



高级网络爬虫 Java Crawler

周钊平[1100012779] & 赵天雨[1100012957]

一、 jCrawler 简介

1、背景

我们平时在上网的时候，都会用到许许多多的搜索引擎，例如百度、谷歌、必应等，这些搜索引擎可以十分方便的对互联网资源进行下载，并且经过一定的排名算法，将结果呈现给我们。我们同样可以使用 Java 实现一个小型的多线程多功能的高级搜索引擎，并且将搜索进度、搜索结果、搜索得到的资源分类、搜索得到的资源分析呈现给使用者。同时提供不同的搜索功能。

2、开发者

开发人员由 Z 氏公司¹的两位创始人周钊平[1100012779]、赵天雨[1100012957]同学合力完成，版权完全属于 Z 氏公司，属于自主知识产权。虽然只有两个人，因此在完成项目的过程中我们收获了很多经验。

3、版本控制

版本控制系统使用了 Git²，代码仓库使用了国内的 Git 仓库 Gitcafe³。代码详见 <https://gitcafe.com/pzi1win2go3/jCrawler>，完全开源。



4、开发环境

¹ Z 氏公司:由于两位创始人的姓都是 Z 开头，公司全称 Z&Z，又称 Z 氏公司。

² Git:一个开源的分布式版本控制系统，用以有效、高速的处理从很小到非常大的项目版本管理，由 Linus Torvalds 开发。详情访问 <http://github.com>。

³ Gitcafe:一个在线托管软件项目的服务平台,可以通过 git 来将他们所写的开源或商业项目的代码托管在 GitCafe 上，与其他程序员针对这些项目在线协作开发。详情访问 <http://gitcafe.com>。

开发在 win7 下使用 eclipse 完成。



5、开发过程

开发方式完全按照项目进度的标准化流程进行。在版本 **release1** 中只有爬虫的算法框架，**release2** 中添加了图形界面，**release3** 中又添加了一些比较新颖的功能。开发过程层次清晰，比较顺利。

6、程序结构

程序由 8 个 package 组成：[总计超过两千行代码]

package main: 主程序入口（14 行）
package crawler: 爬虫内核的总体操控程序 class Configuration: 保存程序相关的配置变量和常量（41 行） class WebCrawler: 负责内核程序流程控制及各组件的调度（185 行）
package fetcher: 获取网页 HTML 内容 class HtmlFetcher: 对于给定的 URL 获取其编码方式及 HTML 内容（123 行）
package parser: 分析获得的 HTML 内容 class Parser: 分析给定的内容字符串，获得其中的 URL； 获得电子邮箱；获得手机号码； 文档、音频、图片、种子、exe、压缩包的资源地址； 去 HTML 标签；计算网页重要度。（339 行）
package url: 存放一条 URL 关联信息 class WebUrl: 提供 URL 的相关操作（42 行）
package queue: 维护待爬取的 URL 队列 class UrlQueue: 维护当前待爬取的 URL 队列及已爬取的 URL 键值对（81 行）
package ui: 界面 AddDialog.java: 添加源网页的窗口（261 行） MyDialog.java: 一些提示窗口的实现（240 行） ViewTable.java: 可视化的表单，呈现出搜索的结果（109 行）

MyFram.java: 主界面 (200 行)

Tool.java: 图片、按钮处理的小工具类 (87 行)

ButtonPane.java: 选择资源的按钮群 (188 行)

ProgressBar.java: 进度条 (103 行)

7、功能简介

(1) 指定爬虫保存下载内容的目录, 指定爬虫的线程数以及最大网页下载数, 指定爬虫爬取的网页源 (此指定方式非常灵活, 后面会有介绍)。

(2) 开始爬取 HTML 内容。

(3) 分析爬取的 HTML 内容, 获得其中的 URL 地址并爬取之, 电子邮箱、电话号码、图片、文档、音频等资源并记录之。

(4) 通过爬取的 URL 的相互关系计算每个 URL 的重要度, 爬取完毕后可给出重要度排名。

(5) 爬取的资源可以通过用户的选择进行呈现, 其中图片、文档、音频资源还可以进行下载到文件夹中。

二、 建立 jCrawler 的界面

0、 简介

JCrawler 的界面用 java 的 Swing 库实现, 符合 MVC 模式, 并且不依赖第三方的 gui, 具有界面平台无关的特点。为了完成这部分工作, 我们花费一天时间深入研究了 Swing。其中构造界面的顺序是: 设计组件模型模型、打出组件代码⁴、添加响应事件。事实证明这样的写法很有效率。

1、 主界面

JCrawler 的主界面主要包括工具栏、资源查看按钮和资源列表三大块, 工具栏用 JToolBar 实现, 资源查看按钮由放置在 JPanel 许多 JButton 实现, 资源列表由 JTable 放在一个 JScrollPane 里实现。

⁴ 没有使用组件拖拽设计的工具, 开发效率比较低, 但是带来的优点是代码的可读性和维护性都提高了。



(图 1: 主界面的三大部分, 注解已经在界面中打出)

主界面对应着 MyFrame.java, 继承了 JFrame 类, 其中这三大都是主界面的成员对象, 并且有自己的实现文件, 下面我们就来介绍这三大块部分。

2、 工具栏



(图 2: 工具栏以及它上面的八个按钮)

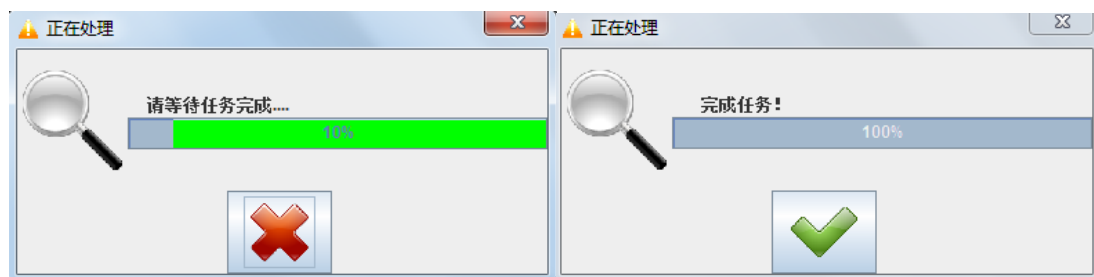
如上图所示工具栏包括了八个实用的按钮, 从左到右依次是开始 (开始进行爬取)、保存 (选择爬取内容的保存目录)、添加 (添加网页)、设置 (设置爬取

时的参数)、分析(统计分析爬取的内容)、下载(下载表单中选取的内容)、帮助(简要介绍使用的流程)、关于(列举版权声明和制作人等)。工具栏的每个按钮都采用了特殊的统一风格的图片,并且都有自己的 ToolTip,将鼠标放在上面即可看见相应的解说内容。


其中工具栏对象放置在主界面类中,但是按钮所触发的窗口放置在 MyDialog.java 以及 AddDialog.java 中。下面我们来介绍这八个按钮。

(1) 开始按钮

点击开始按钮后,会出现一个下载的进度条,进度条在 ProgressBar.java 中实现,继承了 JDialog 类。同时主界面类中实现了一个 progressBar 对象。



(图 3: 点击开始按钮后出现的进度条, 以及任务完成后的样子)

由于对 JDialog 的扩展,开始以后主界面是不会响应任何点击的,进度条在处理中和完成的界面也是有所不同的,其中在处理中时如果点  键就会终止所有的爬虫线程。

```
timer = new Timer(300, new ActionListener()
{
    public void actionPerformed(ActionEvent e)
    {
        int n = WebCrawler.getDownloadN();
        bar.setValue(n);

        if (n >= Configuration.maxHtmlDownloaded)
        {
            timer.stop();

button.setIcon(ImageTool.makeImageIcon("images/dialog/nav-prefs.png"));
            info.setText("完成任务!");
        }
    }
});
```

(图 4: 进度条的实现方式)

如上图，进度条的实现方式就是用一个 Timer 隔一段时间进行总任务量的查询，知道所有任务完成为止，如果终止爬虫，timer 也会提前结束。

(2) 保存按钮

保存按钮所激发后调用了 MyFrame 的 saveFloder() 函数。现实了 JFileChooser 实例化的 saveDialog 窗口。核心代码如下：

```
public void saveFloder()
{
    saveDialog.setFileSelectionMode(JFileChooser.DIRECTORIES_ONLY);
    saveDialog.setDialogTitle("选择你要保存下载内容的路径名");
    int result = saveDialog.showSaveDialog(this);
    if (result == JFileChooser.APPROVE_OPTION)
    {
        Configuration.savePath =
saveDialog.getSelectedFile().getAbsolutePath() + "/";
    }
}
```

(图 5: 保存按钮的相应事件)

通过改变 Configuration.savePath，导致后续的操作都执行在 savePath 路径下，可以说这样的一次选择就是相当于选择了一个 workplace。

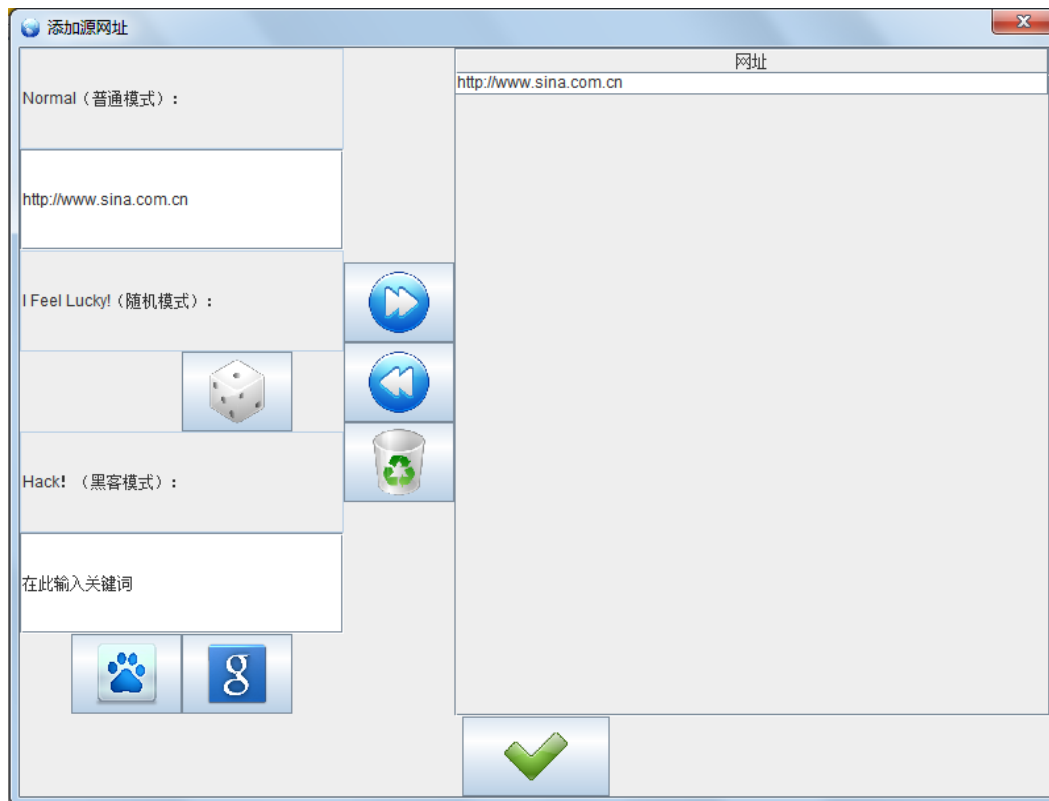
(3) 添加按钮

添加按钮所激发的窗口是一个比较复杂的窗口，实现了相当多的功能，它对应于 AddDialog.java，在 MyFrame 中有一个 addDialog 对象。

其中 AddDialog 由三大部分构成，一是左边的选择方式块，二是中间的选择操作块，三是右边的选择结果块。

选择方式块由相应的按钮标签构成，具有三种添加方式：1、普通模式：直接输入网页。2、随机模式：在内部预存的网址中随机出一个网页。3、黑客模式：模拟百度和谷歌的搜索方式，由键入的关键词生成相应的查询网页。

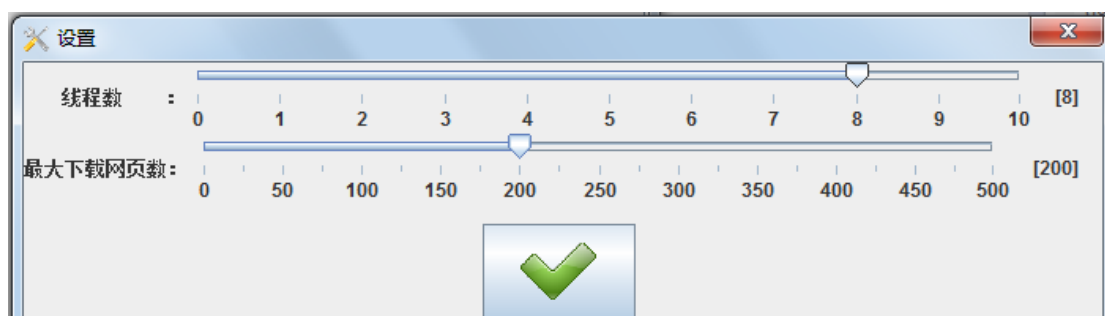
在选择方式块选择好需要添加的网页后，点击中间的选择操作块中的添加按钮，网页则会自动添加到右边的选择结果块中。在选择结果块中选择相应的网页然后点击退回的按钮就会实现所选的网页在选择结果块中被删除。同时还有一个一键清零的按钮，点击之后选择结果块中的所有网页均被删除。



(图 6: 添加窗口 (上) 以及其相应的激发窗口 (下))

在键入网址时还会对其进行正则匹配, 如果网址发生错误还会弹出上图的提示窗口。

(4) 设置按钮

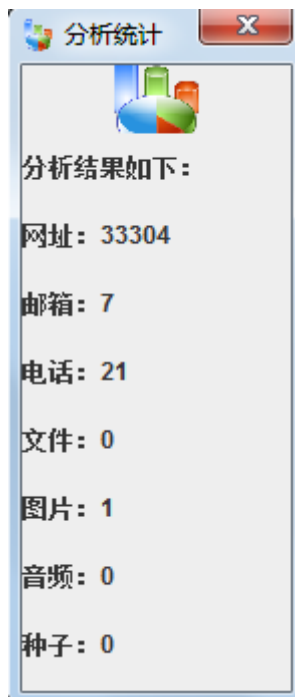


(图 7: 设置按钮激发的窗口)

设置按钮所激发的窗口对应于 MyDialog.java 中的 SettingDialog, 在 MyFrame 中有一个 settingDialogg 对象。里面最核心的部分是两个 JSlider, 对这两个 JSlider 进行读数后直接改变 Configuration 中的 maxThread 和 maxHtmlDownloaded 来改变相应的爬虫参数。

如果 slider 拉到零还会弹出类似图 6（下）中的提示窗口。

(5) 分析按钮



分析按钮所激发的窗口对应于 MyDialog.java 中的 AnalyseDialog, 继承于 JDialog, 在 MyFrame 中有一个 analyseDialogg 对象。

具体执行步骤就是从 INFOs 文件中（由爬虫的分析步骤写入）读出相应的统计信息，然后加以显示。

不过显示的是当前所选择目录下的 INFOs，也就是该 workspace 的爬虫结果。

（图 8：分析按钮激发的窗口）

(6) 下载按钮

选中资源列表中的一些资源后，再点击此按钮，于是就会自动下载相应的资源到当前所选保存目录的 Downloads 文件夹中。

具体实现方法就是调用 ViewTable（就是资源列表的类）的 downLoad() 函数。

(7) 其他按钮

帮助按钮激活的窗口对应于 MyDialog.java 中的 HelpDialog, 在 MyFrame 中有一个 helpDialog 对象；关于按钮激活的窗口对应于 MyDialog 中的 AboutDialog, 在 MyFrame 中有一个 aboutDialog 对象。两个都是 JDialog 继承而来的。由于帮助和关于按钮所激发的窗口都只有简单的文字与图片，因此在这里不多进行介绍。

3、 资源列表

资源列表对应于 ViewTable.java, 在 MyFrame 中有一个实例化的 downloadTable 对象, 继承于 JTable。改写了几个函数, 还添加了诸如下载之类的方法, 以方便我们的操作。

排名	网址	引用系数
1	http://news.sina.com.cn/guid...	1.961129000608502
2	http://corp.sina.com.cn/eng/	1.7233116565211737
3	http://corp.sina.com.cn/chn/	1.7217896810591453
4	http://sports.sina.com.cn/	1.6926853789207676
5	http://www.sina.com.cn/	1.6827155952323571
6	http://www.sina.com.cn/intro/...	1.6573661796680752
7	http://emarketing.sina.com.cn/	1.6249349931380148
8	http://www.sina.com.cn/cont...	1.5654017986285398
9	http://corp.sina.com.cn/chn/s...	1.521403335263334
10	http://video.sina.com.cn/	1.2963438046343647
11	http://tech.sina.com.cn/	1.1942632479158384
12	http://www.sina.com.cn/intro/...	1.182012209463764
13	http://news.sina.com.cn/	1.1206942319279165
14	http://tech.sina.com.cn/focus...	1.1141343244061466
15	http://weibo.com/n/	1.043650793650793
16	http://english.sina.com	0.8831504294058963
17	http://english.sina.com/	0.7805620779415531
18	http://m.sina.com.cn	0.7468004218004218
19	http://blog.sina.com.cn/	0.7212261336689862
20	http://video.sina.com.cn	0.7148149910389343
21	http://search.sina.com.cn/?c...	0.6207015727965653
22	http://help.sina.com.cn/	0.613028653488177
23	http://video.sina.com.cn/mov...	0.5742180358031598
24	http://edu.sina.com.cn/	0.571931321484548
25	http://ent.sina.com.cn/	0.5713132145943359
26	http://video.sina.com.cn/news/	0.5542875442443771
27	http://book.sina.com.cn/	0.5394773055293377
28	http://upload.you.video.sina...	0.5247949137938781
29	http://members.sina.com.cn/...	0.5219326175118179
30	http://m.weibo.com/web/cell...	0.5155038759689923
31	http://corp.sina.com.cn/chn/s...	0.5107418336180567
32	http://www.adobe.com/go/ge...	0.5018456967402504
33	http://www.dqwsgroup.com/	0.5010422094841063
34	http://e.weibo.com/sportsch...	0.5005141388174807
35	http://app.sina.com.cn/appd...	0.5
36	http://m.weibo.com/web/cell...	0.4642857142857142
37	http://qing.blog.sina.com.cn/t...	0.46210317460317346
38	http://video.sina.com.cn/spor...	0.4588810238248226

(图 9: downloadTable 在 pagerank 显示方式下的结果)

downloadTable 能够对应相应的资源查看按钮显示相应的资源, 因此会有不同的 downloadTable 的图案, 上面列举出的是 pagerank 时的图样。在显示其他资源的时候同时也会显示相应的来源网页。

4、 资源查看按钮



(图 10: 资源查看按钮群)

三、爬取网页和网页上的资源

1、HTML 下载

使用字节流下载。

为了判断网页编码, 在正式下载之前, 用默认编码预读 HTML 内容的前 50 行(猜测 charset 信息一定会在 50 行内出现)并正则匹配 HTML head 中的 charset 信息, 如果获得了匹配结果, 则使用该 charset, 否则使用默认的 UTF-8。然后进

行正式读取，并以字符串形式返回 HTML 内容。

2、HTML 分析

所有内容分析都是用正则表达式匹配完成，相应的正则表达式都保存在 Configuration 中。

3、URL、URL 分析内容的判重

都使用 ConcurrentSkipListMap 保存已获得内容，查看是否已经存入来进行判重。

四、 下载指定资源

在版本 release3 中我们还添加了这样的一个功能，那就是在选取资源列表中相应元素的时候，再点击工具栏中的下载按钮，相应的资源能够下载到当前选择的目录的 Downloads 文件夹中。

五、 改进空间

1、 分词

原计划成词语分析并完成搜索功能，于是尝试使用了一个开源的分词器 IKanalyzer，但是发现对于网页内容中的中文支持效果相当差，遂放弃。有可能可以找到一个更好的替代方案。

2、 表单的扩展

列表中的元素可以选择排序方式，可以直接双击下载打开（不用原来的工具栏按钮），同时也可以导出到 excel 文件中。

3、 搜索资源的扩展

这个的扩展相当容易只要添加与原来类似的代码即可。

六、 体会

这次的大作业使我们熟悉的网络编程以及 swing 编程，深入使用了 java 这门语言，虽然只有两个人，但是我们项目的质量还是很高的。希望以后 java 这门语言能够多多派上用场，因为语言只有真正地使用过了才能够深入地学习。