江苏师范大学

JIANGSU NORMAL UNIVERSITY

本科毕业设计

UNDERGRADUATE DESIGN

设 计 题 目： 网上舆情爬取系统的设计与实现

姓 名： 李林

学 院： 智慧教育学院(计算机科学与技术学院)

专 业： 软件工程

年 级、学 号: 2012级、12267067

指 导 教 师： 董永权

江苏师范大学教务处印制

设计原创性声明

本人郑重声明：所呈交的毕业设计，是在导师的指导下，独立进行研究所取得的成果，所有数据、图片资料真实可靠。除文中已经注明引用的内容外，本设计的研究成果不包含他人享有著作权的内容。对本设计所涉及的研究工作做出贡献的个人和集体，均已在设计中以明确的方式标明。本设计的知识产权归属培养单位。

本人签名：　　年　　月　　日

设计版权使用授权书

本设计“网上舆情爬取系统的设计与实现”是本人在校期间所完成学业的组成部分，是在江苏师范大学教师的指导下完成的，因此，本人特授权江苏师范大学可将本毕业论文的全部或部分内容编入有关书籍、数据库保存，可采用复制、印刷、网页制作等方式将论文文本和经过编辑、批注等处理的论文文本提供给读者查阅、参考，可向有关学术部门和国家有关部门或机构呈送复印件和电子文档。本毕业论文无论做何种处理，必须尊重本人的著作权，署明本人姓名。

作者签名：　　指导教师签名：

年　　月　　日　　年　　月　　日

网上舆情爬取系统的设计与实现

摘　要

随着计算机技术的迅猛发展，网络已成为人们对不同社会问题发表看法的重要场所，互联网已成为广大人民群众思想文化信息传播的集散地，网络舆情呈现了多样化的趋势。为了进行正确的舆论导向，网络舆情的监控势在必行，而爬取系统正是其中重要组成部分。本系统针对这一需求进行设计，使用 B/S 架构，选用 Python 语言和 MySQL 数据库进行开发。网上舆情爬取系统总共包括两大模块：前台展示模块和后台爬取模块，其中前台展示模块包括四个部分：帖子展示、帖子搜索、敏感词管理和 URL 设置。后台爬取模块包括两个部分：帖子爬取和帖子存储。本系统具备一定的使用价值，能够稳定运行，帮助用户了解最新舆情，为网络舆情的监控奠定基础。

该论文有图 25 幅，表 2 个，参考文献 20 篇。

关键词：网上舆情爬取系统　舆情爬取系统　爬取系统

Design and Implementation of Crawling Public Online Opinion System

Abstract

With the rapid development of computer technology, the network has become an important place where the people express the views of different social issues. The Internet has become a distribution center for the crowds and the internet public opinion presents the trend of diversification. For the correct guidance of public opinion, public opinion monitoring network is imperative, and the crawling system is an important component. For the needs of development, the system uses B/S architecture, chooses Python language and MySQL database. Online public opinion crawling system comprises a total of two modules: the foreground display module and background crawling module. The display module consists of four parts: the display of posts, the search of posts, the management of sensitive words and the setting of URL. Background crawling module consists of two parts: the storage of posts and the crawling of posts. This system has some value and stable operation. It can help users learn about the latest public opinion, and can lay the foundation for the Internet public opinion monitoring.

This paper consists of twenty-five pictures, two tables and twenty references.

Key Words: Crawling Public Online Opinion System; Public opinion crawling system; Crawling System

目 录

图清单

表清单

1 绪论

1.1 课题背景及研究意义

1.1.1 课题背景

随着计算机技术的应用和发展，网络已经普及到千家万户，人们越来越习惯于在网络上发表自己的看法、观点等，网络舆情也随之迅速兴起。由于每个人的观点和看法不同，所以网络舆情呈现了多样化的趋势，同时网络舆情也越来越复杂，更加难以控制。

随着网上舆情的深入发展，需要一定的舆情监控措施。为了更加方便的监控网络舆情，进行正确的舆论导向，网上舆情爬取系统的开发迫切需要。

1.1.2 研究意义

网上舆情爬取系统的意义重大，主要有经济、文化和技术三方面的意义。从经济层面来看，本系统可以将爬取的数据进行整理分析，通过大量的数据洞察人们的需求，从而产生经济效益。从文化层面来看，通过爬取网上的舆情信息，国家可以进行正确的舆论导向，弘扬正确的文化观，对推动建设文化强国有一定的意义。从技术层面来看，本系统可以为爬取网络其他资源提供有效的示范作用，对于科学、合理的利用网络资源意义重大。

1.2 开发工具的选择及语言介绍

1.2.1 Python 简介

Python[1]是解释性的语言，具有强大的面向对象的特征。Python 有两个较为显著的特点：简洁性和粘合性。

首先介绍 Python 语言的简洁性，除了强制制表符以外，Python 的语法规范十分人性化，简洁清晰，一目了然，没有很多冗余的语法规则，方便新手很容易入门，这也是 Python 语言的一大优势。

其次介绍 Python 语言的粘合性，Python 语言可以结合其他语言的模块，比如 MATLAB 在建模方面非常出色，当 Python 生成了主要程序后，MATLAB 可以进行建模操作，然后打包成一个扩展库，Python 直接调用该库即可，这体现了

Python 语言的强大的粘合性，这也是 Python 语言被称为“胶水语言”的原因。

1.2.2 MySQL 数据库简介

MySQL[2]是一种关系型的数据库管理系统，在当今众多数据库中，MySQL 数据库的影响力仍是独一无二的，MySQL 的优势表现在其性能的优越，同时磁盘占用率低和出色的稳定性也是 MySQL 傲视群雄的一个重要的原因。MySQL 结构图如图 1-1 所示。

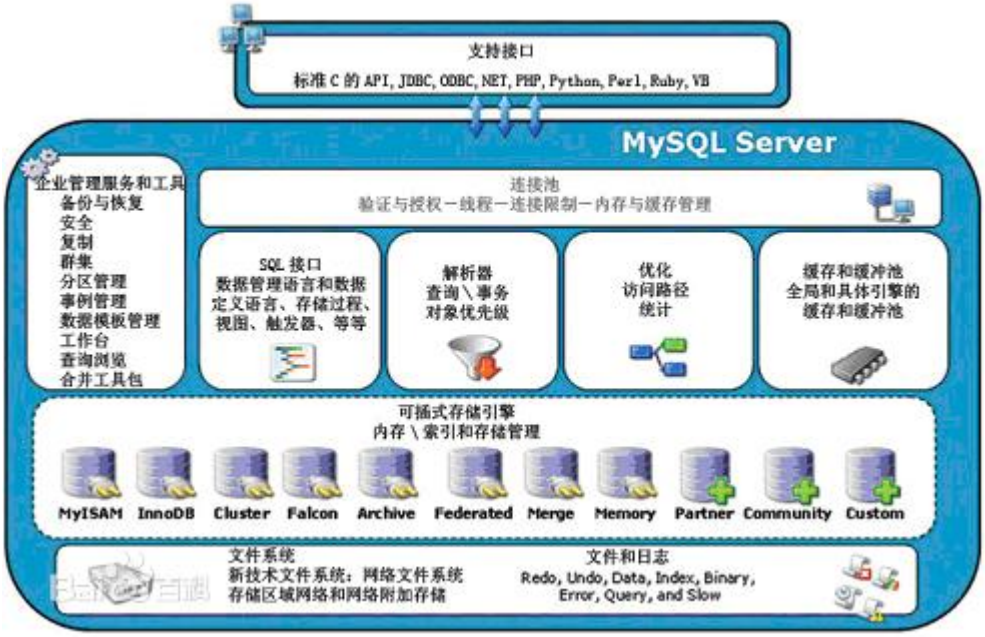

图 1-1 MySQL 结构图

1.2.3 开发工具及运行环境

操作系统：Microsoft Windows 10
开发环境：PyCharm[3]5.0.4，WampServer[4]2.5
数据库：MySQL 5.6.17

1.3 本文的主要内容和组织结构

本文主要大致介绍了网上舆情爬取系统的背景、研究意义、开发语言以及开发工具等。

本系统能够有效的爬取网络资源，首先选取了一个网站作为样例，通过 Scrapy 框架爬取了帖子的相关信息，其中包括发帖人（postMan）、发帖时间（firstTime）、帖子标题（title）、帖子内容（content）、帖子链接（link）、

阅读数量（readCount）和回复数量（replyCount），将爬取的信息存放至数据库。前台使用了 Flask 框架进行展示。用户可以直观的看到帖子的相关信息，可以通过图表来深入了解舆情动向，还可以通过搜索以及添加敏感词来查找自己感兴趣的舆论。

本文的章节内容安排如下：

第 1 章：绪论。主要详细描述了本系统的背景、意义、开发语言的选用及介绍、开发工具的选用，同时介绍了本系统的主要贡献和研究内容。

第 2 章：需求分析。主要介绍了本系统的需求，包括性能需求和相关功能需求。

第 3 章：系统功能模块设计。主要使用了图文的形式展示系统中各个模块的划分和数据库的设计与实现。

第 4 章：系统实现流程。主要介绍了系统前后台的各个功能模块，并且对模块的运行流程以及核心代码进行展示。

第 5 章：关键技术。主要介绍了本系统所采用的核心技术以及相关的配置。

第 6 章：总结与展望。

1.4 本章小结

本章要介绍了该系统的研究背景及意义、开发语言的介绍以及开发工具的选择和研究的主要内容和组织结构。

2 需求分析

2.1 功能需求

2.1.1 前台展示模块

1）帖子展示

首次访问主页面，用户可以看到爬取论坛的帖子（本文以“结合美”论坛 http://www.cxjhm.com/forum.php 为例），分页显示在主界面。

2）图表展示

该模块使用折线图对爬取到的帖子进行展示，折线图按照月份进行分类。

3）敏感词管理

用户可以添加自己感兴趣的敏感词，同时也可以删除不感兴趣的敏感词。

4）帖子搜索

用户可以根据自己的意向搜索感兴趣的帖子，同时也提供了搜索用户的功能。

5）URL 设置

用户可以设置自己感兴趣的 URL，重启程序，就可以根据输入的 URL 进行爬取。

2.1.2 后台爬取模块

1）帖子爬取

该爬取模块主要是将结合美上的帖子爬取下来，提供了发帖人、发帖时间、发帖内容、帖子标题、阅读和回复数量等信息，同时利用 Scrapy 框架循环爬取下一页帖子。

2）帖子存储

按照适当的格式存储数据库，同时添加去重功能，防止相同的帖子存入数据库。

2.1.3 用例模型

1）用例图（用户）

用户用例图描述了一个用户的操作权限。用户可以进行帖子展示、帖子搜索、

敏感词管理和 URL 设置，用户用例图如图 2-1 所示。

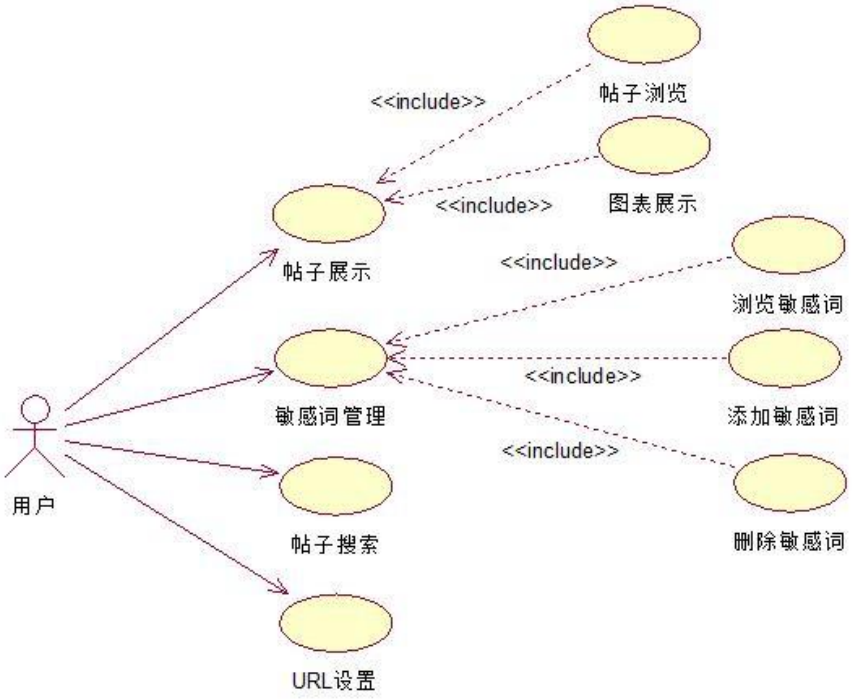

图 2-1 用例图（用户）

2）用例图（管理员）

管理员用例图描述了一个管理员的操作权限。管理员可以进行帖子爬取和帖子存储操作。管理员用例图如图 2-2 所示。

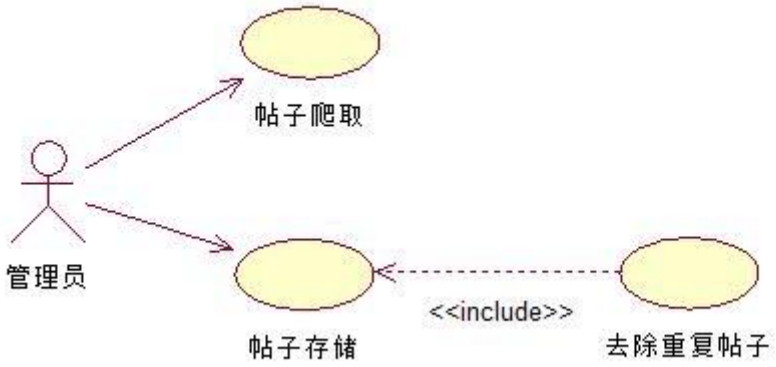

图 2-2 用例图（管理员）

2.2 性能需求

2.2.1 系统的软件环境

- 后台服务器。
  1）Windows 10
  2）python2.7.11
- 客户端计算机。
  1） Windows 10
  2） Chrome 49.0
- 数据库服务器。
  MySQL+WampServer

2.2.2 系统硬件环境

- CPU: Intel Core-i3
- 内存: 6GB
- 硬盘容量: 512GB

2.3 可行性分析

2.3.1 概述

可行性分析在系统开发过程中有着举足轻重的地位，可行性分析包括经济可行性分析和技术可行性分析，如果该项目无法通过成本效益分析或者在技术上无法实现，则该项目没有开发的必要，所以说可行性分析可以避免开发人员浪费大量的人力、物力和财力，只有通过了可行性分析，项目才可以实施，可行性分析是高效的开发项目必不可少的前提和重要的基础。

2.3.2 系统业务流程调查

在开发本系统前，本人进行了系统业务流程调查，从业务流程来看，系统是可行的。本系统的业务流程从大体上来说可以分为四个部分。第一个部分是帖子搜索，用户根据需求进行搜索。第二个部分是敏感词管理，用户可以根据需求添加敏感词。第三个部分是 URL 设置。第四个部分是爬取业务。帖子搜索业务如图 2-3 所示，敏感词管理业务如图 2-4 所示，URL 设置业务如图 2-5 所示，爬取业务如图 2-6 所示。

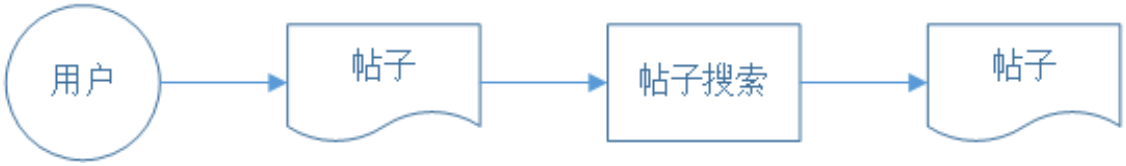

图 2-3 网上舆情爬取系统帖子搜索业务

图 2-4 网上舆情爬取系统敏感词管理业务

图 2-5 网上舆情爬取系统 URL 设置业务

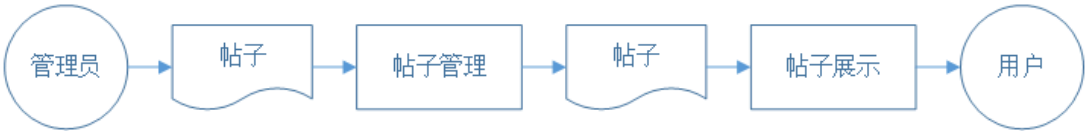

图 2-6 网上舆情爬取系统爬取业务

2.3.3 系统可行性调查

1）技术可行性：本系统采用的是 B/S 架构进行开发。这种架构以其良好的开放性、可扩展性以及共享性获得众多开发人员的青睐，也便开发人员日后的维护。此外，本系统采用目前最为流行的爬取框架 Scrapy 框架以及 beautiful soup[5]技术，维护起来方便简单。

2）经济的可行性：成本效益的分析是经济可行性分析当中最为重要的内容。当开发一个系统时，如果他在经济方面不适用，就完全不需要开发这个系统。本系统的开发只需要一台计算机，开发成本低廉，因此在经济方面是完全可行的。

2.4 本章小结

本章主要介绍了该系统的需求，其中包括软硬件环境，功能需求、性能需求以及系统可行性调查。

3 系统总体功能模块设计

3.1 系统功能模块的划分

本系统分前台展示模块和后台爬取模块。前台展示模块有五个模块：帖子展示、图表展示、帖子搜索、敏感词管理和 URL 设置。后台爬取模块包括两个模块：帖子爬取和帖子存储。

系统前台展示模块功能如图 3-1 所示。

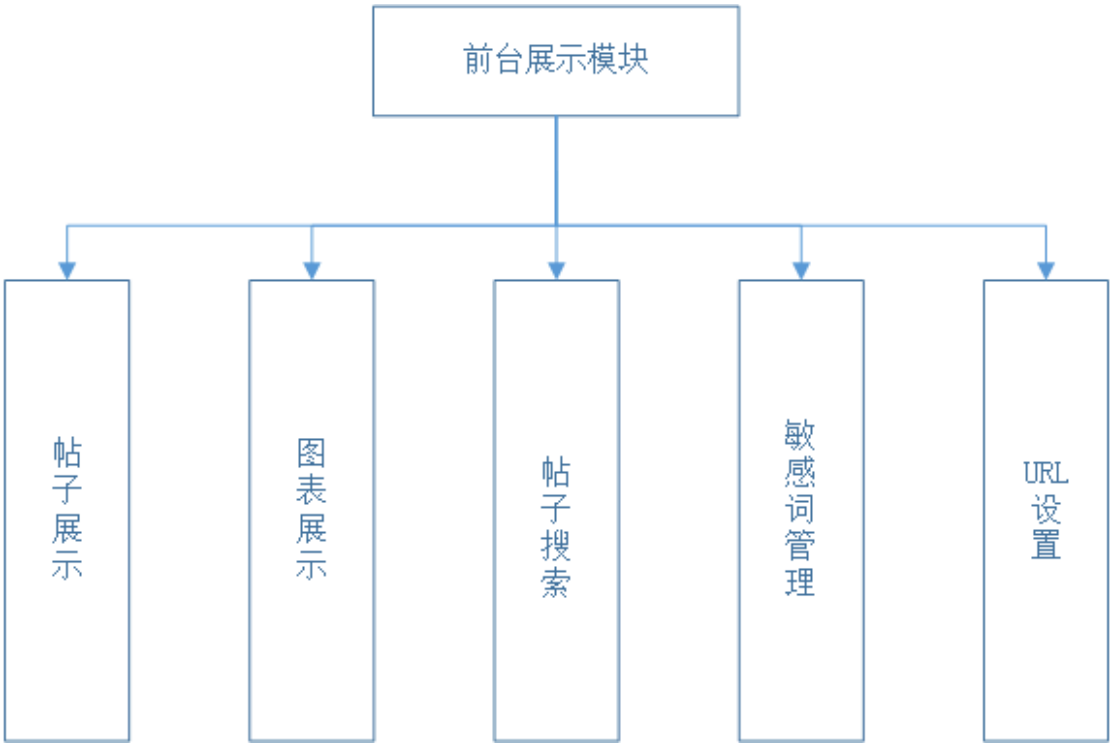

图 3-1 系统前台展示模块

系统后台爬取模块功能如图 3-2 所示。

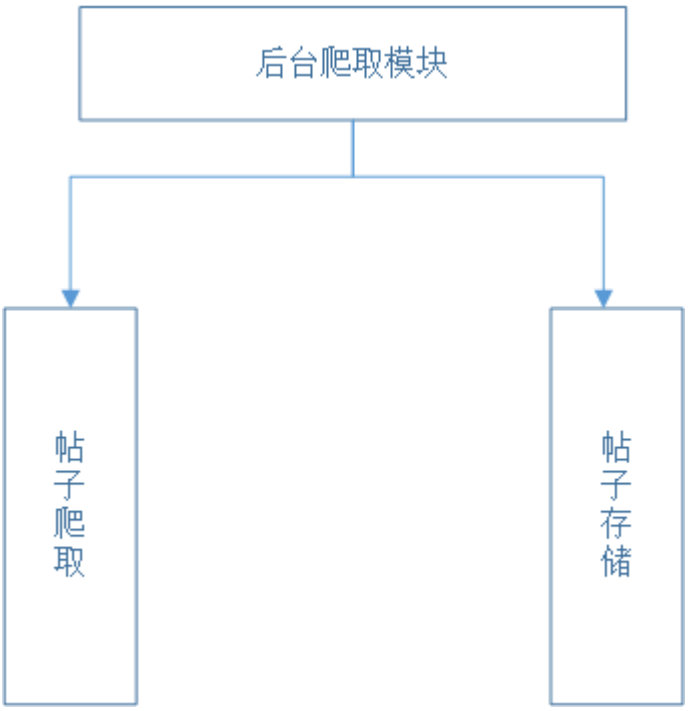

图 3-2 系统后台爬取模块

3.2 数据库设计

在设计爬取系统时，数据库的设计相当重要，数据库的设计关系到整个系统的设计，所以一个好的数据库是一个好系统的开端，本系统使用了当今非常流行的 MySQL 数据库，3.1 节已经介绍了本系统的需求分析和系统总体的设计，故数据库按照需求分析和总体设计进行，下面是数据库中一些关键的表。

3.2.1 实体

数据库中的实体可以指的是人也可以指的是物。经分析，本系统的实体主要有两类：帖子和敏感词。本系统的实体属性图如图 3-3 和 3-4 所示。

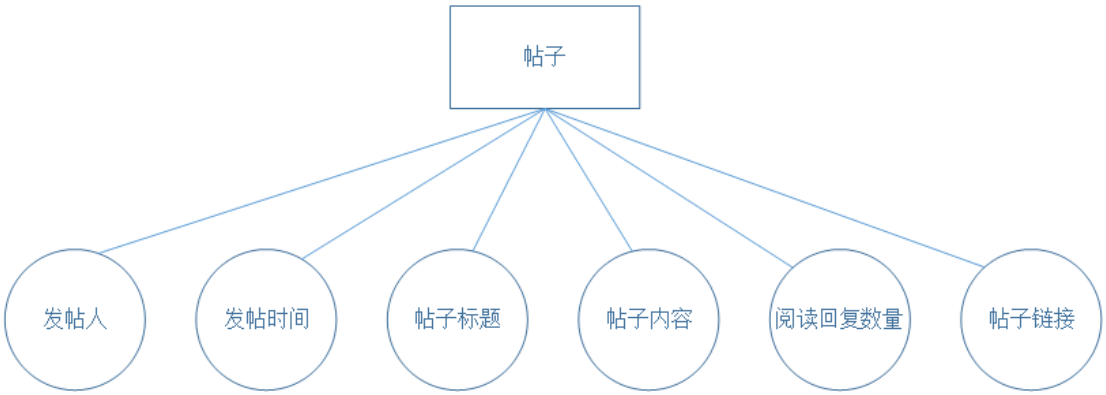

图 3-3“帖子”属性描述图

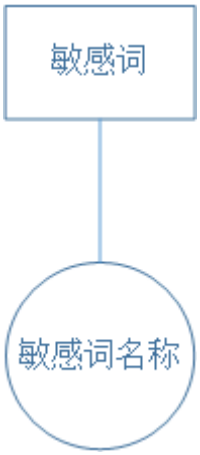

图 3-4“敏感词”属性描述图

3.2.2 关系模型

本系统的关系模型如下：

1）帖子表：（编号，发帖人，发帖时间，帖子标题，阅读和回复数量，帖子内容，帖子链接）

2）敏感词表（编号，敏感词名称）

3.2.3 数据库中的主要表结构

根据需求分析，系统使用的表如表 3-1 和 3-2 所示。

表 3-1 网上帖子表（jihemei）

列名	数据类型	类型长度	主键	允许空	默认值	说明
id	int	11	是	NO	Null	编号
postman	varchar	20		NO	Null	发帖人
firstTime	date			NO	Null	发帖时间
Title	text			NO	Null	标题
Content	text			Yes	Null	内容
readCount	int	11		NO	0	阅读次数
replyCount	int	11		NO	Null	回复次数
Link	text			NO	Null	发帖链接

表 3-2 敏感词表（sensitive_words）

列名	数据类型	类型长度	主键	允许空	默认值	说明
id	int	11	是	NO	Null	编号
word	varchar	20		NO	Null	敏感词名称

3.3 本章小结

本章阐述了系统的总体设计。主要对系统的功能模块进行了划分，同时对数据库的设计进行了详细的描述。

4 系统实现过程

4.1 帖子展示子模块

帖子展示子模块主要是将爬取的帖子用表格的形式展示在网页上，主要包括帖子列表页面和帖子详细页面。帖子列表页面主要展示发帖人(posMan)、帖子标题(title)和帖子链接(link)。帖子详细页面主要展示帖子标题、帖子作者和帖子内容。帖子列表页面如图 4-1 所示，帖子详细页面如图 4-2 所示。

Data Show

id	postMan	title	link
1	aksuenfy	michael kors包包 ALCAO Z6WI BNd5	文章链接
2	shekplos9y	Cheap NFL Jerseys China Wholesale	文章链接
3	pgyamywo	michael kors 單肩包 HoA54 c5ah mCI4	文章链接
4	东东	超市偷东西累了 男子吸根烟歇歇	文章链接
5	东东	企业及个人可用手机缴税	文章链接
6	东东	加快凝聚侨心侨力提升对外开放水平	文章链接
7	绿色青春	人社部:生育险和医疗险将合并	文章链接
8	绿色青春	完善安全监管体系推动安全产业发展	文章链接
9	绿色青春	徐州成"最文艺十大城市"	文章链接

图 4-1 帖子列表页面

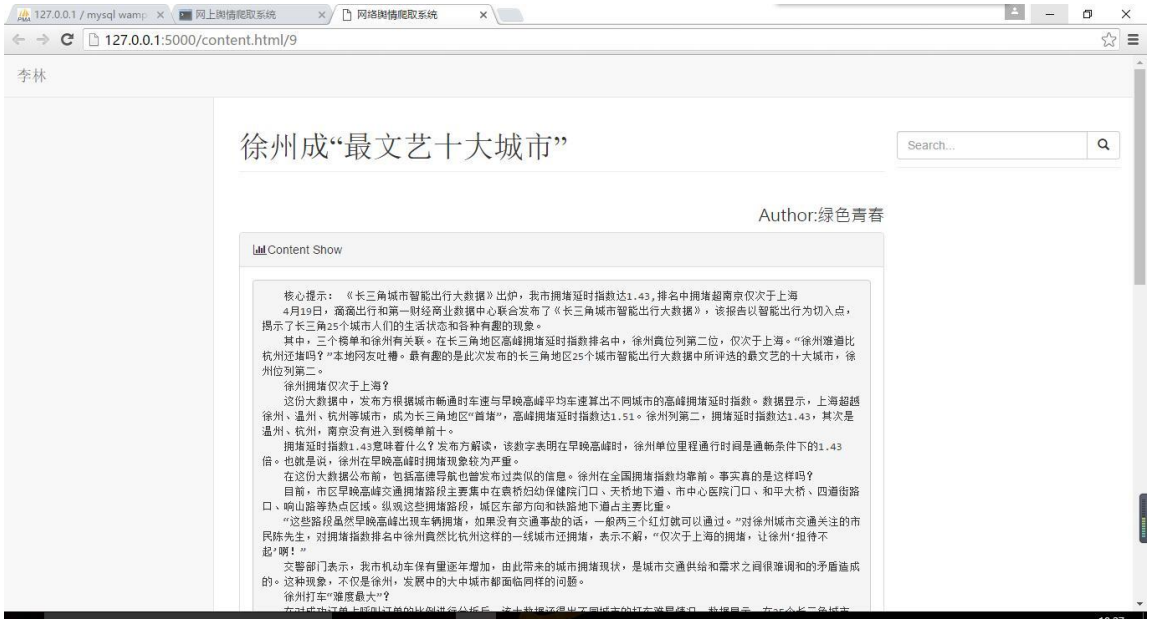

图 4-2 帖子详细页面

帖子展示的代码如下所示：

```
<table border="0" class="table table-bordered" id="bootstrap-table">
    <tr>
        <th>id</th>
        <th>postMan</th>
        <th>title</th>
        <th>link</th>
    </tr>
    {% for item in items %}
        <tr>
        <td id="getId">{{ loop.index }}</td>
        <td>{{ item.postMan }}</td>
        <td>{{ item.title }}</td>
        <td><a href="http://127.0.0.1:5000/content.html/{{ item.id }}"target="_blank">文章链接</a></td>
        </tr>
    {% endfor %}
</table>
```

4.2 图表展示子模块

在图表展示子模块中，照月份分类，统计出每个月份的帖子数量，使用折线图进行展示。图表展示如图 4-3 所示。

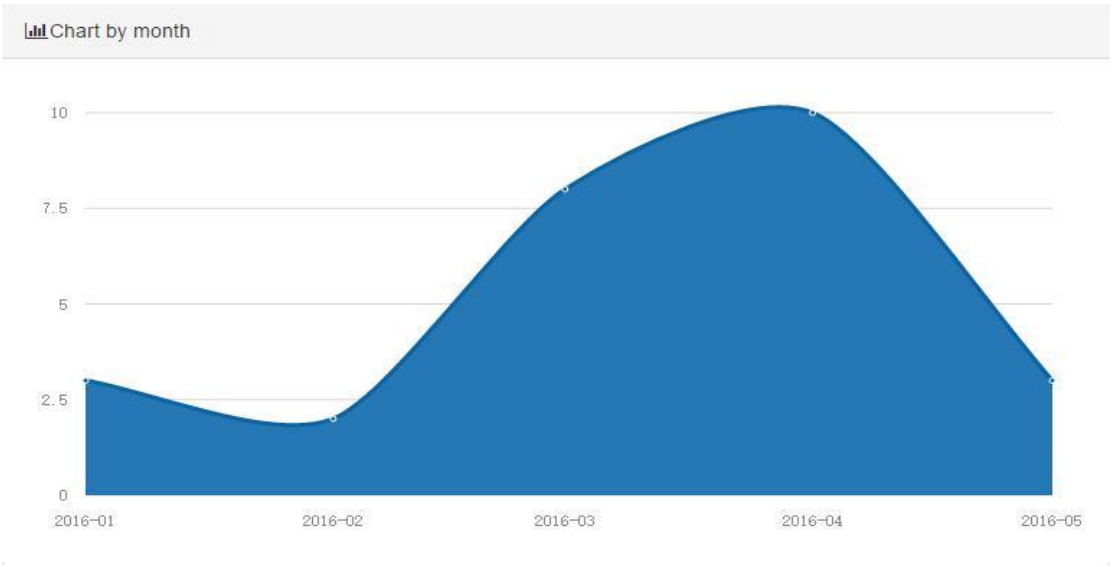

图 4-3 图表展示

4.3 敏感词管理子模块

敏感词管理子模块主要包括两个子模块：添加敏感词子模块和删除敏感词子模块。用户可以根据喜好定制自己喜欢的敏感词，方便查询相关的信息。

敏感词管理如图 4-4 所示。

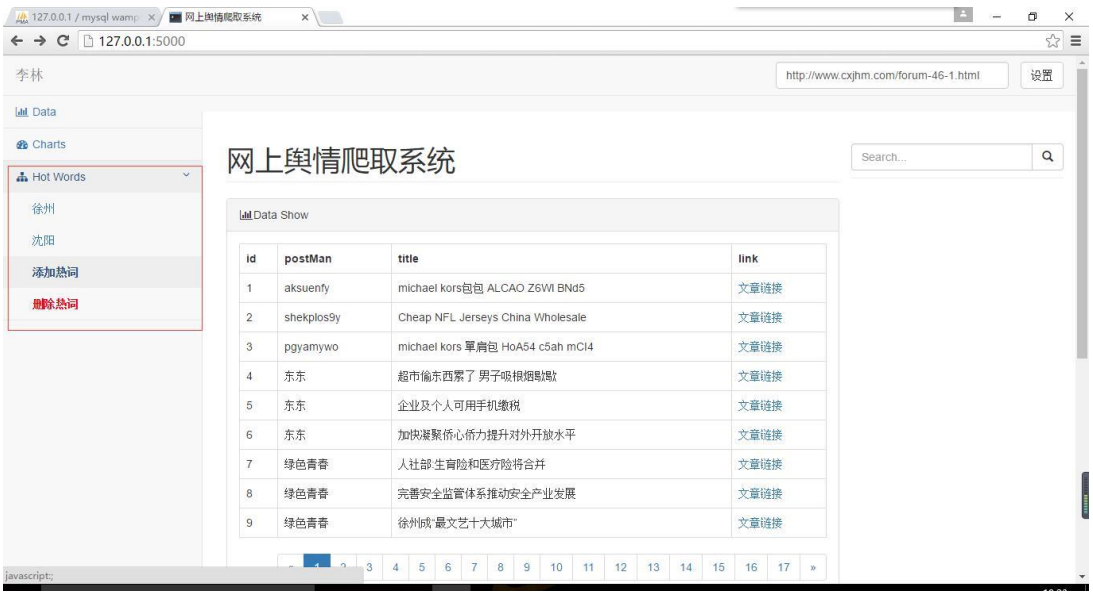

图 4-4 敏感词管理

4.3.1 添加敏感词子模块

添加敏感词子模块采用提示框的方式让用户输入。

添加敏感词如图 4-5 和 4-6 所示。

图 4-5 添加敏感词

图 4-6 添加敏感词成功

添加敏感词的代码如下所示：

```
@app.route('/add_word', methods=['POST', 'GET'])
def Add_Word():
    word = request.form.get('word')
    db.addWord(word)
    return jsonify(word=word)
```

4.3.2 删除敏感词子模块

删除敏感词子模块采用提示框的方式让用户输入

删除敏感词如图 4-7 和 4-8 所示。

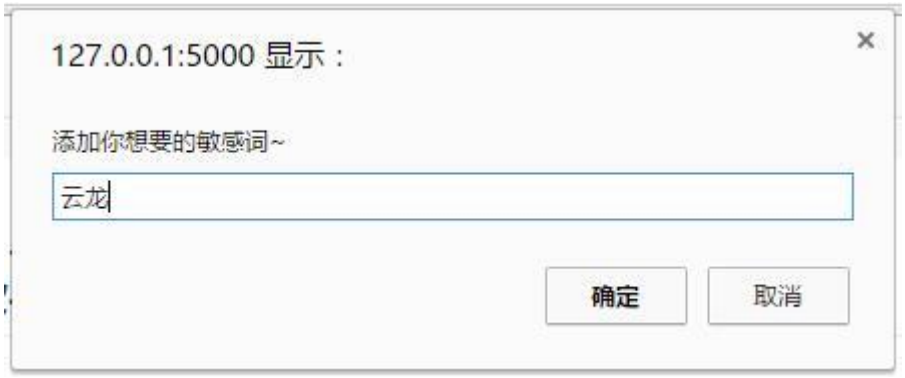

图 4-7 删除敏感词

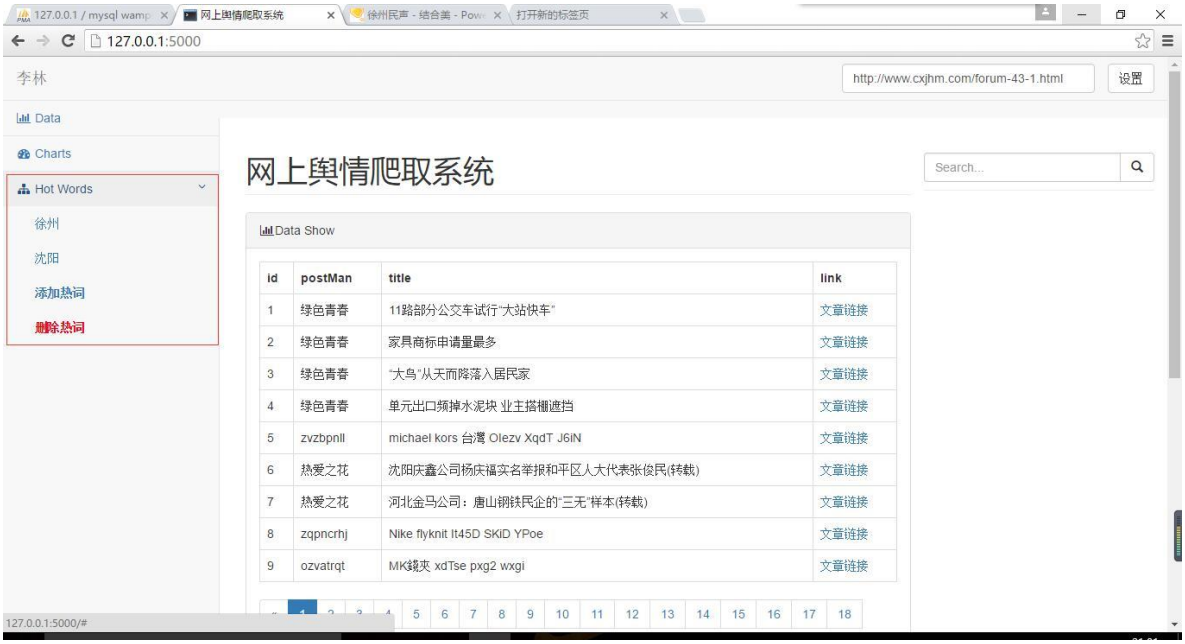

图 4-8 删除敏感词成功

删除敏感词的代码如下所示：

```
@app.route('/delete_word', methods=['POST', 'GET'])
def Delete_Word():
    word = request.form.get('word')
    db.deleteWord(word)
    return jsonify(word=word)
```

4.4 帖子搜索子模块

帖子搜索子模块放置在右上角，用户输入自己想查询的信息，后台会根据查询的信息搜索发帖人和帖子标题，将帖子展示在主页。

帖子搜索如图 4-9 所示。

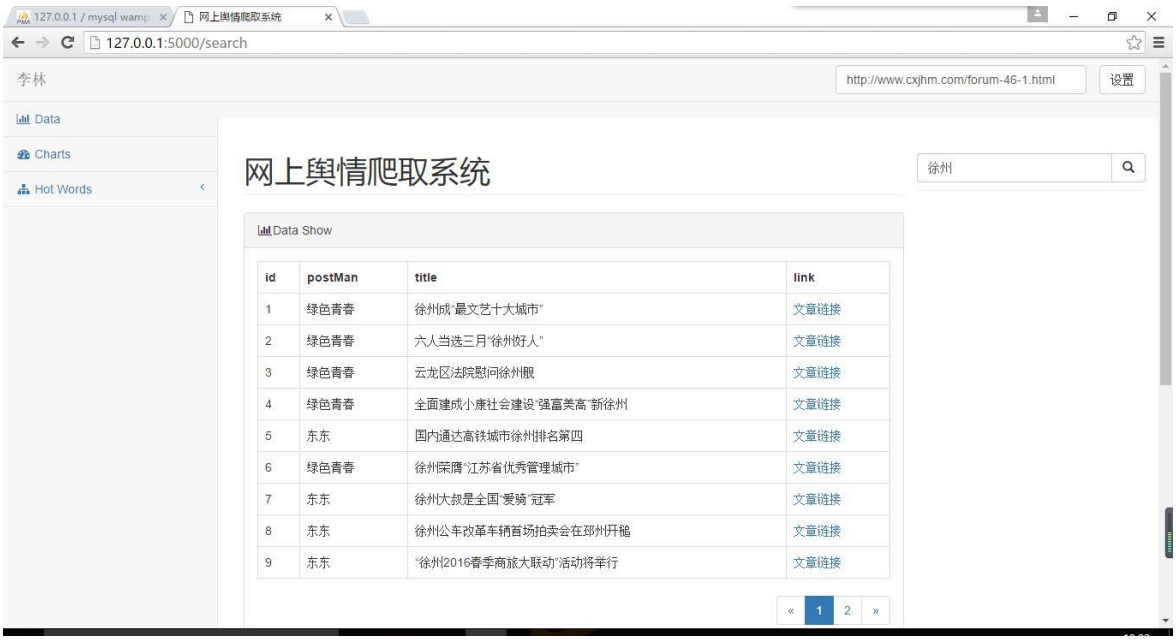

图 4-9 帖子搜索

帖子搜索的代码。

```
@app.route('/search/<word>', methods=['POST', 'GET'])

def Search_Word(word):

    items = db.searchTitle(word)

    words = db.getWord()

    URL = None

    with open('E:\Python Code\Code\Crawler\URLs', 'r') as f:

        URL = f.readline()


    return render_template('index.html',

                           items=items,

                           words=words,

                           URL = URL,

                           data=str(db.getCount_byMonth_byContent(word)))
```

4.5 URL 设置子模块

URL 设置子模块是让用户输入待爬取的 URL（在 http://www.cxjhm.com/域名之下），后台获取到 URL 之后，开始爬取用户输入的网页，刷新一下页面就可以获取到新爬取的数据。

设置 URL 模块如图 4-10 和 4-11 所示。

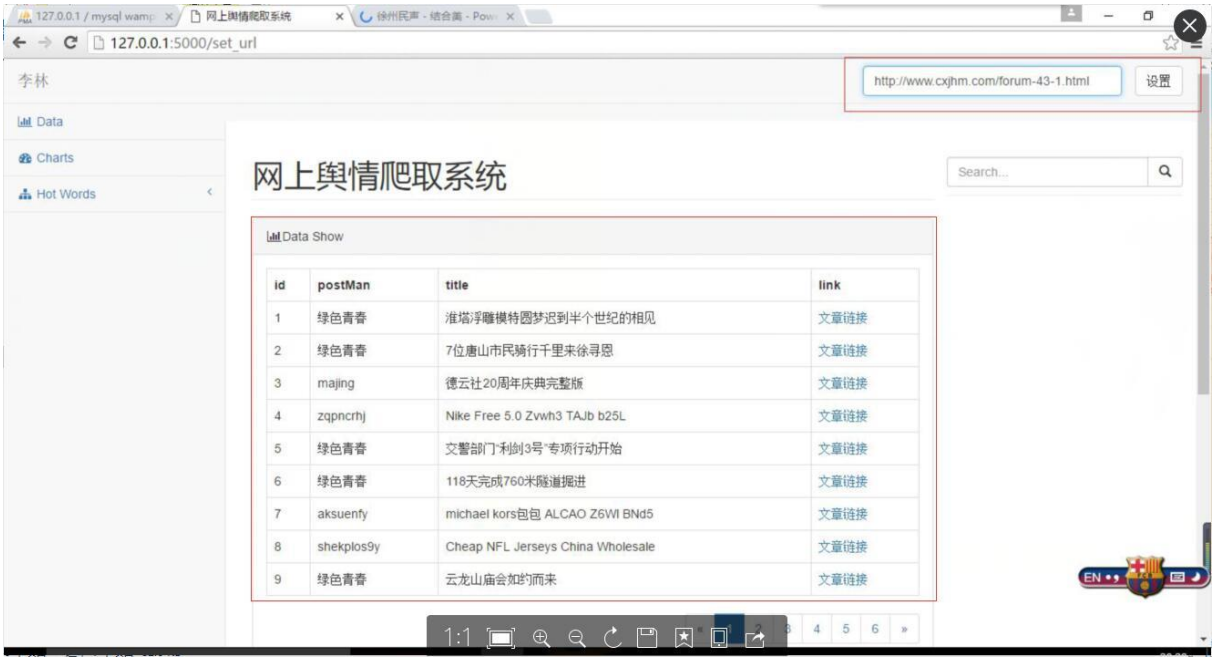

图 4-10 设置 URL 之前

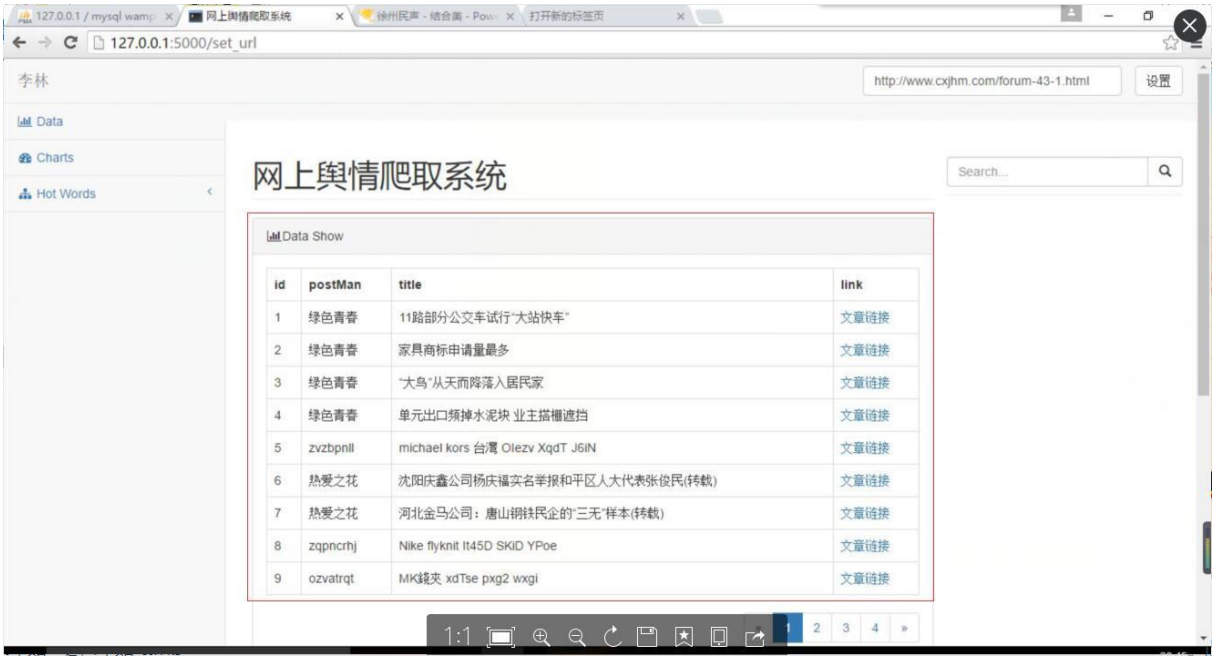

图 4-11 设置 URL 之后刷新的页面

URL 设置的代码如下。

```
@app.route('/set_URL', methods=['POST'])

def set_URL():

    data = request.form

    URL = data.get('URL', None)

    if URL is None:
```

```
        flash('URL 为空', 'danger')


    # 写入文件
    with open('E:\Python Code\Code\Crawler\URLs', 'w') as f:
        f.write(URL)


    items = db.getItem()
    words = db.getWord()
    return render_template(
        'index.html',
        items=items,
        words=words,
        URL = URL,
        data=str(db.getCount_byMonth())
    )
```

4.6 系统后台子模块

系统后台子模块主要是帖子爬取和帖子存储。

4.6.1 帖子爬取模块

首先是爬取的代码：（jihemei_spider.py）主要分三个小模块：URL 设置、爬取帖子相关信息、爬取下一页。

1)URL 设置的功能函数，代码如下所示：

```
 def __init__(self):                    #设置 URL 函数
        # self.start_URLs.append(URL)
        with open("E:\Python Code\Code\Crawler\URLs", "r") as f:
            self.start_URLs.append(f.readline())
```

2)爬取下一页功能函数，代码如下所示：

```
for href in response.css("#fd_page_bottom a::attr('href')"):  # 抓取 button 所有的链接
        URL = response.URLjoin(href.extract())  # 加入队列
        m = re.search(r'www\.cxjhm\.com/forum\-\d+\-(\d+)\.html', URL)
        page = m.group(1)
        if int(page) <= 10:  # page 是字符串，装换为 int
            yield Scrapy.Request(URL, callback=self.parse)  # 再次调用该函数
```

3)爬取帖子的功能函数（核心函数），代码如下所示：

```
def parse(self, response):
    # 首先选择大范围
    sel = Selector(response)
    sites = sel.css('.bm_c tr')
  # 循环逐个获取每个标签下的数据
    for site in sites:
        #新建 PostItem 类
        item = PostItem()
        item['title'] = site.css('.s.xst::text').extract_first()  # 取出其中的文本,取出第一个
        item['postMan'] = site.css('cite a::text').extract_first()
        # 先选择第一个 by，防止第二个 by 干扰。
        cols = site.css('.by')
        col = cols[0]
        # 专门处理时间
        print '----------------------',col is None
        if col != None:
            time = col.css('em span::text').extract_first()
            if time != None:
                time_utf8 = time.encode("utf-8")
                if time_utf8.find("天") != -1:
                    item['firstTime'] = col.css('em span::attr(title)').extract_first()
                elif time_utf8.find("小时") != -1:
                    item['firstTime'] = col.css('em span::attr(title)').extract_first()
```

```
                elif time_utf8.find("分钟") != -1:
                    item['firstTime'] = col.css('em span::attr(title)').extract_first()
                else:
                    item['firstTime'] = col.css('em span::text').extract_first()
            item['replyCount'] = site.css('.xi2::text').extract_first()
            item['readCount'] = site.css('.num em::text').extract_first()
            item['link'] = site.css('.s.xst::attr(href)').extract_first()  # 选择 href
            # 针对该链接爬取 content
            if item['link']:
                link = "http://www.cxjhm.com/" + item['link']
                html = get_content(link, my_headers)
                soup = BeautifulSoup(html)
                content = soup.find(attrs={'class': 't_f'}).get_text()  # 仅需要文本
                # content.replace(" "," ")
                # content = content.replace("\n","<br/>")
                # content = '<pre>'+content+'</pre>'
                item['content'] = content
            #返回
            yield item
```

4.6.2 帖子存储模块

帖子存储模块的代码如下所示：

```
Base = declarative_base()
#初始化数据库连接          防止中文乱码
engine = create_engine('mysql://root:1234@localhost/spider?charset=utf8', echo=True)
#pipeline 对象
class CrawlerPipeline(object):
    def process_item(self, item, spider):
        #创建 Session 类型
        DBSession = sessionmaker(bind=engine)
        #创建 session 对象
        session = DBSession()
```

```
#从 item 中获取帖子的属性，创建 Post 对象。
admin = Post(postMan = item['postMan'],
                    firstTime = item['firstTime'],
                    title = item['title'],
                    content = item['content'],
                    readCount = item['readCount'],
                    replyCount = item['replyCount'],
                    link = item['link'])
#去掉重复
flag = 0                                    #标记 flag，初始值为 0，找到相同的帖子置 1。
items = session.query(Post).all()           #query 方法需要加 Post
for item in items:                          #在已经存在的数据库里面查找当前帖子。
    if item.link == admin.link:
        flag = 1
#判断重复
if flag==0:
    session.add(admin)                      #加入数据库
    session.commit()
    session.close()                         #提交并关闭 session 对象
#返回
return item
```

4.7 本章小结

本章对前台、后台中的模块的功能进行了详细的分析，同时使用代码和截图来让用户更加清晰的了解整个系统的实现过程。

5 关键技术

5.1 系统开发模式

网络程序的开发模式有两种：B/S 模式[6]和 C/S 模式[7]，本系统采用了 B/S 模式。

B/S 模式是一种具有三层结构的技术系统：第一层，客户端发送请求，这些请求会被封装起来，然后通过网络发送出去。第二层，Web 服务器接收到客户机发送的信息，将其暂存在服务器上，服务器通过一系列程序将这些请求分发出去。第三层，数据库服务器负责存储 Web 处理过的数据，主要进行的是增删改查操作。B/S 模式的结构图如图 5-1 所示。

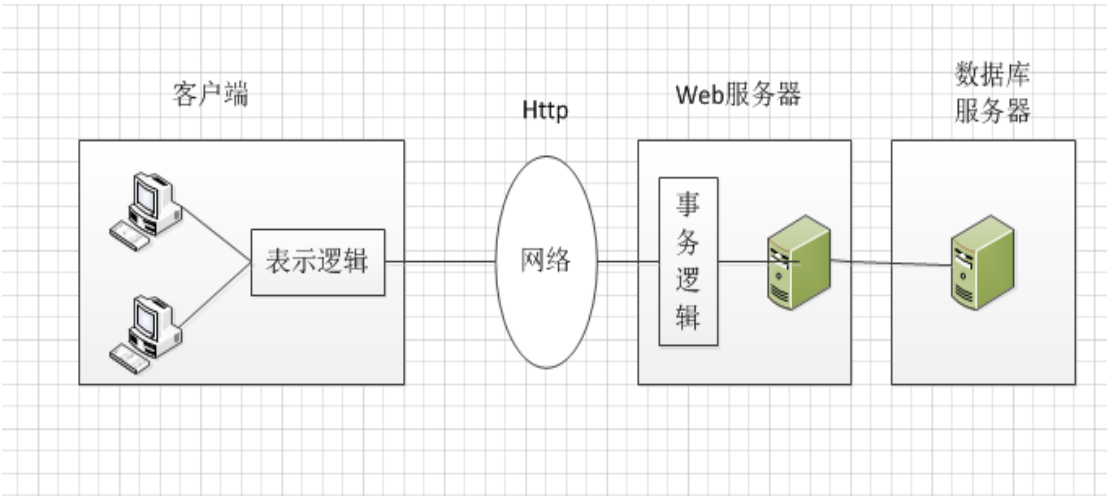

图 5-1 B/S 架构结构图

B/S 模式的优势主要有：

1）开发简单轻便。

2）便捷性强，能够随时浏览。

3）服务器提供了安全的存储。

4）维护简单，维护成本低。

5）网络通信量低。

6）和 C/S 相对比， B/S 模式速度更快。

5.2 DIV+CSS

本系统采用了 DIV+CSS[9]进行页面布局。

DIV+CSS 是“Web 标准”较为常用的专业术语之一。DIV 主要负责页面的布局，DIV 使得整个页面框架结构清晰，CSS[10]（层叠样式表）主要负责页面的美

化，使得页面更加具有亲和力，更加人性化。

使用 DIV+CSS 布局的优势主要体现在内容和形式相分离，也就是 HTML 代码和 CSS 相分离，这样控制更加灵活，使得代码看上去清晰，易于代码的移植和维护。大大降低了网站的成本。

DIV+CSS 的优点（优势）主要体现在以下几个方面：

1）代码简洁易懂，容易上手。

2）形式和内容相分离。

3）提升了 Web 的浏览速度，提升了用户体验。

4）易于维护

5.3 jQuery 和 Ajax 技术

5.3.1 jQuery 技术

本系统采用 jQuery[11]技术，jQuery 的文档通俗易懂，提供了许多优美的插件，jQuery 和 CSS 一样采用代码和内容相分离的技术来设计网页。它的出现一定程度上解放了系统的开发者，提供了极佳的用户体验。

jQuery 具有的重要特性如下：

1）改进了 Ajax 技术，同时引入很多 JSON[12]和 Ajax[13]处理方面的更新。

2）设置函数操作方便。

3）重写了大部分函数，使得这些函数的性能有了较大幅度的提升。

5.3.2 Ajax 技术

本系统在开发过程中采用了 Ajax 技术，Ajax 也就是异步的 Javascript 和 XML[14]，Ajax 技术使得网页只需要局部刷新，很大程度上提高了浏览效率，同时 jQuery 库提供了 Ajax 方法，使得调用 Ajax 十分的方便，本系统正是采用了 jQuery 的 Ajax 方法。Ajax 的特点如下：

1）使用了 web 标准。

2）使用 CSS 的标准和 XHTML。

3）使用 DOM 对象进行交互。

4）调用方便。

5）绑定在 Javascript[15]上。

5.4 Scrapy 框架

Scrapy[16]框架是一个非常成熟的框架，该框架主要用来爬取指定网站的数据，Scrapy 框架应用是十分广泛的，可以爬取数据，信息检索等，Scrapy 框架对于大数据的意义也是不可估量的。

Scrapy 框架的爬取的步骤大致如下：首先定义一个爬取的起始 URL，也就是 start_urls（元组），那么一般这个其实网页内部会有很多 URL，通过这些 URL 会连接到很多其他页面，所以他从当前页面的 URL 开始爬取，然后将这个网页内的其他 URL 存放到一个队列中去，然后进入一个新的页面爬取，然后递归执行上面的操作就可以完成爬取工作。

Scrapy 框架结构图如图 5-2 所示。

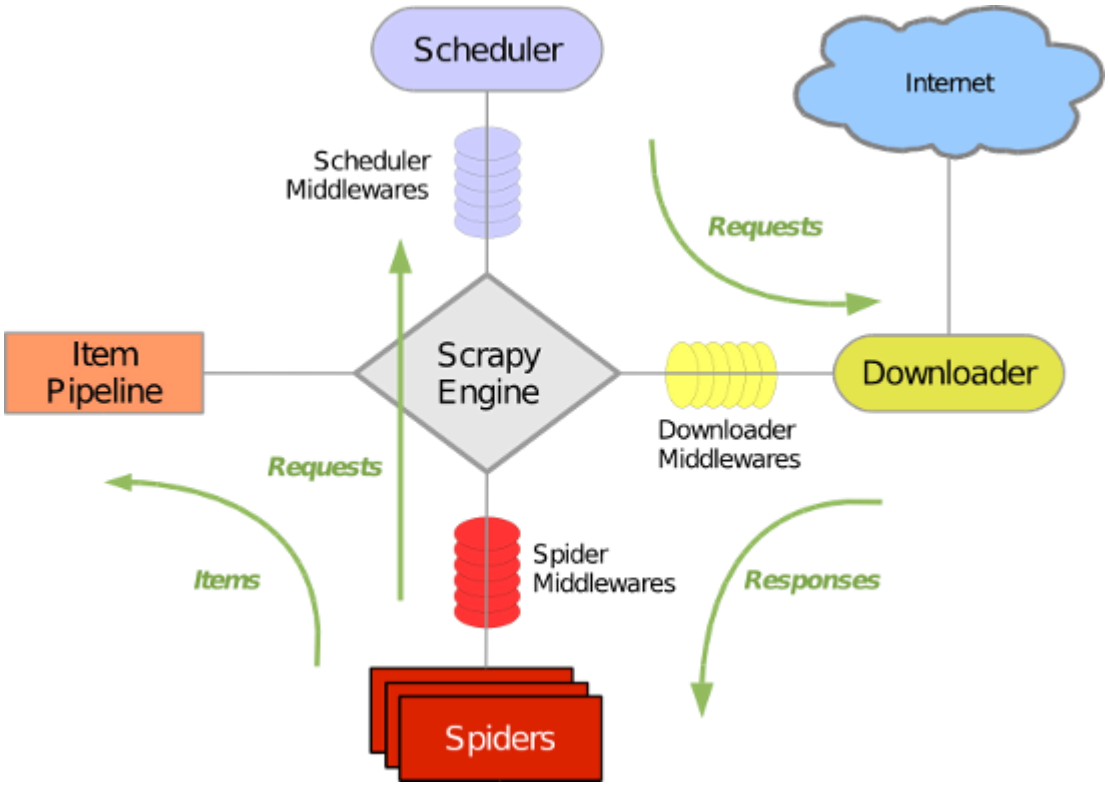

图 5-2 Scrapy 框架结构图

5.5 Flask 框架

Flask[17]是一个微型的、用 Python 编写的 Web 开发框架。尽管 Flask 框架是一个很小的框架，但是其功能不容小觑。Flask 框架真正诠释了短小而精悍。Flask 没有默认使用的数据，因此为了扩展其功能，Flask 加入了 Flask-extension，包括下面介绍的 SQLAlchemy（ORM 工具），WERKZEUG WSGI[19]工具箱和 Jinja2[18]模板引擎的出现使得 Flask 如虎添翼，大大提高了 Flask 的

易用性。Flask 使用 BSD 授权。

5.6 SQLAlchemy

SQLAlchemy [20]是一个开放源代码的软件，SQLAlchemy 的开发使用了当今较为流行的 Python 语言。在 Python 中有很多 ORM 工具，包括 peewee，pyorm，strom，SQLObject 等等，但是 SQLAlchemy 仍可以称得上所有框架中最为优异的框架。SQLAlchemy 提供了必要的对象关系映射（ORM）工具和 SQL expression，SQLAlchemy 的发行使用了 MIT(The MIT License)的许可证。

5.7 本章小结

本章重点介绍了本系统使用的核心技术，其中核心技术包括：B/S 模式，DIV+CSS 技术，jQuery 技术，Ajax 技术，Scrapy 框架，Flask 框架和 SQLAlchemy 技术。

6 总结与展望

6.1 总结

网上舆情爬取系统网站由前台和后台两个部分组成。前台主要提供了帖子查询、图表展示、详细内容展示、敏感词管理和搜索功能。系统的前台部分采用了DIV+CSS 进行页面设计，使得代码清晰，同时又使得页面更加的人性化。本系统还采用 jQuery 技术和 Ajax 技术进行设计，减少代码的冗余，提高了代码的运行效率。系统的后台部分，主要实现的功能是帖子爬取和帖子存储，其中帖子爬取模块使用 Scrapy 框架和 Beautiful Soup 技术，帖子存储模块使用 SQLAlchemy 工具，使得存储数据库的过程很方便。

通过本次的毕业设计，本人感受到了系统开发是个较为复杂的过程。本次毕业设计极大的提高了本人的动手能力和逻辑思维能力，这次开发经历让本人不仅学会使用 Python 语言，还领略到数据库设计对于系统开发的重大意义。

本系统的特色有：

1）系统操作简单，功能模块清晰。

2）前台展示多角度，不仅有数据展示，同时也有图表展示，还有敏感词管理和搜索功能，从各个角度展示爬取的帖子。

3）技术上借助 Scrapy 和 Flask 开发框架，便于日后系统的维护、更新以及功能上的扩展。

4）系统前台界面风格统一、清晰、美观、易用。

6.2 展望

随着网上舆情越来越复杂，本系统在某些方面仍有待改善，主要有以下两个个问题：

1）定制爬取功能有待进一步改善。本系统将爬取的 URL 固定化，虽然可以设置 URL，但是必须修改后台代码，不具有灵活性。

2）舆情推送机制有待进一步提高。本系统并没有实现舆情推送，后续将借鉴推荐技术，将敏感舆情的推送做进一步研发。

参考文献

[1]Magnus Lie Hetlang（挪）. Python 基础教程[M]. 北京：人民邮电出版社，2010，9-19.
[2]贝尔（美）.深入理解 MySQL[M]. 北京：人民邮电出版社，2010，50-85.
[3]Zheng Cirino(美). Pycharm. 中国国际图书贸易集团公司. 2005,66-68
[4]何富贵 JSP 开发案例教程[M]. 北京，机械工业出版社，2013.
[5]亓晓静 计算机应用与软件技术专业：基于 C/S 架构的软件项目实训[M] 北京，电子工业出版社, 2010,13-17
[6]白勇.用 B/S 模式构建学校管理信息系统[J].重庆电力高等专科学校学报，1999（03）:66-69..
[7]Tsui，Frank F. Python P em dash a research processor in josephson technology[C]. IBM Journal of Research and Development，Vol34， No2,1140：222-235.
[8]Tsui，Frank F. Python em dash a research processor in josephson technology[C]. IBM Journal of Research and Development，Vol44， No2,1980：243-252.
[9]Bear(美), Bibeault,Yehuda Katz.jQuery 实战[M].人民邮电出版社，2010,24-46.
[10]帕里(美) Ajax Hacks[M]. 电子工业出版社，2014,5-8.
[11]廖雪峰. JSON 入门指南[M]. 电子工业出版社. 2008，60-66
[12]亨特(美) XML 入门经典(第 4 版)[M]. 清华大学出版社. 2009, 10-15
[13]弗拉纳根(美) javascript 权威指南[M]. 机械工业出版社. 2007，5-10
[14]Romanoff(美) Scrapy 入门教程[J]. 人民邮电出版社. 2009,20-32
[15]Miguel Grinberg. flask web development. O'Reilly Media.2014,20-25
[16]Miguel Grinberg. flask web development. O'Reilly Media.2014,67-75
[17]Miguel Grinberg. flask web development. O'Reilly Media.2014,144-147
[18] Kong Michael. An environment for secure SQLAlchemy [M]. Oxford University Press Inc, 1993: 149.
[19] Zhang, L. and W. Zhang. Implement of e-government system with data persistence of beautiful soup[M].. Hong Kong,2010:66-76.
[20]Mark Ramm(美). SQLAlchemy，Addison-Wesley Professional. 2010，55-66

致谢

本系统的开发和实现均在董永权老师的悉心指导下完成，特别是在本系统实现的过程中，董老师对我系统的整体框架和功能提出了许多宝贵的意见，并且指出了本次实现过程的重点和难点，让我不再惧怕困难，努力完成。他严谨治学的态度和对于学生的悉心指导，都让我对完成本系统信心十足。在系统即将结束时，董老师仍不忘悉心指导我，对整个系统的完善提出了很多建设性的意见，并且对于以后系统开发提出了很多宝贵的意见。在此，我要向董老师致以最诚挚的谢意。董老师对于计算机事业的追求和热爱，使我引发了真挚的思考和无穷的启发。在此，我要向董老师真诚地感谢。

感谢江苏师范大学智慧教育学院（计算机科学与技术学院），给了我本科四年的成长和学习专业知识平台，为我提供了良好的学习环境和学习氛围。

感谢我同窗思念的同学，在我遇到困难和遭到挫折的时候，给予了我莫大的鼓励和关怀。

本系统定有很多的不足，恳请各位老师指正。