浙江大学计算机学院

Java 程序设计课程报告

2017—2018 学年秋冬学期

题目	建议 WebCrawler 设计开发
学号	3150103457
学生姓名	邓墨琳
所在专业	计算机科学与技术
所在班级	1502 班

目　录

1 引言

本次编写的是一个简单的 Web Crawler，这是一个比较简单的项目，主要考察类的建立、jar 包导入以及开源软件库的使用，对今后的学习十分有帮助。

1.1 编写目标

(1) 写一个简单的 Web Crawler,爬取某个网站的网页。

(2) 注意重复网页的问题。

(3) 爬取深度的问题。

(4) 通过命令行进行网页内容的检索，并展示网页列表。

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) 获取页面中的每个 URL，并解析 html 信息；

(2) 基于于网页 dom 树来解析 html 信息，提取标题和正文；

(3) 创建索引和使用索引查询；

程序的总体功能如图 1 所示：

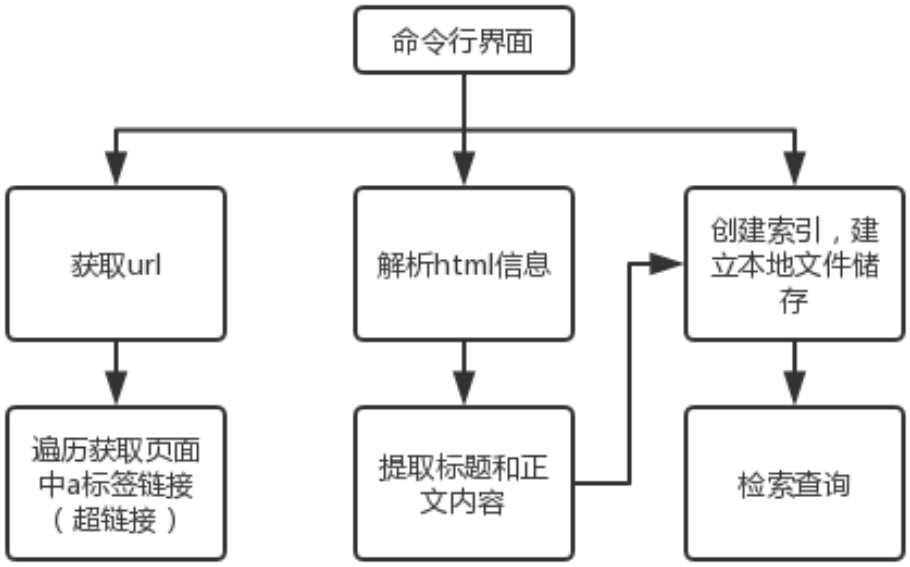


图 1 总体功能图

2.2 流程图设计

程序总体流程如图 2 所示：

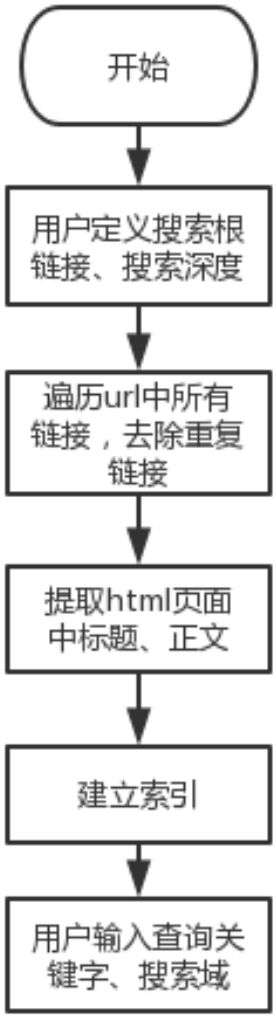


图 2 总体流程

3 详细设计

3.1 类设计

ParseURL 类使用 Jsoup 库实现遍历某个 url 链接中所有<a>标签对应超链并储存这些连接的类，实现了 processPage 接口，成员变量类型为 ArrayList<String>储存的每一个解析得到的 url 地址。

(1)成员变量

①uList 是储存的 String 类型的 ArrayList，储存解析得到的 url 地址。

(2)方法

①processPage(String URL, int depth, int d)根据用户指定深度递归调用，遍历 url 地址中所有<a>标签对应的链接地址，储存在 uList 中。

Crawler 类中通过 Lucene 接口实现了创建索引和使用索引查询。

(1)成员变量

①depth 为 int 类型，用户指定搜索深度。

②url 为 String 类型，用户指定搜索根 url 地址。

③u 是 ParseURL 类型的成员变量，通过该成员调用 processPage 方法，访问 uList 提取标题正文，创建索引。

(2)方法

①CreateIndex(String filePath)通过 Lucene 接口创建索引，同时调用 getWebInfo(int k)创建并添加 docement 文件。

②getWebInfo(int k)获取 uList 中第 k 个链接，提取标题、正文，创建 document，并返回单个 document。

③search(String filePath)根据用户指定检索条件和搜索域进行检索，在命令行显示目标 url 的标题和正文内容。

ParseURL 类和 Crawler 类是聚合关系，UML 类图如图三所示。

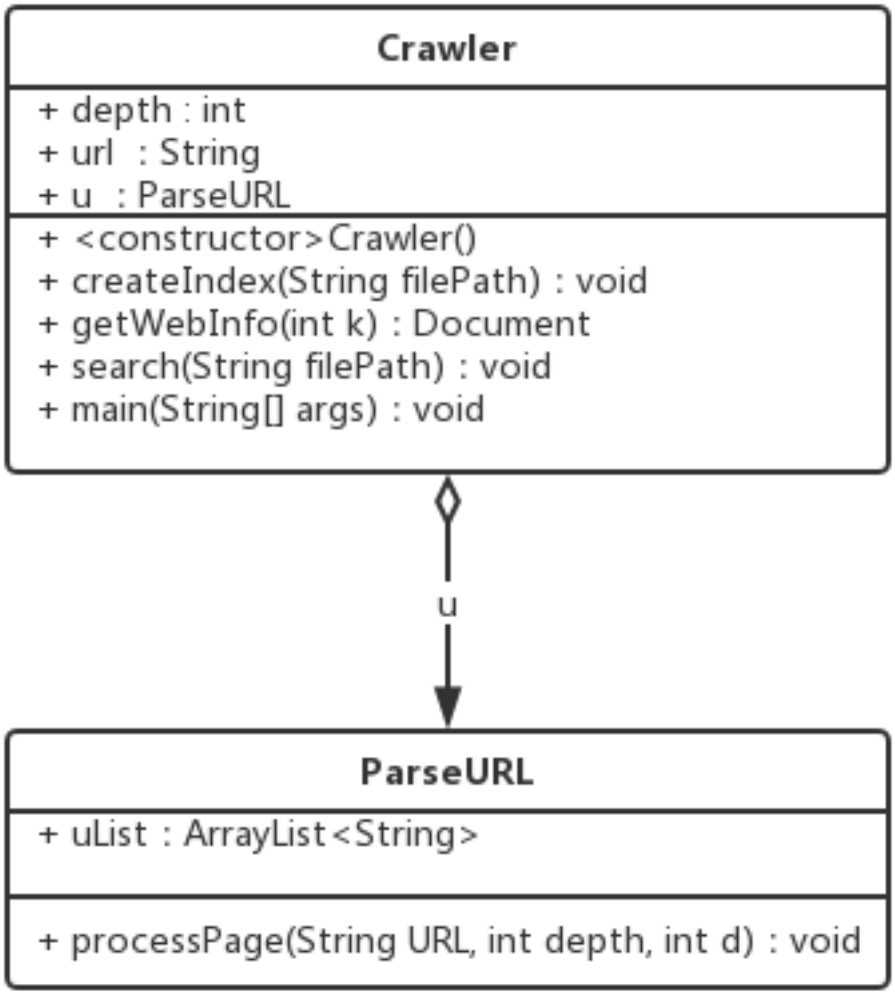


图 3 类的 UML 图

3.2 方法实现

(1)processPage(String URL, int depth, int d):void

URL:String 为 url 地址

depth:int 为搜索深度

d:int 为 depth-当前深度，即剩余遍历深度

```
if(d == depth) {
    uList.add(URL);
    System.out.println(URL);
}
if(d == 0)      return;
try {
    Document doc = Jsoup.connect(URL).timeout(0).get();
    Elements links = doc.select("a[href]");
        for(Element link: links) {
            if(uList.size() == 200) break;
            String s = link.attr("abs:href");
            if(!uList.contains(s)) {
                for(int i = 0; i < (depth-d+1); i++)
                    System.out.print(">");
                    uList.add(s);
                    System.out.println(s);
                 }
             processPage(s, depth, d-1);
            }
        }
}catch (Exception e) {
    return;
}
```

①利用 jsoup 解析 HTML 页面

②过滤出 a 标签链接添加进 ArrayList

③对每个 url 绝对地址递归调用该函数，直到到达指定深度

Note：uList 的大小限制在 200 条，因为数量过多造成创建索引和 doc 文件速度较慢。

(2)createIndex(String filePath):void

filePath:String 为索引文件路径，默认路径为工程目录下的 index 文件

```
File f=new File(filePath);
IndexWriter iwr=null;
try {
   Directory dir=FSDirectory.open(f);
   Analyzer analyzer = new IKAnalyzer();
   IndexWriterConfig conf=new
IndexWriterConfig(Version.LUCENE_4_10_0,analyzer);
   iwr=new IndexWriter(dir,conf);//建立 IndexWriter。固定套路
   //添加 doc
   iwr.deleteAll();
   for(int i = 0; i < u.uList.size(); i++) {
      Document doc=getWebInfo(i);
      iwr.addDocument(doc);//添加 doc，Lucene 的检索是以
document 为基本单位
   }
} catch (IOException e) {
   // TODO Auto-generated catch block
   e.printStackTrace();
}
try {
   iwr.close();
} catch (IOException e) {
   // TODO Auto-generated catch block
   e.printStackTrace();
}
```

①指定索引存放目录

②创建中文分词器

③创建 indexwriter，遍历 uList 中所有链接地址，调用 getWebInfo()创建文件，并添加文件。

(3)getWebInfo(int k):Document

```
try {
   final HTMLDocument htmlDoc = HTMLFetcher.fetch(new
URL(u.uList.get(k)));
   final TextDocument web_doc = new
BoilerpipeSAXInput(htmlDoc.toInputSource()).getTextDocument
();
   //前面几行代码就是一些固定方式，修改不同的 url，获取不同的值。
使用时用 doc.getTitle(),doc.getContent()即可。
   //web_doc.getTitle()和 web_doc.getContent()也可直接访问
   String title = web_doc.getTitle();
   String content =
ArticleExtractor.INSTANCE.getText(web_doc);
   if(title == null) title = "NULL";
   if(content == null) content = "NULL";
   Document doc=new Document();
   Field f1=new
TextField("url",u.uList.get(k),Field.Store.YES);
   Field f2=new TextField("title",title,Field.Store.YES);
   Field f3=new
TextField("content",content,Field.Store.YES);
   doc.add(f1);
   doc.add(f2);
   doc.add(f3);
   return doc;
}catch(Exception e) {
   //e.printStackTrace();
   return new Document();
}
```

①获取 url 链接的 HTML 信息

②提取标题、正文

③创建搜索域，添加搜索域名称

④返回 Document 文件

(4)search(String filePath):void

```
File f=new File(filePath);
try {
   IndexSearcher searcher=new
IndexSearcher(DirectoryReader.open(FSDirectory.open(f)));
   Scanner in = new Scanner(System.in);
   System.out.print("Input query string: ");
   String queryStr= in.nextLine();//用户输入查询条件，filed
中的内容，可以使用正则表达式
   Analyzer analyzer = new IKAnalyzer();//分词器
   System.out.print("Input query field: ");
   String queryFie= in.nextLine();//用户输入查询 field

   //指定 field 为“name”，Lucene 会按照关键词搜索每个 doc 中的

name。即搜索域
   QueryParser parser = new
QueryParser(Version.LUCENE_4_10_0, queryFie, analyzer);
   Query query=parser.parse(queryStr);
   TopDocs hits=searcher.search(query,u.uList.size());//前
面几行代码也是固定套路，使用时直接改 field 和关键词即可
   for(ScoreDoc doc:hits.scoreDocs){
      Document d=searcher.doc(doc.doc);
      System.out.println(d.get("title"));
      System.out.println(d.get("url"));
      System.out.println(d.get("content"));
   }
} catch (IOException | ParseException e) {
   e.printStackTrace();
}
```

①创建 indexSearcher，指定索引库的地址

②指定查询条件、指定搜索域

③遍历得出结果

④关闭资源

4 测试与运行

4.1 程序运行

查询 root URL：www.zju.edu.cn/english

深度 depth = 2

查询条件 queryString = “research”

搜索域 field = “content”

根据深度、内容多少，访问速度，建立索引、创建文件预计等待时间会增加

```
Input root URL: www.zju.edu.cn/english
Input searching depth: 2
http://www.zju.edu.cn/english
>http://www.zju.edu.cn/english/main.htm
>>http://www.zju.edu.cn/english/2016/1020/c2905a205077/page.htm
>>http://www.zju.edu.cn/english/2016/1020/c2906a205325/page.htm
>>http://www.zju.edu.cn/english/main.htm#
>>http://www.zju.edu.cn/
>>http://www.zju.edu.cn/russisch
>>http://www.zju.edu.cn/deutsch
>>http://www.zju.edu.cn/english/wbout/list.htm
>>http://iczu.zju.edu.cn/english/redir.php?catalog_id=225
>>http://www.zju.edu.cn/english/wcademics/list.htm
>>http://www.zju.edu.cn/english/wesearch/list.htm
>>http://www.zju.edu.cn/english/wampuswwife/list.htm
```

```
Crawler [Java Application] C:\Program Files\Java\jre1.8.0_131\bin\javaw.exe (2018年1月5日 下午8:50:12)
>>http://www.hanyushi.zju.edu.cn/redir.php?catalog_id=263
>>http://study%20of%20language%20and%20cognition%2C%20center%20for/
>>http://www.css.zju.edu.cn/english/
>>http://www.z2hospital.com/cms/EnglishDefault.htm
>>http://english.srrsh.com/
>>http://www.zjkq.com.cn/English/index.html
>>http://www.ceu.zju.edu.cn/itpe_en/index.asp
>>http://bksy.zju.edu.cn/english/
>>http://vlsi.zju.edu.cn/english/overview.htm
>>http://www.cab.zju.edu.cn/IVS/english/
>>http://www.womanhospital.cn/root/english/
>>http://zcni-test.zju.edu.cn/en/main_en.html
>>http://www.zupuc.org/
>>http://www.zjdxghy.com/en/
>>http://www.zju.edu.cn/english/2016/1020/c2906a205325/page.htm#
>http://www.zju.edu.cn/english/#
>>http://www.zju.edu.cn/main.htm
>>http://www.zju.edu.cn/551/list.htm
>>http://www.zju.edu.cn/552/list.htm
>>http://www.zju.edu.cn/553/list.htm
>>http://www.zju.edu.cn/554/list.htm
>>http://www.zju.edu.cn/555/list.htm
>>http://www.zju.edu.cn/556/list.htm
>>http://120.zju.edu.cn/
>>http://www.zju.edu.cn/xxgk
>>http://www.zju.edu.cn/560/list.htm
>>http://my.zju.edu.cn/main/loginIndex.do
>>http://www.zju.edu.cn/533/list.htm
>>http://www.zju.edu.cn/512/list.htm
>>http://www.zju.edu.cn/513/list.htm
>>http://www.zju.edu.cn/514/list.htm
>>http://www.zju.edu.cn/572/list.htm
>>http://www.zju.edu.cn/573/list.htm
>>http://www.zju.edu.cn/574/list.htm
>>http://www.zju.edu.cn/575/list.htm
```

父节点前面的>符号比子节点少一个，如上图红框标识所示

遍历完成，输入查找条件和查询域

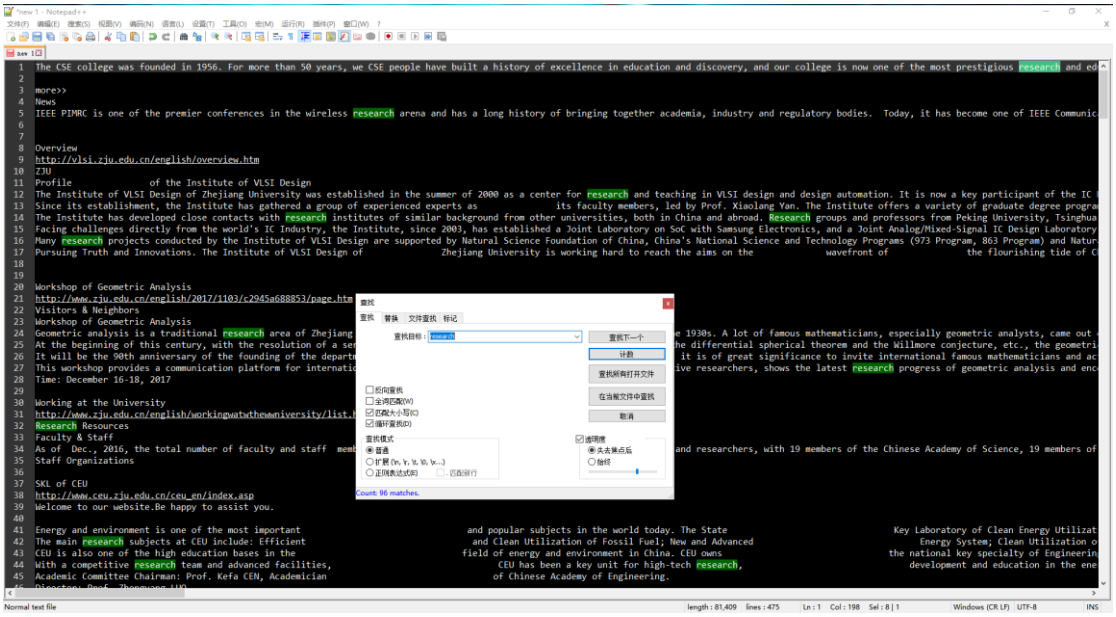
```
>>http://www.zju.edu.cn/512/list.htm
>>http://www.zju.edu.cn/513/list.htm
>>http://www.zju.edu.cn/514/list.htm
>>http://www.zju.edu.cn/572/list.htm
>>http://www.zju.edu.cn/573/list.htm
>>http://www.zju.edu.cn/574/list.htm
>>http://www.zju.edu.cn/575/list.htm
>>http://www.zju.edu.cn/576/list.htm
>>http://www.zju.edu.cn/515/list.htm
>>http://www.zju.edu.cn/glysj/list.htm
>>http://www.zju.edu.cn/583/list.htm
>>http://www.zju.edu.cn/584/list.htm
>>http://www.zju.edu.cn/585/list.htm
>>http://www.zju.edu.cn/588/list.htm
>>http://www.zju.edu.cn/589/list.htm
>>http://www.zju.edu.cn/519/list.htm
>>http://www.zju.edu.cn/590/list.htm
>>http://www.zju.edu.cn/591/list.htm
>>http://www.acv.zju.edu.cn/page.html?$$$originPortlet=mashuplisttopic:xg:mainLeft.view&m=wszt8
>>http://www.news.zju.edu.cn/
>>http://www.zju.edu.cn/zhfw
>>http://www.zju.edu.cn/536/list.htm
>>http://www.zju.edu.cn/593/list.htm
>>http://www.zju.edu.cn/594/list.htm
>>http://www.zju.edu.cn/595/list.htm
>>http://www.zju.edu.cn/596/list.htm
>>http://www.zju.edu.cn/597/list.htm
>>http://www.zju.edu.cn/598/list.htm
>>http://www.zju.edu.cn/599/list.htm
>>http://www.zju.edu.cn/600/list.htm
>>http://www.zju.edu.cn/601/list.htm
>>http://www.zju.edu.cn/602/list.htm
>>http://www.zju.edu.cn/603/list.htm
Input query string: research
Input query field: content
```

```
<terminated> Crawler [Java Application] C:\Program Files\Java\jre1.8.0_131\bin\javaw.exe (2018年1月5日 下午8:50:12)

ZJU hosts International Youth Forum on Belt and Road
http://www.zju.edu.cn/english/2018/0104/c2944a773810/page.htm
ZJU hosts International Youth Forum on Belt and Road
2018-01-04
"Belt and Road" are undoubtedly among the top media buzzwords of 2017, but what are their implications for the youth around the world? Scholars and students di
The International Youth Forum on the Belt and Road was held at ZJU's Zhijiang campus on December 28-29. The forum aims to inspire innovative ideas among global
The forum was jointly organized by the Academy of International Strategy and Law (AISL) at ZJU and the International Academy of the Belt and Road ( IABR ) in H
At the opening ceremony, AISL President Professor WANG Guiguo delivered a keynote address. From the perspective of national strategy, Professor WANG analysed t
Mr. SUN Jungong, vice president of Alibaba and research fellow of AISL, said it was cultural and emotional bonds that drew young people from different countrie
As a major highlight of the forum, a new book titled "The Belt and Road Initiative and its Oppurtunities for the Youth" was launched. The book was completed by
Besides, AISL signed a strategic cooperation agreement with the Mainland-Hong Kong Joint Mediation Center to jointly commit themselves to  disseminating Chines
In the following sessions, the participanting students made presentations on various topics on political, economic, cultural, educational and legal issues pert
Presiding over the closing ceremony, Professor ZHAO Jun, vice dean of Guanghua Law School, gave a summary of the forum. He expressed his gratitude to the organ
The Belt and Road Initiative  (the Silk Road Economic Belt plus the 21st-century Maritime Silk Road) is a development strategy first proposed in 2013 by China'
Media Contact
HONG Jiaying , Foreign Affairs Secretary at Guanghua Law School
ZHONG Yuan , Secretary, Center for Education & Teaching at Guanghua Law School
EVENTS

College of Computer Science and Technology, Zhejiang University
http://cspo.zju.edu.cn/english/
The International Symposium on Visual Information Communication and Interaction (VINCI) is a multi-disciplinary conference that provi...
01/17/2012
Recently, National Natural Science Foundation of China published the list of the winners of National Science Fund for Distinguished Y...
Links:
Search:
Introduce to CS
Research on computer science and technology of Zhejiang University can be traced back to early 1960's. In 1973 the Department of Radio Electronic Engineering s
```

我们将输出的内容复制到 notepad++中，利用查找功能可以清晰的看到，所有 content 中都包含有 resaerch 关键词。



5. 总结

通过 WebCrawler 的设计实现，我熟悉了解 Jsoup、Lucene 以及 Boilerpipe 的使用方法和相关原理，提高了编程技巧。通过该课程设计，全面系统的理解了程序构造的一般原理和基本实现方法。把死板的课本知识变得生动有趣，激发了学习的积极性。把学过的编程思想的知识强化，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。

参考文献

[1] luceneAPI 的简单使用（java）

http://blog.csdn.net/HANLIPENGHANLIPENG/article/details/53143656

[2] 如何使用 Java 语言实现一个网页爬虫

http://blog.csdn.net/uniquewonderq/article/details/50619899