PaperFree检测报告简明打印版

相似度：13.34%
编　号：R8BDUIWSH4FGQOHJ
标　题：网上舆情爬取系统的设计与实现
作　者：李林
长　度：17680字符
时　间：2016-05-21 18:09:05
比对库：中国学位论文全文数据库；中国学术期刊数据库；中国重要会议论文全文数据库；英文论文全文数据库；互联网资源；自建比对库

相似资源列表(学术期刊，学位论文，会议论文，英文论文等本地数据库资源)
1. 相似度：1.11%　篇名：《进销存管理系统》
来源：《南昌大学硕士论文》　年份：2010　作者：蔡雯
2. 相似度：0.58%　篇名：《基于.Net三层架构高校户籍管理系统设计与实现》
来源：《软件导刊》　年份：2011　作者：纪洲鹏
3. 相似度：0.47%　篇名：《提高数字图书馆性能的问题探讨》
来源：《天津职业院校联合学报》　年份：2013　作者：郝素敏
4. 相似度：0.37%　篇名：《Ajax技术在高校学生管理系统的应用》
来源：《科技创新导报》　年份：2014　作者：李佳凝
5. 相似度：0.25%　篇名：《基于PHP的包装企业门户网站设计与实现》
来源：《包装工程》　年份：2013　作者：杨凌云
6. 相似度：0.25%　篇名：《医院医疗设备管理数字化设计》
来源：《中国管理信息化》　年份：2013　作者：杜昱铿
7. 相似度：0.23%　篇名：《建立规范高效的税收数据质量管理体系--以某省地税部门为例》
来源：《网友世界》　年份：2014　作者：于众
8. 相似度：0.23%　篇名：《浅谈通过多媒体展示化学实验》
来源：《中学课程辅导：教学研究》　年份：2014　作者：吴怡生
9. 相似度：0.23%　篇名：《面向对象的编程思路》
来源：《福建电脑》　年份：2004　作者：王文陵
10. 相似度：0.23%　篇名：《基于SE的高职学院设备管理系统的分析与设计》
来源：《装备制造与教育》　年份：2014　作者：张衍志
11. 相似度：0.23%　篇名：《《数据库原理及应用》的多层次系统化实验教学研究》
来源：《实验科学与技术》　年份：2013　作者：牛新征
12. 相似度：0.22%　篇名：《领导干部如何提升网络舆情的应对能力》
来源：《山西青年：下半月》　年份：2013　作者：饶晓娜
13. 相似度：0.22%　篇名：《供应链库存管理系统设计——基于中小型制造企业ERP系统的库存管理研究之二》
来源：《轻工科技》　年份：2013　作者：陈璐
14. 相似度：0.22%　篇名：《校内毕业论文检查系统的设计与实现》
来源：《科技风》　年份：2011　作者：程杰
15. 相似度：0.22%　篇名：《动态语言Python探讨与比较》
来源：《企业科技与发展：上半月》　年份：2012　作者：张茗芳
16. 相似度：0.22%　篇名：《新媒体环境下创新高校宣传思想工作路径探究》
来源：《学校党建与思想教育：下》　年份：2013　作者：张岳君
17. 相似度：0.22%　篇名：《基于web的高校毕业论文档案管理信息系统的设计》
来源：《科技资讯》　年份：2013　作者：江小华
18. 相似度：0.22%　篇名：《网络舆情文化治理研究》
来源：《湖北社会科学》　年份：2013　作者：李鸣
19. 相似度：0.21%　篇名：《企业舆情分析系统的设计与实现》
来源：《西安电子科技大学硕士论文》　年份：2013　作者：贾利娟
20. 相似度：0.2%　篇名：《论计算机基础教育中的网络文化》

来源：《吉林省教育学院学报：中旬》 年份：2013 作者：韩雪松
21. 相似度：0.18% 篇名：《基于Struts技术的网上购物系统的设计与实现》
来源：《商场现代化》 年份：2013 作者：赵忠华
22. 相似度：0.17% 篇名：《高中学生信息管理系统的设计开发与应用》
来源：《祖国：教育版》 年份：2013 作者：杜建峰
23. 相似度：0.16% 篇名：《物资仓库管理系统信息化设计措施》
来源：《科学与财富》 年份：2013 作者：张新梅
24. 相似度：0.16% 篇名：《培训档案管理系统设计与开发》
来源：《软件导刊.教育技术》 年份：2013 作者：陆峰
25. 相似度：0.14% 篇名：《家具销售管理系统的设计与探索》
来源：《黑龙江科技信息》 年份：2014 作者：罗俭
26. 相似度：0.14% 篇名：《基于ASP．NET高校科研管理信息系统的开发及应用》
来源：《高教论坛》 年份：2013 作者：万荣泽
27. 相似度：0.14% 篇名：《基于Python语言的面向对象程序设计课程教学》
来源：《计算机工程与科学》 年份：2014 作者：狄博
28. 相似度：0.14% 篇名：《进销存管理系统的设计与实现》
来源：《海峡科学》 年份：2012 作者：吴章贵
29. 相似度：0.12% 篇名：《议在线考试系统的研发与应用》
来源：《神州》 年份：2013 作者：宋洁心
30. 相似度：0.12% 篇名：《软件工程专业静态网页制作课程教学内容改革》
来源：《内蒙古财经大学学报》 年份：2013 作者：王鑫
31. 相似度：0.11% 篇名：《基于ASP．NET技术的项目任务管理系统》
来源：《中国科技博览》 年份：2014 作者：陈应彬
32. 相似度：0.11% 篇名：《“网页设计”课程中DIV＋CSS布局技术的教学》
来源：《计算机时代》 年份：2013 作者：孟庆轩
33. 相似度：0.11% 篇名：《图书馆微信公众号建设》
来源：《图书馆杂志》 年份：2014 作者：黎邦群
34. 相似度：0.1% 篇名：《本科毕业论文（设计）管理系统的设计研究》
来源：《中国科技纵横》 年份：2015 作者：张亦秋
35. 相似度：0.1% 篇名：《基于ASP．NET的高校学生成绩管理系统》
来源：《商情》 年份：2013 作者：伦冠民
36. 相似度：0.09% 篇名：《云计算技术在企业电子商务中的应用探讨》
来源：《商场现代化》 年份：2013 作者：龚谨
37. 相似度：0.09% 篇名：《互联网时代下商业银行经营的新变化》
来源：《时代经贸》 年份：2013 作者：周珍

相似资源列表(百度文库，豆丁文库，博客，新闻网站等互联网资源)

1. 相似度：0.55% 标题：《biyexinxiguanlixitong 本系统是采用B/S模式进行开发的, 的用户...》
来源：http://www.pudn.com/downloads487/doc/project/detail2030841.html
2. 相似度：0.47% 标题：《Flask -- 使用Python和OpenShift进行即时Web开发 - lgphp - 推酷》
来源：http://www.tuicool.com/articles/Nr6R3a
3. 相似度：0.46% 标题：《爬虫框架Scrapy实战之批量抓取招聘信息 - Python框架教程 - ...》
来源：http://www.pythontab.com/html/2015/pythonweb_0410/943.html
4. 相似度：0.43% 标题：《写一个标志变量int flag ,int flag =1//代表true ,0//代表false?》
来源：http://zhidao.baidu.com/link?url=zBH92GGfz-pWQpaxUdj1RFn2Nnvp77dX_-n-2GXxGbxLwQQ9NOHC8-wFSc6MclA7KwUVRm3zL9nMp6hfc7uU0q
5. 相似度：0.36% 标题：《Python资源大全- 方倍工作室- 博客园》
来源：http://www.cnblogs.com/txw1958/p/python-tutorial-list.html
6. 相似度：0.33% 标题：《The Guidance of Public Opinion_百度文库》
来源：http://wenku.baidu.com/link?url=9wT9U-p38Akk8raIXIGFyZ8guIbw5grk9Ds4RUbt7QpffKr__bNIg_toAZqZ2w8SOmy_uC3i9pCSmPNDS4QSnB4IU
7. 相似度：0.32% 标题：《【scrapy】学习Scrapy入门 - 简书》
来源：http://www.jianshu.com/p/a8aad3bf4dc4
8. 相似度：0.3% 标题：《SQLAlchemy 简单笔记 - 简书》
来源：http://www.jianshu.com/p/e6bba189fcbd

9. 相似度：0.27% 标题：《开发工具 | CODE开源知识库 | CODE》
来源：http://code.csdn.net/openkb/c-244
10. 相似度：0.25% 标题：《Item Pipeline — Scrapy 1.0.6 documentation》
来源：http://doc.scrapy.org/en/latest/topics/item-pipeline.html
11. 相似度：0.23% 标题：《DIV+CSS是什么意思呢?实质是什么?_百度知道》
来源：
http://zhidao.baidu.com/link?url=B0RzYueb14wo3YD36vzwgTMSkIWCUJovmBrO0Ms2a6fAZ0BOLW72
GJI9WP1Y-5AqX67DWXAGDvY6v13qX90c4G03y5ZG
12. 相似度：0.18% 标题：《python简介_百度文库》
来源：
http://wenku.baidu.com/link?url=e06ccx42SRSEISUdmBgZUQiZHRklgSm-5SxbXPNjeiUsnUfrkNTERv6u
13. 相似度：0.18% 标题：《myEclipse:开发Java, J2EE的 Eclipse 插件集合 for Mac 8.6.0下...》
来源：http://www.macappbox.com/myeclipse/
14. 相似度：0.17% 标题：《scrapy-redis 和 scrapy 有什么区别? - 知乎用户的回答 - 知乎》
来源：https://www.zhihu.com/question/32302268/answer/55724369
15. 相似度：0.11% 标题：《又来求助了,大神求解答 python类继承的问题_百度知道》
来源：http://zhidao.baidu.com/link?url=MrBEcE6f8bAsTY3Y6G3ZoYrKw-
u84OrKlB8NealQg5QcGirXwpkduoyuNUBc5B6cy7PbbrvPTXl3nWN49f5GB0p_CLbsINQ3Uu2Nahiym8C
16. 相似度：0.1% 标题：《Flask框架学习笔记（一）安装篇（windows安装与centos安装）_...》
来源：http://www.169it.com/tech-python/article-539019800.html

全文简明报告

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,{80% : 互联网已成为广大人民群众思想文化信息的集散地, }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开发,使用了B/S架构,选用了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:爬取帖子和存储帖子。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

{ 60% : With the rapid development of computer technology, the network has become the people to express the views of different social issues important place, } the Internet has become a distribution center for the masses, the opinion of the Internet presents the trend of diversification.{ 62% : For the correct guidance of public opinion, } public opinion monitoring network is imperative, and the crawling system is the one important component.{ 63% : The system for the needs of development, the use of B / S architecture, } the choice of the Python language and MySQL database development. Online public opinion crawling system comprises a total of two modules: the foreground and background display module crawling module, which shows the front desk module consists of four parts: the post display, post search for sensitive words and URL management settings. Background crawling module consists of two parts: the storage and crawling Posts Posts.{82% : This system has some value, stable operation, } to help users learn about the latest public opinion, to lay the foundation for the Internet public opinion monitoring.

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

目 录

{ 62% : 随着计算机技术的应用和发展,网络以及普及到千家万户, }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,网上爬取系统可以将

爬取的数据进行整理分析,通过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,该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

{ 66% : 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,{ 68% : 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,{ 59% : 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,{ 61% : 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,{ 78% : 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

{83% : 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 本文的研究内容及贡献

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。

网上舆情爬取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),{ 66% : 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。{ 62% : 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。主要详细描述了网上舆情爬取系统的背景、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。主要介绍了网上舆情爬取系统的需求,包括性能需求和相关功能需求。

{ 74% : 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。{ 63% : 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。主要介绍了网上舆情爬取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

本文主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内容及主要的贡献。

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

{ 64% : 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,{ 62% : 同时也提供了搜索用户的功能。 }

5)URL设置

{ 64% : 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)爬取帖子: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)存储帖子: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。管理员可以爬取帖子,存储帖子。{ 68% : 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

ï•后台服务器。

1)Windows 10

2)python2.7.11

ï•客户端计算机。

1) Windows 10

2) Chrome 49.0

ï•数据库服务器。

MySQL+WampServer

2.2.2 系统硬件环境及要求

本系统对于计算机的容量和CPU有一定的要求。下面从这两个方面阐述。

1)容量要求:对于容量需求主要通过两个方面来讨论:磁盘容量需求和内存容量需求。本系统磁盘容量需求有限,主要是爬取和展示的代码已经数据库存储的帖子,总共占用20M左右。本系统需要一定的内存容量,由于使用了Scrapy框架,系统爬取网页时需要占用一定的内存空间,本系统虽然采用了多线程技术,但是测试计算机的内存容量是6GB,故爬取网站时不会影响计算机的其他操作。

2)CPU要求:Scrapy框架是支持多线程的,同时Scrapy框架也是默认多线程爬取的,当然Python语言也是支持多线程的,故本系统采用了多线程技术,对CPU有一定的要求。

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,{ 59% : 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,{ 62% : 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

在开发本系统钱,笔者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,同样,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。系统业务流程如图2-3、2-4、2-5和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:{ 59% : 本网上舆情爬取系统采用的是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,使用B/S结构方便开发人员日后的维护。此外,本系统采用目前最为流行的Scrapy框架以及beautiful soup[5]技术,维护起来方便简单。

{ 59% : 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。此网上舆情爬取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

{ 66% : 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

网上舆情爬取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。后台爬取模块有爬取帖子和存储帖子。

网上舆情爬取系统前台展示功能模块如图3-1所示。

图3-1系统前台展示模块

网上舆情爬取系统后台的爬取功能模块如图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,{ 61% : 上文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

根据本网站的需求,系统使用的表如表3-1和3-2所示。

表3-1网上舆情爬取系统帖子表(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网上舆情爬取系统敏感词表(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

本章节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,{ 61%：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

浏览帖子子模块主要是将爬取的帖子用表格的形式展示在网页上,主要展示发帖人、发帖时间、帖子标题以及帖子的内容。

帖子展示模块如图4-1和4-2所示。

图4-1帖子展示

图4-2内容展示

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除敏感词子模块。这样可以根据用户的喜好定制自己喜欢的敏感词,这样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

敏感词管理的代码如下所示:

^li^

^a href="#"^^i class="fa fa-sitemap fa-fw"^^/i^ Hot Words ^span class="fa arrow"^^/span^^/a^

^ul class="nav nav-second-level" id="hot_words"^

{% for word in words %}

^li^

^a href=http://127.0.0.1:5000/search/{{ word.word }} target="_self"^{{ word.word }}^/a^

^/li^

{% endfor %}

^li^

^a href="javascript:;" onclick="add_word()" id="add_word"^^strong^添加敏感词^/strong^^/a^

^/li^

^li^

^a href="javascript:;" onclick="delete_word()" id="delete_word" style="font-weight:bold; color:red"^删除敏感词^/a^

^/li^

^/ul^

^!-- /.nav-second-level --^

^/li^

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

删除帖子子模块采用提示框的方式让用户输入

删除帖子如图4-7和4-8所示。

图4-7删除敏感词

图4-8删除成功

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

搜索帖子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。

搜索帖子如图4-9所示。

图4-9搜索帖子

搜索帖子的代码。

```
@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 是让用户输入自己想爬取的URL(在http://www.cxjhm.com/域名之下),后台获取到URL之后,清空一下数据库,重新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

设置URL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爬取帖子和存储帖子。

4.6.1 爬取帖子模块

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:

```
{100% : 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 存储帖子模块

存储帖子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
{ 67% : class CrawlerPipeline(object): }
```

{100% : 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
{ 60% : 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

{ 62% : 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[6] 和C/S的[7],他们分别是也浏览器/服务器模式和客户机/服务器模式,{ 63% : 本系统采用了浏览器/服务器模式。 }

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,这些请求会被按照一定的语法封装起来,准备通过网络发送。第二层,Web服务器接收到客户机发送的信息,通过一定的语法暂存在服务器上,服务器按照一定的程序将这些请求分发出去。第三层时,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架构的结构图如图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。

综上所述,本网上舆情爬取系统采用B/S模式设计实现。

5.2 DIV+CSS

本系统采用了DIV+CSS[9]进行页面布局。

{ 71% : 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。{ 66% : 大大降低了网站的成本。 }

{ 68% : 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

本系统用了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。

5.3.2Ajax技术

本爬取系统网站在开发过程中采用了Ajax技术,{ 64% : 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

{ 62% : 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,然后递归上面的操作就可以完成爬取工作。

Scrapy框架结构图5-2如下。

图5-2 Scrapy框架结构图

5.5 Flask框架

{ 78% : Flask[17]是一个微型的、用Python编写的Web开发框架。 }{ 67% : 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。{ 62% : Flask使用BSD授权。 }

5.6 SQLAlchemy

SQLAlchemy [20]是一个开放源代码的软件,{ 69% : 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,{ 73% : 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。{ 63% : 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

本章节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,{ 64% : jQuery技术,Ajax技术, }Scrapy框架,Flask框架和SQLAlchemy技术,通过介绍这些技术,可以是用户更为方便的了解系统的开发。

6 总结与展望

6.1总结

网上舆情爬取系统已全部完成。该系统后台采用了Scrapy框架和beautiful soup技术,前端采用了html、css、javascript技术和Flask框架。开发工具选用了PyCharm。

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。在前台方面,笔者自学了一些前端技术,包括HTML,{100% : CSS,Javascript, }jQuery。笔者采用了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。同时,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,主要实现的功能是爬取帖子,存储帖子,其中爬取帖子笔者是用了Scrapy框架和beautiful soup技术,存储帖子笔者使用了sqlalchemy工具,使得存储数据库的过程很方便。

通过本次的课题设计,笔者感受到了系统开发是个较为复杂的过程。本次设计极大的提高了笔者的动手能力和逻辑思维能力,这次开发经历让笔者不仅学会使用Python语言,还让笔者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,前台展示不仅有数据展示,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

{ 63% : 本课题虽然达到了预期的效果, }但是随着网上舆情越来越复杂,本系统在某些些方面仍有待改善。主要有几下三个问题:

1)个性化:该程序将URL固定,如果用户希望爬取自己想爬取的模块,需要修改后台代码,需要添加一个功能让用户输入自己想爬取的网址,启动程序爬取网站即可。

2)网上舆情爬取系统网站的安全性:本系统的数据库安全性需要进一步加强,因此,可以采取一些必要的加密手段,比如通过MD5算法或者sha1算法进行加密。

3)网站的交互性:该网站未提供登陆和注册功能,缺少一些人性化的推送。

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,董老师对我的系统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难点,让我不再惧怕困难,努力完成。他的严谨的治学态度和对于学生的悉心指导,都让我对完成本系统更为有信心。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。

本系统一定有很多的不足,恳请各位老师指正。