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．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．． | 1 |
| 1.1 编写目标．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．． | 1 |
| 2 设计思路．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．． | 1 |
| 2.1 功能模块设计．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．． | 1 |
| 2.2 流程图设计．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．． | 2 |
| 3 详细设计．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．． | 3 |
| 3.1 类设计．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．． | 3 |
| 3.2 方法实现．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．． | 5 |
| 4 运行结果．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．． | 9 |
| 5 总结．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．． | 11 |
| 参考文献．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．．． | 12 |

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。

1. **总体设计**

2.1 功能模块设计

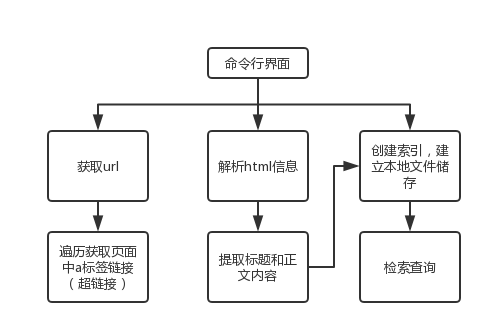
本程序需实现的主要功能有：

(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；

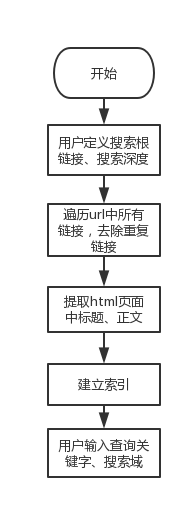
程序的总体功能如图1所示：

****

**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**

1. **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。

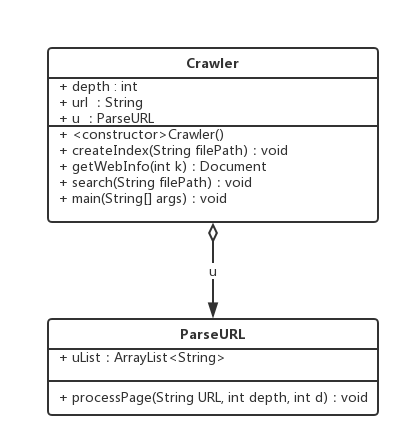
1. 方法

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;

}

利用jsoup解析HTML页面

过滤出a标签链接添加进ArrayList

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();

}

指定索引存放目录

创建中文分词器

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();

}

获取url链接的HTML信息

提取标题、正文

创建搜索域，添加搜索域名称

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();

}

创建indexSearcher，指定索引库的地址

指定查询条件、指定搜索域

遍历得出结果

关闭资源

1. **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

zju . edu . cn/english 
Input LIRL: 
Input searching depth: 
http://•eee.•w.zju.edu.cn/english 
>http://•eee.•w.zju.edu.cn/english/main.htm 
>>http://•eee.•w.zju.edu.cn/english/2ß16/Iß2ø/c29ß5a2ß5ß77/page.htm 
>>http://•eee.•e'.zju.edu.cn/english/2ß16/Iß2ø/c29ß6a2ß5325/page.htm 
>>http://•eee.•w.zju.edu.cn/english/main.htm# 
>>http• 
.//mw.zju.edu.cn 
>>http://•eee.•w.zju.edu.cn/russisch 
>>http://•eee.•w.zju.edu.cn/deutsch 
>>http://•eee.•w.zju.edu.cn/english/wbout/list.htm 
>>http://iczu.zju.edu.cn/english/redir.php?cataIog_id=225 
>>http://•eee.•w.zju.edu.cn/english/wcademics/list.htm 
>>http://•eee.•w.zju.edu.cn/english/wesearch/list.htm 

Crawler [Java Application] .8.0 131 (201EÉ5± 
>>http://•eee.•w.hanyushi.zju.edu.cn/redir.php?cataIog_id=263 
>>http://study%2øof%2øIanguage%2ßand%2øcognition%2C%2øcenter%2øfor/ 
>>http://•eee.•w.css.zju.edu.cn/english/ 
>>http://•eee.•w.z2hospitaI.com/cms/EnglishDefauIt.htm 
>>http• 
. //english . . com/ 
>>http://•eee.•w.zjkq.com.cn/EngIish/Index.htmI 
>>http://•eee.•w.ceu.zju.edu.cn/itpe_en/index.asp 
>>http://bksy.zju.edu.cn/english/ 
>>http://vlsi.zju.edu.cn/english/ovetwiew.htm 
>>http://vwwv.cab.zju.edu.cn/IVS/engIish/ 
>>http://•eee.•e'.womanhospital.cn/root/english/ 
>>http• 
.//zcni-test.zju.edu.cn/en/main_en.htmI 
>>http• 
>>http://•eee.•w.zjdxghy.com/en/ 
>>http://•eee.•e'.zju.edu.cn/english/2ø16/Iß2ø/c29ß6a2ß5325/Page.htm# 
>http://•eee.•w.zju.edu.cn/english/# 
edu.cn 
zyu.edu.cn/xxgk 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
://12ø. • 
zyu. 
zyu. 
edu.cn/main.htm 
edu.cn/551/Iist.htm 
edu.cn/552/Iist.htm 
edu.cn/553/Iist.htm 
edu.cn/554/Iist.htm 
edu.cn/555/Iist.htm 
edu.cn/556/Iist.htm 
edu.cn/56ø/Iist.htm 
>>http://my.zju.edu.cn/main/loginlndex.do 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
edu. 
cn/533/1ist.htm 
edu. 
cn/512/1ist.htm 
edu. 
cn/513/1ist.htm 
edu.cn/514/Iist.htm 
edu. 
cn/572/1ist.htm 
edu. 
cn/573/1ist.htm 
edu. 
cn/574/1ist.htm 
edu. 
cn/575/1ist.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

zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
acv. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
zyu. 
edu. 
cn/512/1ist.htm 
edu. 
cn/513/1ist. 
edu.cn/514/Iist.htm 
edu. 
cn/572/1ist.htm 
edu. 
cn/573/1ist.htm 
edu. 
cn/574/1ist.htm 
edu. 
cn/575/1ist.htm 
edu. 
cn/576/1ist. 
edu. 
cn/515/1ist.htm 
edu.cn/glysj/list.htm 
edu. 
cn/583/1ist.htm 
edu.cn/584/Iist.htm 
edu. 
cn/585/1ist.htm 
edu. 
cn/588/1ist.htm 
edu.cn/58g/Iist.htm 
edu.cn/519/Iist.htm 
edu.cn/59ø/Iist.htm 
edu.cn/591/Iist.htm 
edu.cn/zhfw 
edu. 
cn/536/1ist. 
edu. 
cn/sg3/1ist.htm 
edu.cn/594/Iist.htm 
edu. 
cn/595/1ist. 
edu.cn/596/list.htm 
edu. 
cn/sg7/1ist.htm 
edu.cn/598/Iist.htm 
edu.cn/5gg/Iist.htm 
edu.cn/6øø/Iist.htm 
edu.cn/6ßI/Iist.htm 
edu.cn/6ß2/Iist.htm 
edu.cn 
68/ list. htm 
: xg:mainLeft. 
Input query string: 
research 
Input query field: 
content 

Crawler [Java Application] 131 (201EÉ5± 
ZJLI hosts International Youth Forum on Belt and Road 
http://•eee.•w.zju.edu.cn/english/2ß18/ßIø4/c2944a77381ß/page.htm 
ZJLI hosts International Youth Forum on Belt and Road 
2ø18-ø1-ø4 
'Belt and Road" are undoubtedly among the top media buzzwords of 2ø17, 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 ZJLl'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 ZJLI and the International Academy of the Belt and Road ( IA8R ) in H 
At the opening ceremony, AISL President Professor WANG Guiguo delivered a keynote address. FAY the perspective of national strategy, Professor WANG analysed t 
Mr. SLIN 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 connit 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 surmary 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 2ß13 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 Connunication and Interaction (VINCI) is a multi-disciplinary conference that provi. 
al/17/2ø12 
Recently, National Natural Science Foundation of China published the list 
Links: 
Search: 
Introduce to CS 
Research on computer science and technology of Zhejiang University can be 
of the 
tra ced 
"Inners 
back to 
of National Science Fund for Distinguished 
early 196ø '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。

1 
2 
3 
4 
5 
6 
7 
8 
9 
10 
11 
12 
13 
14 
15 
16 
17 
18 
19 
20 
21 
22 
23 
24 
25 
26 
27 
28 
29 
30 
31 
32 
33 
34 
35 
36 
37 
38 
39 
41 
42 
43 
44 
45 
*new I - Notepad++ 
Ln:l 
The CSE college was founded in 1956. For more than 50 years we CSE 
eople have built a history of excellence in education and discovery. and our college is now one of the mostprestigious and ed 
more>> 
News 
IEEE PIMRC is one of the premier conferences in the wireless 
Overview 
http: / /vlsi. zju . edu . cn/english 'overview. htm 
sear 
arena and has a long history of bringing together academia, 
industry and regulatory bodies . 
Tod ay , 
it has become one of IEEE Communic. 
Profile 
of the Institute of VLSI Design 
The Institute of VLSI Design of Zhejiang University was established in the summer of 2ØØØ as a center for 
and teaching in VLSI design and design automation. It is now a key participant of the IC 
sear 
Since its establishment, the Institute has gathered a group of experienced experts as 
its faculty members, led by Prof. Xiaolang Yan. The Institute offers a variety of graduate degree progral 
The Institute has developed close contacts with research institutes of similar background from other universities, both in China and abroad. Research groups and professors from Peking University, Tsinghua 
Facing challenges directly from the world's IC Industry, the Institute, since 2003, has established a Joint Laboratory on SOC with Samsung Electronics, and a Joint Analog/Mixed-Signa1 IC Design Laboratory 
Many Fesearch projects conducted by the Institute of VLSI Design are supported by Natural Science Foundation of China, China's National Science and Technology Programs (973 Program, 863 Program) and Natur. 
Pursuing Truth and Innovations. The Institute of VLSI Design of 
Workshop of Geometric Analysis 
http://www.zju.edu . cn /english / 2017 / 1103 /c2945a688853 / page . htm 
Visitors & Neighbors 
Workshop of Geometric Analysis 
Geometric analysis is a traditional research 
area of Zhejiang 
At the beginning of this century, with the resolution of a se 
It will be the 90th anniversary of the founding of the depart 
This workshop provides a communication platform for internati 
Zhejiang University is working hard to reach the aims on the 
fiie 
wavefront of 
the flourishing tide of CI 
e 1930s. A lot of famous mathematicians, especially geometric analysts, came out 
he differential spherical theorem and the Willmore conjecture, etc., the geometri. 
it is of great significance to invite international famous mathematicians and ac 
ive researchers, 
and researchers, 
shows the latest 
sear progress of geometric analysis and enc. 
Time: December 16-18, 2017 
Working at the University 
htt : www.z •u.edu.cn en lish workin 
Research Resources 
Faculty & Staff 
atwthewwniversit 
list . 
(D) 
o 
Count: 96 matches. 
Z] i%BÉg 
As of Dec., 
2016, the total number of faculty and staff mem 
Staff Organizations 
SKL of CELI 
http: / /www.ceu . zju.edu.cn/ceu enl index. asp 
Welcome to our website. Be happy to assist you. 
Energy and environment is one of the most important 
with 19 members 
of the 
Chinese Academy of Science, 
19 members of 
research 
The main 
CEU is also one of 
With a competitive 
Academic Committee 
subjects at CEU include: Efficient 
the high education bases in the 
research team and advanced facilities, 
Chairman: Prof. Kefa CEN, Academician 
and popular subjects in the world today. The State 
and Clean Utilization of Fossil Fuel; New and Advanced 
field of energy and environment in China. CEU owns 
Key Laboratory of Clean Energy Utilizat 
Energy System; Clean Utilization o 
the national key specialty of Engineerin: 
development and education in the ene 
Normal text file 
CEU has been a key unit for high-tech 
research 
of Chinese Academy of Engineering. 
length • 
. 81,409 
lines 
: 475 
col 
: 198 
Sel 811 
Windows (CR LF) 
UTF-8 
INS 

**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)

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