再次进行查找,也可以输入 quit,退出程序。

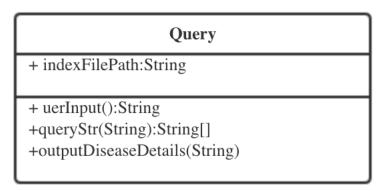
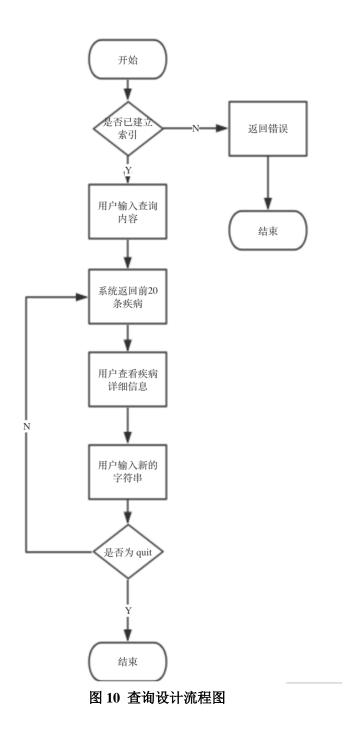


图9 查询 UML 图

以下是 UML 图中有关数据和方法的详细说明:

- (1) 成员变量
- a) indexFilePath 存放了 Lucene 建立的索引的位置。
  - (2) 方法
- b) userInput()方法是读入用户的输入,返回字符串。
- c) queryString(String)是在索引中查找该关键词所在的文档,调用 CreateIndex 类中的 search 方法,并返回前 20 个相关度最高的文档名。
- d) outputDiseaseDetails(String)是输出该文档的详细内容,供用户查看。

### 用户查询的流程图如下:



### 4 测试与运行

### 4.1 程序测试

在程序代码基本完成后,经过不断的调试与修改,最后测试本次所设计的 Web 爬虫和搜索引擎能够正常运行,没有出现明显的错误和漏洞,但是在一些细节方面仍然需要完善,比如用户交互方面,可以进行优化。总的来说本次设计在功能上已经基本达到要求,其他细节方面有待以后完善。

### 4.2 程序运行

爬虫爬取的疾病信息图所示:



图 11 疾病文档

● ● 0-艾森门格综合征.txt

简介

艾森门格综合征是什么? 艾森门格综合征(Eisenmenger syndrome, E5)是由艾森门格医生在 1897 年发现并总结的一种临床综合征。三个重要的因素共同组成了
艾森门格综合征: 存在先天性心脏病,造成肺循环和体循环的额外通路。 肺动脉高压为主要表现的肺动脉病变。 出现紫绀。 艾森门格综合征可以治愈鸣? 一旦出现艾森门格综合征,肺动脉和心脏的损害是不可逆转的,无法完全治愈。需要在生活中长期监测病情变化,使用合适的药物治疗,以避免病情恶化,延长期望寿命。严重的艾森门格综合

艾森门格综合征: 存在先天性心脏病,造成肺循环和体循环的额外通路。 肺动脉高压为主要表现的肺动脉病变。 出现紫绀。 艾森门格综合征可以治愈吗? 一旦出现艾森门格综合征,肺动脉和心脏的损害是不可逆转的,无法完全治愈。需要在生活中长期监测病情变化,使用合适的药物治疗,以避免病情恶化,延长期望寿命。严重的艾森门格综合征可以考虑通过心肺移植来争取生存的机会。 艾森门格综合征会影响寿命吗? 病人的生存时间是受疾病影响的,没有控制的艾森门格综合征病人,往往死于心功能衰竭,预期平均寿命 37 岁,也有病人存活至 60 岁。 艾森门格综合征常见吗? 艾森门格综合征往往在先天性心脏病患者进入成人期后才会出现症状。随着先天性心脏病手术治疗的发展,艾森门格综合征已经逐渐在减少了,但仍不容忽视。

班·VA 有什么症状? 突出的症状是紫绀,皮肤发黑,包括手脚皮肤、甲床、嘴唇发紫,这种皮肤颜色的变化不受活动影响,24 小时都可以观察到。另外由于慢性缺氧,手指末端指甲 所在那个指节会变得突出肥大,像捣药的杵子,也叫做杵状指。还有一些不典型的症状包括乏力、心悸、晕厥。 (杵状指) 病因 引起艾森门格综合征有哪些常见的原因? 常见的引起艾森门格综合征的先天性心脏病包括:室间隔缺损、房间隔缺损、动脉导管未闭。其他由于血管发育异常引起的血流通道异

引起艾森门格综合征有哪些常见的原因? 常见的引起艾森门格综合征的先天性心脏病包括:室间隔缺损、房间隔缺损、动脉导管未闭。其他由于血管发育异常引起的血流通道异常也会引起艾森门格综合征。 (室间隔缺损) (房间隔缺损) (动脉导管未闭) 正常的生理循环是由体循环和肺循环组成的,这两个循环互相不会直接交通。如果有发育异常,比如心脏房室间隔上出现小洞,血流就会在体循环和肺循环之间流动。这个是造成艾森门格症的最基本的原因。 什么时候容易出现艾森门格综合征? 未经治疗的先天性心脏病之后十余年,多见于接近成人期发病。不会在刚出生的时候就发病。因为长时间的异常血流才会造成肺动脉的损伤,最后出现肺动脉高压。 先天性心脏病越复杂,缺损越大,发生艾森门格综合征的机会越大。由于目前手术、心导管介入治疗等治疗手段的丰富,一部分先天性心脏病在儿科期间得到纠正,使艾森门格综合征的发病率下降。

大,女生又解〕格综合征的机会越大。由于自削手术、心导管介入治疗等治疗手段的丰富,一部分先大性心脏病在几种期间得到纠正,便又解〕格综合征的发病率下降。 诊断 怀疑艾森门格综合征需要做哪些检查? 心脏彩超、心电图、心导管检查、胸片、心脏磁共振或者 CT、肺部 CT、肺血管 CT、肺功能测定、血常规、心功能检查、凝血功能、 生化检查、血氧饱和度监测、血气分析、贫血检测。 为什么要做这些检查? 心脏彩超、心脏磁共振、心脏 CT: 了解心脏的结构变化,判断心脏发育异常的种类和血管走向。 心电图、心功能检查: 了解心脏的功能变化。 心导管检查: 可以显示心脏分流的特点,并确定肺血管阻力以及肺血管对舒张剂的反应。 胸片、肺部 CT、肺部血管 CT: 了解 肺部和肺部血管受影响的状态,寻找有无血栓形成。 血常规、凝血功能、生化检查: 评估全身状态,有无感染、凝血障碍,了解肝肾功能。 肺功能: 排除肺部原因引起的紫 绀。 血氧饱和度监测、血气分析: 判断体内氧气的情况,了解吸氧前后的血氧变化可以更好的判断病情。 贫血检测: 艾森门格综合征患者容易出现缺铁、所以需要了解铁代谢 相关的指标,包括铁蛋白。 艾森门格综合征的并发症有哪些? 心律失常、心功能衰竭、血栓形成,这些并发症可能会威胁患者的生命,造成病人的死亡。 治疗

治疗 艾森门格综合征需要去看哪个科? 心脏外科、心脏内科。 如何治疗艾森门格综合征? 由于没有根治的手段,需要综合治疗,评估心功能,减少发生并发症的机会,改善缺氧 对身体的损害,特别需要预防血栓的形成。 新的研究认为针对病因,使用肺血管扩张剂对于改善疾病有一定疗效,可以延长生存时间。但仍然需要更多的临床研究数值支持。可 以考虑使用的肺血管扩张剂包括波生坦、西地那非、依前列醇。 非常严重的病人可以考虑心肺联合移植,或者肺移植联合心脏修补术。 艾森门格尔综合征期的患者为什么不单 统行心脏修补术进行治疗? 艾森门格综合征则 最根本的原因是整个肺的动脉损伤。肺动脉挛缩增厚,形成持续性的肺动脉高压。 进入艾森门格综合征期,单纯修补心脏的缺损 是没有效果的。因为肺动脉高压已经让血流的方向出现改变,如果修补心脏缺损,那反而进一步加重了心脏负担,容易出现心脏功能衰竭。所以进入艾森门格尔综合症期,需要 解决的关键是肺动脉的异常压力升高。 艾森门格尔综合征出院后需要复查吗? 诊断艾森门格综合征后,需要每年在心脏专科门诊随访一次,评估心脏功能和有无并发症出现。 如果出现心悸或者发烧一定要马上前往医院就诊。

工文解门格尔综合征患者在生活中有哪些注意事项? 女性艾森门格综合征病人需要绝对避孕。因为艾森门格综合征的病人,一旦妊娠,死亡率可达 30%~50%,同时胎儿死亡率 和发育异常率也非常高。死亡风险最大的时期是分娩前后一周,孕期也有风险,故而需要一格避孕。 可以考虑不可逆的避孕方案,或者宫内节育器,选择避孕手法时应和妇产科 医生—起讨论:禁止使用含有复方(健孕激素)的口服避孕药、避孕贴,因为会加重血栓形成风险;正确使用避孕套的情况下,仍然有怀孕和6,所以不建设使用避孕套作为避 孕手段。如果万一怀孕,为了孕妇的生命考虑,建议终止妊娠。需在正规医院,采取安全的方式,密切监护生命体征。 避免过于劳累,避免中到重度的剧烈运动,避免蒸桑拿, 热水盆浴、泡温泉等处于高温环境的活动,避免过量饮酒。由于艾森门格综合征的病人处于体循环缺血状态。如果处于上述状态。可能会造成体循环血管阻力下降,导致右向左 分流增加,继而导致脑部缺血、心功能衰竭,威胁生命。 避免去高海拔的地区旅游。特别是处于海平面 5000 英尺以上的高原。因为高原地区空气含氧量低,会加重身体缺 氧。 避免进行手术操作。任何手术操作都可能会危及生命,围手术期死亡率为 19%。所以只有在威胁生命的情况下,才考虑进行手术。尽量避免一切外科手术操作。

#### 图 12 疾病信息内容

#### 建立索引如图所示:



图 13 索引文件

12

### 用户查询如图所示:

### 图 14 用户查询界面

### 用户查看详细信息,如图所示:

#### 图 15 查看疾病详细信息

### 用户持续查询如图所示:

#### 图 16 用户持续查询

### 用户退出查询如图所示:

```
You can type the number in front of the item to check the details of the disease or new string to query!

1:口唇疱疹; 2:表皮囊肿; 3:脾功能亢进; 4:嘯酸性粒细胞性胃肠炎; 5:小儿种痘后脑脊髓炎; 6:脐带绕颈; 7:鼻窦炎; 8:晕动病; 9:蜂蜇伤; 10:妊娠期糖尿病;

11:布鲁岛病; 2:水痘; 13:胆囊切除术后综合征; 14:胎儿窘迫; 15:專麻疹; 16:妊娠期肝内胆汁淤积症; 17:鎮虫性皮炎; 18:宫颈腺囊肿; 19:阴茎珍珠状丘疹; 20:心动过缓; quit

Good bye!
```

#### 图 17 用户退出查询

### 5. 总结

此次的 Web 编程是我第一次使用 Java 的库进行爬取网站和建立搜索引擎,第一感觉是 Java 的库太方便了。我之前也有用 python 的 beautifulsoup4 进行过爬取百度百科并建立搜索引擎的尝试,以为已经是比较方便的了。此次使用 Java 进行爬虫和建立索引,体会到了 Java 的方便之处。

使用 Jsoup 对网页内容进行爬取并没有遇到太大的困难,我比较感兴趣对是 lucene。引用一段网上对 lucene 重要概念的介绍:

Document:索引包含多个 Document。而每个 Document 则包含多个 Field 对象。
Document 可以是从数据库表里取出的一堆数据,可以是一个文件,也可以是一个网页等。
注意,它不等同于文件系统中的文件。

Field: 一个 Field 有一个名称,它对应 Document 的一部分数据,表示文档的内容或者文档的元数据(与下文中提到的资源元数据不是一个概念)。一个 Field 对象有两个重要 属性: Store (可以有 YES, NO, COMPACT 三种取值) 和 Index (可以有 TOKENIZED, UN TOKENIZED, NO, NO NORMS 四种取值)

Querv: 抽象了搜索时使用的语句。

IndexSearcher: 提供 Query 对象给它,它利用已有的索引进行搜索并返回搜索结果。 Hits: 一个容器,包含了指向 一部分搜索结果的指针。

我一开始按照老师给的样例进行了代码编写,觉得 lucene 建立索引的流程如下:将输入的数据源统一为字符串,然后从数据源提取数据,创建合适的 Field 添加到对应数据源的 Document 对象之中。我只使用了 TextField,因为我在把文档中的数据放到 StringField 时会无法建立索引,我还没有查到为什么,有可能是超过长度了。

当从 IndexSearcher 对象得到搜索结果(Hits)之后, 可以直接从中获取 需要的值,再格式化予以输出。但是为了交互的简洁性和美观,我选择了列出前 20 个相关度最高的疾病并对他们编号,用户可以输入编号进行更加详细的查看,我认为我是给用户做了一个接口,这个接口提供了便捷性,正如 lucene 包一样 简洁。

通过该次作业,我对 Java 调用库有了更深入对认识,理解了程序构造的一

般原理和基本实现方法。能够把课堂上学的知识通过自己设计的程序表示出来,加深了对理论知识的理解。在写代码调试的过程中,对 Java 的特性也有了更加深入的理解。

# 参考文献

- [1] 耿祥义. Java 大学实用教程[M]. 北京:清华大学出版社,2009.
- [2] 耿祥义. Java 课程设计[M]. 北京:清华大学出版社,2008.
- [3] <a href="https://lucene.apache.org/core/">https://lucene.apache.org/core/</a>
- [4] https://jsoup.org/