**存档资料 成绩：**

**华东交通大学理工学院**

**课 程 设 计 报 告 书**

**课程名称 Pyhton数据库编程**

**题 目 爬虫（采集）**

**分 院 电信分院**

**专业班级 19计科三班**

**学 号 20190210230313**

**学生姓名 刘海**

**指导教师 吴忠胜**

**2020年6月28日**

目录

[第一章 绪论 3](#_Toc1819)

[1.1 研究背景 3](#_Toc2698)

[1.2 研究意义 4](#_Toc21336)

[1.3 研究目标 4](#_Toc7122)

[第二章 系统结构 4](#_Toc17702)

[2.1 主题介绍 4](#_Toc16016)

[2.2数据流图 5](#_Toc14611)

[第三章需求分析 5](#_Toc117)

[3.1功能需求分析 6](#_Toc1018)

[（1）网页下载功能 6](#_Toc12576)

[（2）网页分析能力 6](#_Toc29826)

[3.2系统性能分析 6](#_Toc13325)

[3.3系统各功能流程图 7](#_Toc22983)

[（1） 主模块功能流程图 7](#_Toc30520)

[（2） 功能模块流程图 9](#_Toc32761)

[第四章 详细设计 10](#_Toc23465)

[4.1 开发工具 10](#_Toc23346)

[4.2 系统设计原则 10](#_Toc1917)

[4.3系统数据库设计 11](#_Toc6619)

[4.3.1 概念结构设计 11](#_Toc5424)

[4.3.2 数据库E-R图 11](#_Toc25429)

[4.3.3 数据库表的设计 12](#_Toc29229)

[4.4测试与结果 12](#_Toc27837)

[第五章 系统实现 13](#_Toc3828)

[5.1代码实现 13](#_Toc7431)

[5.2运行界面 15](#_Toc21845)

[致谢 1](#_Toc712)

[参考文献 2](#_Toc15754)

# 第1章 绪论

## 1.1 研究背景

近年来。随着网路应用的逐渐扩展和深入，如何搞笑的获取网上数据已经成为公司和个人的追求，在大数据时代 谁掌握了更多的数据 谁就可以 获得更高的利益，而网络爬虫是其中最为常见的一种从网上获取数据的手段。

网络爬虫，即Web Spider，是一个很形象的名字，如果把互联网比喻成一个蜘蛛网，那么Spider就是在网上爬来爬去的蜘蛛，网络蜘蛛是通过网页的链接地址来寻找网页的，从网站某一个网页（通常是首页）开始，读取网页的内容，找到网页中的链接地址，通过链接地址寻找下一个网页，知道把所有网页抓取完毕，把整个互联网当成一个网站，那么网络蜘蛛就可以用原理把互联网上所有网页都抓取下来

在网络爬虫如此先进快捷的特点下，如何实现它。在面对对象的高级语言中，首选python，python是一种“解释型，面向对象的，带有动态语义的高级语言”，使编程者在编程使保持风格，写的程序清晰易懂，有广阔的应用前景

## 1.2 研究意义

通过本次课程设计，完成一个对百度新闻中游戏信息的爬取，让需要游戏新闻的人，更加快捷方便的获取新闻的内容，来将版本的更替和英雄的强度来进行匹配

## 1.3 研究目标

针对新闻的具体内容，和爬取方式，将新闻中的内容获取到数据库，使需要新闻内容的人更加方便快捷的，从繁重的工作中解脱出来,大大减轻了工作量,减少人为的工作失误，全面提高读取新闻的效率，从而使玩家对游戏的更新内容的认知跃上一个新的台阶。

# 系统结构

## 2.1 主题介绍

提示框：来让用户来判断 爬取的内容是否 事宜

信息：在用户确定之后能够爬取到 ，然后存储到数据库中

## 2.2数据流图

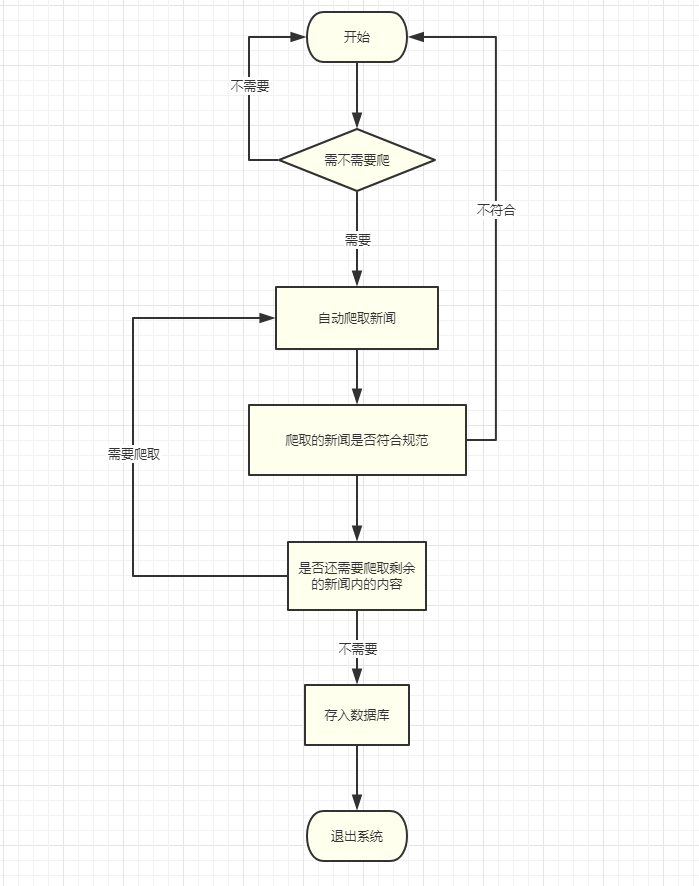


图 2-2-1

# 第三章需求分析

网络爬虫是一个专门从万维网上下载网页并分析网页的程序，它将下载的网页和收集到的信息存储到本地数据库中以供搜索引擎使用，网络爬虫的工作原理是从一个或若干初始网页的链接开始进而得到一个链接队列。伴随网页的抓取又不断的从网页里获取新的链接放到链接队列中，知道爬虫程序满足系统的某一条件时停止，它是搜索引擎最重要的组成部分。

## 3.1功能需求分析

### （1）网页下载功能

① 能够下载任何http协议和https协议的链接的网页。

② 构造HTTP请求中的GET请求。

③ 分析HTTP相应请求。

### （2）网页分析能力

① 提取网页标题

② 提取网页关键字

③ 提取网页摘要

④ 提取网页链接并统计数量

⑤ 把新链接加入到URL队列中

（3）保存功能

① 能够正确的保存网页以及网页信息到文件

②功能把系统运行中的异常写入日志文件。

1. 提示界面

① 管理爬虫功能

② 查看运行结果

## 3.2系统性能分析

当用户使用系统时，系统需要能够对于一些异常状态系统记录并跳过此异常继续执行。系统需要具有较高的可移植性和可靠性，系统需要具有很好的可测试性和可维护性。网络爬虫系统时不停地从万维网上下载网页和采集网页信息的系统，由于网络爬虫系统时搜索引擎的组成部分，搜索引擎要利用到爬虫系统的信息，所以系统要设计合理的存储文件并简历索引

（1）网络爬虫系统 流程如图 3-2

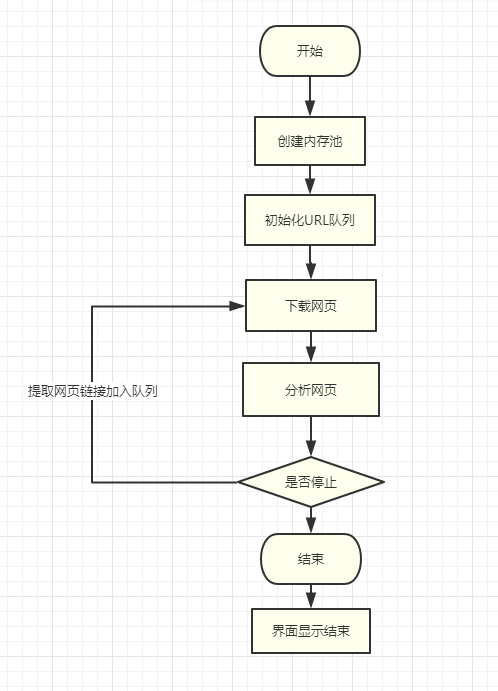


图3-2

## 3.3系统各功能流程图

### 主模块功能流程图

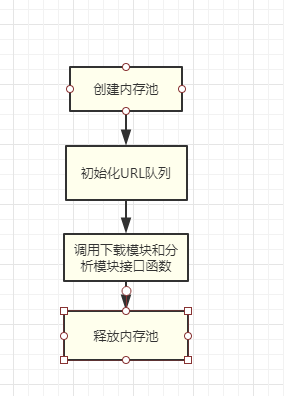


图3-4-1

需要说明的问题：

①指向内存池的是一个全局变量指针

②初始的URL必须从文件中读取，然后用字节流读取出来

③必须有全局变量指向URL的 队列头

### 功能模块流程图

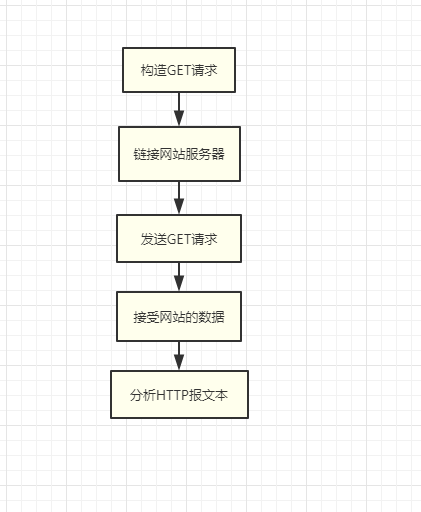


图3-4-2

需要说明的问题：

①利用系统函数把网页内容读入内存

②利用表达式来提取相关信息

③把具体信息分配出来 获取想要的信息

④把一个网页的URL写入数据库保存

# 第四章 详细设计

## 4.1 开发工具

3.1.1 Chrome

Google Chrome，又称Google浏览器，是一个由Google(谷歌）公司开发的网页浏览器。该浏览器是基于其他开源软件所撰写，包括WebKit，目标是提升稳定性、速度和安全性，并创造出简单有效率的使用者界面。软件的名称是来自于称作Chrome的网络浏览器图形使用者界面(GUI)软件的 beta测试版本在2008年9月2日发布，提供50种语言版本，有 Windows、Mac oS X、linsx、bndicidl、以及iRS版本提供下载。"a0)1.3年9月，C.hcne已达全球份额的4尝.?%，成为全球使用最广的浏览器~

在本次课程设计中，Chrome不仅是查找网页的工具，也是审阅Html寻找关键<tag>的重要工具。

3.1.2 Pycharm

Pycharm是一种 Python IDE，带有一整套可以帮助用户在使用 Python 语言开发时提高其效率的工具，比如调试X语法高亮K Project管理,代码跳转、智能提示、自动完成、单元测试、版本控制。此外该IDE提供J文些高级功能高激用承夜持Django框架下的专业 Web开发。

本次课程设计项目中，Pychram 承担了全部的Python开发工作，其优秀的自动关联和记忆关键词功能，很好的提高了效率，减少不必要的错误。

## 4.2 系统设计原则

通过对凤凰网网页的 html格式分析，来具体获取想要的字段的内容，在对获取到的网页的内容进行分析和分类，将新闻信息全部解析到字节流中。后 将数据存入数据库中，使 使用者更加方便的能够获取和分析新闻内容，提高获取信息的能力和效率

## 4.3系统数据库设计

### 4.3.1 概念结构设计

针对本系统的需求，设计出如下面所示的数据项和数据结构：

新闻信息保存表：信息编号，信息标题，信息链接

### 

### 4.3.2 数据库E-R图

(1).新闻信息保存表E-R图

标题

Id

链接

MySQL数据库

### 4.3.3 数据库表的设计

(1). 新闻信息保存表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 长度 | 是否主键 | 描述 |
| id | Int |  | 是 | 自动编号 |
| Title | varchar | 20 |  | 标题 |
| LinkUrl | varchar | 20 |  | 链接 |

## 4.4测试与结果

对于界面的测试：能够通过界面把数据写入数据库中，并且能够通过数据库读取在文件上显示，在mysql的图形化界面中可以看出数据的变化，数据的删除和修改功能能够正常的使用，但对于异常的数据没有进行判断以致于不合法的数据也可以写入配置文件。

对于爬虫程序的测试：输入不合法的URL，可以在日志中显示出具体出错误。对于系统的测试：经过多次运行计算平均数值。得到系统的运行效率不是很高，大概每秒两个网页。测试移植性：把系统移植到其他的windows系统上。除开mysql系统的密码错误以外，更多的是缺少库函数 ，以及解析不成功出错。没有系统无响应，陷入死锁状态，经过多次测试和修改已经能够正常运行。

# 系统实现

## 5.1代码实现

import requests

from bs4 import BeautifulSoup

import pymysql

import tkinter as t

import tkinter.messagebox as tm

news\_title = []

news\_link = []

def s():

url = "http://news.baidu.com/game"

str1 = requests.get(url).content

soup = BeautifulSoup(str1, "lxml")

for x in range(1, 5):

path = "#col-left > div:nth-child(3) > div.bd > div > div.one-pic-pics > div:nth-child(1) > div > div > ul > " \

"li:nth-child({}) > a".format(x)

data = soup.select(path)

news\_title.append(data[0].text)

news\_link.append(data[0].get('href'))

def mysql():

db = pymysql.connect(host="localhost", port=3306, user="root", password="123456", database="bbs", charset="utf8")

if db:

print("连接")

cursor = db.cursor()

cursor.execute("drop table if exists kcsj")

sql = """create table kcsj(

id int unsigned primary key auto\_increment,

title varchar(5000),

linkUrl varchar(5000))

"""

cursor.execute(sql)

sum = len(news\_title)

for d in range(0, sum):

sql = "insert into kcsj (title,linkUrl) values('{}','{}')".format(news\_title[d], news\_link[d])

cursor.execute(sql)

db.commit()

num = cursor.rowcount

db.close()

def create\_windows():

w = t.Tk()

w.geometry('1x1')

ask\_1 = tm.askyesno('提醒', '即将进行爬取操作，是否继续？')

if ask\_1:

s()

ask\_2 = tm.askyesno('提醒', '是否将信息存入数据库')

if ask\_2:

mysql()

ask\_3 = tm.askyesno('提醒', '信息已爬取进数据库')

w.mainloop()

def main():

create\_windows()

if \_\_name\_\_ == '\_\_main\_\_':

main()

## 5.2运行界面

（1）爬取功能：

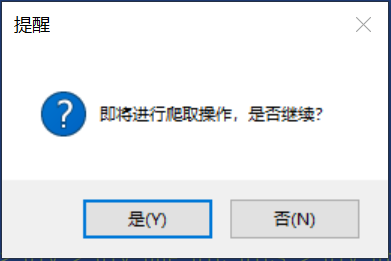


图5-2-1

（2）连接MySQL数据库功能：

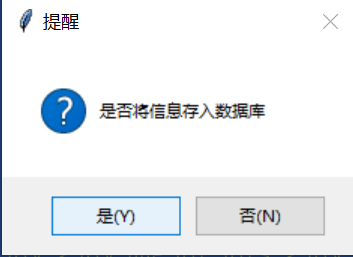


图5-2-2

（3）存储功能

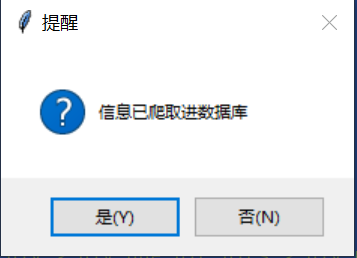


图5-3-1

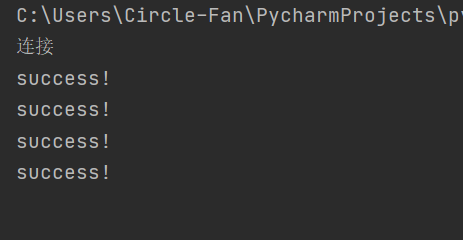


图5-3-2

# 致谢

这是我第一次用Python爬取新闻信息，对于爬虫而言早有耳闻，在java的学习过程中也有一点触碰，在大二上学期也学习过Java的爬虫。对于两个语言写爬虫的方便程度来说，Pyhton实在是让我耳目一新，极致的简单和方便是最让我影响深刻的，在课程设计之前我对爬虫停留在一知半解的程度，在自己和伙伴做爬虫的过程中，让我对Python有了很多新的理解，对于编程的集合也有了更深层次的理解。

在爬虫的过程当中，很明显的可以感觉到，互联网时代任何东西都可以看成资源，一个网站就是HTML+CSS组成，图片就是xxx.jpg文件，无数的资源在互联网上，通过URL来访问资源，在做课设的过程中，检验了我一学期所学习的知识，并且也培养了我做事有规则，有方向，先定义好目标再去做的习惯。在课程设计中，和伙伴积极分配任务和主动去寻找问题的答案。都是在这次课程设计中所得到的东西，并且会伴随我一生。

他又一次将大学大部分学习的东西整合起来，比如数据库，与网页制作，在制作爬虫的过程中都有利用到，但是正是通过此次的课程设计让我知道了我的不足，知道了我的哪些有些知识还是没学好，让我又重新端起自己的课本重拾自己以前的知识，让我查缺补漏。再是让我学会如何将课本自己学会的知识运用到实际问题来，真正应用到python爬虫的爬取网页，数据库的构建维护还有网页的开发。

这次的课程设计终于顺利结束，在设计中遇到了很多的编程问题，最后在吴老师的辛勤指导下，终于迎刃而解。同时，在吴老师的身上我学到了很多实用的知识，在此表示由衷的感谢，也感谢我的同伴对我的鼓励让我没有放弃。

# 参考文献

[1] 百度文献

[2] Python语言程序设计，电子科技大学出版社

[3]嵩天，Python语言程序设计基础，高等教育出版社，2017

[4][美]埃里克.马瑟斯，Python编程从入门到实践，人民邮电出版社，2020十月第二版

[5]明日科技，Python从入门到精通，清华大学出版社，2021升级版