

**网络爬虫工具的设计与实现**

**阶段报告（二）**

|  |  |  |
| --- | --- | --- |
| 小组成员 | 姓名 | 学号 |
| 组长 | 王奕超 | 1501211000 |
| 组员 | 张晗 | 1501211029 |
| 组员 | 胡玉 | 1501210913 |

‌

2015年12月25日 星期五

1. **系统技术点划分**

经过讨论，我们将系统分为三个模块（对应着三个主要技术点）去实现，三个任务可同时进行：

* 数据获取
* 数据解析、组织与存储
* 数据展示

1. **具体分工及简要分析**
2. **数据获取**

* 主要负责人：张晗，王奕超
* 技术点：

由Apache开发的HttpClient(官网http://hc.apache.org)，以及HttpClient自带的框架Fluent，主要运用Fluent.Request和Fluent.Response抓取页面，运用Regex和Parse提取分析页面信息

* 处理过程简要说明：

详见文末demo（此处为了展示，先通过regex和parse将输出文件做了与处理，生成了主要数据信息）

* 子模块预期效果**：**

通过HttpClient以及Fluent获取到ITEye、CSDN等相关技术博客网站的编程语言（C、C++、java、python等）、大数据技术（Hadoop、openstack等）等技术博客的博文数据，并输出为HTML文本文件，供数据库进一步处理。

1. **数据解析、组织及存储**

* 主要负责人：王奕超，张晗
* 技术点：

Jsoup、持久化

* 处理过程简要说明：

使用Jsoup对第一步抓取的HTML文本文件进行解析，获取所需数据（标签，网页链接，标题等），并将这些数据持久化到MySQL数据库中；建立表格，新增字段parentId和id以此来确定某一博客所属类别及博客的动态查询展示。

* 子模块预期效果：

将获取的博客信息按类别存储到数据库中，方便之后web展示查询调用。

1. **数据展示**

* 主要负责人：胡玉
* 技术点：

Jsp、HTML、mysql数据库

* 处理过程简要说明：

从数据库中获取数据，并展示到web页面；分析数据库中表格里面的id和parentId号对内容进行分类并展示

* 子模块预期效果：

类似博客分类网

**通过HttpClinet以及Fluent框架获取数据demo:**

代码：

import java.io.BufferedReader;

import java.io.BufferedWriter;

import java.io.File;

import java.io.FileReader;

import java.io.FileWriter;

import java.util.jar.Attributes.Name;

import java.util.regex.Matcher;

import java.util.regex.Pattern;

import javax.lang.model.type.NullType;

import org.apache.http.client.fluent.Request;

public class webCrawler {

private static String test(int pageNum) throws Exception{

return Request.Get("http://blog.csdn.net/index.html?&page="+pageNum)

.addHeader("Accept", "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,\*/\*;q=0.8")

.addHeader("Accept-Encoding", "gzip, deflate, sdch")

.addHeader("Accept-Language", "zh-CN,zh;q=0.8")

.addHeader("Cache-Control", "max-age=0")

.addHeader("Connection", "keep-alive")

.addHeader("User-Agent", "Mozilla/5.0 (X11; Linux x86\_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/46.0.2490.80 Safari/537.36")

.connectTimeout(3000).socketTimeout(3000).execute().returnContent().asString();

}

private static String findAll(String ori, String regex,int groupNum) {

Pattern p = Pattern.compile(regex, Pattern.DOTALL);

Matcher m = p.matcher(ori);

String resu = "";

while(m.find()) resu = resu+m.group(groupNum);

return resu;

}

public static void main(String[] args) throws Exception{

try {

for (int i=1; i<2; i++){

String test = test(i);

Pattern p = Pattern.compile("<div class=\"blog\_list\">(.\*?)</div>", Pattern.DOTALL);

Matcher m = p.matcher(test);

while (m.find()) {

String info = m.group();

System.out.println(findAll(findAll(info, "<h1>(.\*)</h1>", 1), "<a .\*?>(.\*?)</a>", 1));

}

}

}

catch (Exception e) {

e.printStackTrace();

}

}

}

实现效果：在控制台打印如下信息，包括分类、标题、tags、作者、阅读、评论：



原网页信息：

