**网页爬虫\_链家网二手房（LianjiaSpider）代码说明文档**

图1爬虫系统爬取信息流程图

数据抓取地址address（https://bj.lianjia.com/ershoufang/）

获取address对应的html页面信息(String resultHtml)

解析resultHtml信息，得到房子实体列表list

将房子实体列表信息写入到数据库writeHouseInfo2DB(list)

GetWebPage.java

PageAnalyze.java

表1功能模块

|  |  |  |
| --- | --- | --- |
| package | class | 注释 |
| main | LianjiaSpider.java | 程序入口 |
| entity | HouseInfo.java | 定义房源信息实体 |
| RegHouseInfo.java | 定义房源信息正则实体 |
| jdbcUtil | JdbcUtil.java | 数据库连接类 |
| webspider | GetWebPage.java | 获取HTML页面信息 |
| PageAnalyze.java | HTML页面解析 |

**1.Package entity**

①HouseInfo.java 定义房源基本信息

**private** String title;

**private** String info;

**private** String floor;

**private** String totalPrice;

**private** String unitPrice;

**private** String imgUrl;

②RegHouseInfo.java 定义房源正则化实体

(regXXXX为网页中房源信息的正则表达式，书写依据图2)

**private** String regHousinfo;

**private** String regImg;

**private** String regTitle;

**private** String regTotalPrice;

**private** String regUnitPrice;

**private** String regInfo;

**private** String regFloor;

图2 获取房源数据目标及html代码



**2.Package webspider**

①GetWebPage.java 获取HTML页面信息

根据给定的地址address,获取对应的html页面信息，并以字符串的方式返回,具体流程如图3

图3.GetWebPage.java代码流程

将字符串地址转换为url地址

url = **new** URL(address);

获取HttpURLConnection实例

conn = (HttpURLConnection) url.openConnection();

打开连接会话

conn.connect();

将BufferReader对象中的数据转换为String字符串

String line = **null**;stringBuffer.append(line);

读取网页信息到InputStream对象in = conn.getInputStream();

从InputStream对象中读取网页信息到BufferReader对象

reader = **new** BufferedReader(**new** InputStreamReader(in));

关闭连接会话

返回网页信息字符串

return stringBuffer.toString()

②PageAnalyze.java HTML页面解析

根据给定的html页面，按照正则表达式规则进行正则化匹配解析出里边的数据信息，并返回字符串房源数据。

**匹配过程：**

1>首先将正则表达式进行编译

Pattern pattern= Pattern.*compile*(reg.getReg());

2>再通过调用模式的 matcher 方法从模式创建匹配器对象

Matcher matcher = pattern.matcher(strHtml);

3>最后使用matcher.find()方法扫描输入序列以查找与该模式匹配的下一个子序列，匹配成功返回true

图4.PageAnalyze.java代码流程

对爬取到的HTML页面（resultHtml）依据regHousinfo规则进行正则匹配，得到每套房源信息。

每匹配到一套房源信息，再对其每一项信息（title,imgurl...）按照其reg规则进行单个匹配以获取每套房子的每一项信息。

将获取到的每一项信息（title,imgurl...）添加至单个房子实体houseInfo

将循环外层的所有房子实体添加到一个ArrayList数组中（房子实体列表）

解析完所有房子实体后，返回房子实体列表

**3.Package jdbcUtil**

①JdbcUtil.java 数据库连接类

图5.JdbcUtil.java代码流程

声明并初始化本地数据库的url user password 以及数据库驱动方式Driver

*driver* = "com.mysql.jdbc.Driver"

用static静态代码块来完成mysql数据库驱动包的加载工作，使用

Class.*forName*(*driver*)加载driver变量对应的类

利用DriverManager类中的getConnection()方法会根据参数和服务器建立连接信息，然后返回给Conn

conn=DriverManager.getConnection(url,user, password);

释放连接所需要的ResultSet,Statement, Connection对象，以防占用系统资源

**4.Package main**

①LianjiaSpider.java 程序入口

主要流程见图1.