PaperFree检测报告简明打印版

相似度：18.53%
编　号：KR4M6OSF793TWBDQ
标　题：网上舆情爬取系统的设计与实现
作　者：李林
长　度：19891字符
时　间：2016-05-15 16:52:52
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：1.41%　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
2. 相似度：0.55%　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
3. 相似度：0.52%　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
4. 相似度：0.36%　篇名：《以全媒体思维建设中国科技媒体集团的大数据技术平台》
来源：《中国传媒科技》　年份：2014　作者：史晓波
5. 相似度：0.32%　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
6. 相似度：0.32%　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
7. 相似度：0.26%　篇名：《物流管理系统的设计和开发》
来源：《西部教育研究》　年份：2013　作者：刘镁煜
8. 相似度：0.24%　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
9. 相似度：0.23%　篇名：《基于AJAX和CSS技术的教师在线评价系统设计》
来源：《昆明学院学报》　年份：2013　作者：何国英
10. 相似度：0.22%　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
11. 相似度：0.2%　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
12. 相似度：0.2%　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
13. 相似度：0.2%　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
14. 相似度：0.18%　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
15. 相似度：0.17%　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
16. 相似度：0.17%　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
17. 相似度：0.17%　篇名：《基于SSH整合技术的土壤-茶系统B / S研究》
来源：《山东农业科学》　年份：2014　作者：杨玉建
18. 相似度：0.15%　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
19. 相似度：0.15%　篇名：《论高校微博舆情传播路径及引导控制》
来源：《西南农业大学学报：社会科学版》　年份：2013　作者：郭峰
20. 相似度：0.13%　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
21. 相似度：0.13% 篇名：《资产评估教学实验系统分析与设计》
来源：《内蒙古财经大学学报》 年份：2014 作者：乔永峰
22. 相似度：0.13% 篇名：《基于ASP.NET的电子商务系统》
来源：《中国科技信息》 年份：2011 作者：李文坚
23. 相似度：0.12% 篇名：《RIA技术综述》
来源：《都市家教：上半月》 年份：2013 作者：何建
24. 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
25. 相似度：0.1% 篇名：《网上毕业论文管理系统的设计与实现》
来源：《微型电脑应用》 年份：2013 作者：徐远棋
26. 相似度：0.1% 篇名：《基于ArcSDE和SQLServer的新农村建设数据库设计与实现》
来源：《安徽农业科学》 年份：2013 作者：张洪吉
27. 相似度：0.09% 篇名：《竞赛管理平台的设计与实现》
来源：《产业与科技论坛》 年份：2014 作者：姜秀辉
28. 相似度：0.09% 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
29. 相似度：0.09% 篇名：《利用HTML5的本地存储实现图书馆网站的个性化》
来源：《科技资讯》 年份：2013 作者：许中博
30. 相似度：0.09% 篇名：《校园网建设的新思路》
来源：《新课程：教育学术》 年份：2013 作者：朱琪
31. 相似度：0.08% 篇名：《Ajax技术实现在线智能化考试系统》
来源：《管理观察》 年份：2013 作者：谢会娟

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：2.78% 标题：《开发工具 | CODE开源知识库 | CODE》
来源：http://code.csdn.net/openkb/c-244
2. 相似度：1.52% 标题：《【scrapy】学习Scrapy入门 - 简书》
来源：http://www.jianshu.com/p/a8aad3bf4dc4
3. 相似度：1.23% 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
4. 相似度：1.15% 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
5. 相似度：0.82% 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
6. 相似度：0.63% 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
7. 相似度：0.56% 标题：《python简介_百度文库》
来源：
http://wenku.baidu.com/link?url=e06ccx42SRSEISUdmBgZUQiZHRklgSm-5SxbXPNjeiUsnUfrkNTERv6u
8. 相似度：0.46% 标题：《JavaScript和jQuery实战手册(原书第2版)(china-pub首发) - china-...》
来源：http://product.china-pub.com/3022607
9. 相似度：0.39% 标题：《自学Python十二 战斗吧Scrapy! - 我的代码会飞 - 博客园》
来源：http://www.cnblogs.com/jixin/p/5158177.html
10. 相似度：0.29% 标题：《[ASP.NET MVC 小牛之路]14 - Unobtrusive Ajax - Liam Wang - 博...》
来源：http://www.cnblogs.com/willick/p/3418517.html
11. 相似度：0.26% 标题：《MySQL高级特性----对比与其他数据库-Mysql-华夏名网资讯中心 虚...》
来源：http://www.sudu.cn/info/index.php?id=321521&op=article
12. 相似度：0.24% 标题：《AJAX简介 (缩写:Asynchronous JavaScript and XML) - min..._博客园》
来源：http://www.cnblogs.com/min10/archive/2009/03/18/1415492.html
13. 相似度：0.2% 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=B0RzYueb14wo3YD36vzwgTMSkIWCUJovmBrO0Ms2a6fAZ0BOLW72GJI9WP1Y-5AqX67DWXAGDvY6v13qX90c4G03y5ZG

14. 相似度：0.14% 标题：《触碰jQuery:AJAX异步详解 - 滴答的雨 - 博客园》
来源：http://www.cnblogs.com/heyuquan/archive/2013/05/13/js-jquery-ajax.html
15. 相似度：0.14% 标题：《python是个什么东西_百度知道》
来源：http://zhidao.baidu.com/link?url=5NU_OzMHJyHu3FvneXEQKc85gS-yVwHEPe0EUI2igDGIUzQwS9KFuHQfjU0NVfi-stxMJh8WWW1OW0qufWZ5RL4CF2WFYwPd9bnYyW3A6rm
16. 相似度：0.09% 标题：《AJAX技术中Session服务的改进--《计算机技术与发展》2006年12期》
来源：http://www.cnki.com.cn/Article/CJFDTotal-WJFZ200612024.htm

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,网络已成为们对不同社会问题发表看法的重要场所,{89% : 互联网也成为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的重要的组成部分。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

{100% : With the rapid development of computer technology, } the network has become to express their views on various social issues important place, the Internet has become the ideological and cultural hub of information, the network of public opinion presents a trend of diversification. For the correct guidance of public opinion, public opinion monitoring network is imperative, and the crawling system is the important part of it.{ 65% : The system for the needs of development, the use of B / S structure, } using 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

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

{ 69% : 随着计算机技术的应用和发展,网络作为信息的载体,已经越来越普及,它已经成为公众参与社会生活的重要信息通道。 }因为互联网是开放的,每个人都有权利在互联网上面发表自己想要发表的信息,信息的内容涵盖了很多方面,{ 61% : 小到日志心情,大到全国性活动。 }{ 63% : 互联网已拥有传统媒体无法比拟的优势:方便,虚拟,互动性,多样性。 }

{ 67% : 网络舆情热点通常是迅速形成,更多的人发表对于日常生活问题的各种看法,意见,态度,情绪等,伴随着事件和变化的发展,网络已然成为社会热点的重要载体。 }

随着网上舆情的深入发展,网上舆情呈现了多样化、复杂化的趋势,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指日可待。

1.1.2研究意义

随着网络技术的不断发展,微博、贴吧和论坛等社交工具发展迅猛,用户量越来越大,网络舆情时常出现一些错误的导向,为了监控舆情,使得网络舆情走上一个正确的轨道,爬取系统势在必行。同时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

{ 79% : Python[1]是一种解释性的程序设计语言, }{ 68% : 同时Python还具有面向对象的特征。 }它由吉多·范罗苏姆在1989年发明,Python的第一个版本在1991年发行。

{ 70% : Python语法十分简洁清晰,强制用空白符(White Space)作为语句缩进是Python最为显著的特点。 }

{ 79% : Python有一个丰富而强大的库。 }人们经常称之为胶水语言,是因为Python能够很容易地与其它的语言(尤其是C/ C ++)的各种模块连接在一起。{ 64% : 常见的应用场景是Python快速生成原型程序(有时甚至是最后的程序界面),其中有一些特殊的需求,可以使用其他合适的语言进行改写,比如3D游戏已经3D电影渲染的模块,渲染模块对于性能要求是很高的,必须使用C/C ++重写,然后打包成Python的扩展库, }然后就可以被调用了。{ 68% : 需要注意的是该平台可能需要当您使用扩展库来考虑的一个问题,有些实现可能无法提供跨平台的功能。 }

1.2.2 MySQL数据库简介

{92% : MySQL[2]是一个关系型数据库管理系统, }并且在今天的数据库管理系统中,MySQL的影响力是最大的,MySQL的速度和健壮性是其优势的集中体现。{ 64% : 它被称为关系型数据库管理系统, }是因为它将数据存放在分立的表格中,而不会将所有的数据不加分类叠加在一起。因此,{ 59% : 这样很大程度上增加了灵活性, }同时也提高了存储速度。

与其他数据库相比,MySQL有很多更为先进的功能。其特点如下:

1)性能

因为MySQL是没有线程创建开销,所以MySQL将在一些有以下几个方面的速度更快:

{98% : ①CREATE TABLE和DROP TABLE。 }

②在不属于同一个索引的东西上进行SELECT操作。

③几个键插入列到一个简单的表格中去。

2)磁盘空间效率

MySQL的表在内存中占用很小。例如,MEDIUMINT只需要占用3个字节的长度。在大数据量的情况下,节省一个字节是很重要的。

3)稳定性

MySQL有很多优点:高便携性、较为快速的安装过程、紧凑等一系列优点,这使得中小网站通常将MySQL作为数据库。

MySQL的可以支持多种操作系统,如:{ 62% : FreeBSD,MacOS, }Linux,AIX和Windows,并为许多编程语言提供了应用开发程序接口。为了保证代码的可移植性,MySQL采用较为底层的语言(C/C++)进行编写。 MySQL不仅能够支持多线程,提升SQL的查询速度,同时还优化搜索算法。

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

本文主要介绍了网上舆情爬取系统的背景、意义、整体设计思路以及关键技术等。

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(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readCount和回复replyCount的数量),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。主要详细描述了网上舆情爬取系统的背景、意义、开发语言、开发工具的选用和介绍,同时介绍了本文的主要贡献和研究内容。

第2章:需求分析。主要介绍了系统的性能需求和相关的功能需求。

{ 74% : 第3章:系统功能模块设计。 }阐述了系统功能的模块的划分和数据库的设计与实现。

第4章:系统实现流程。介绍了系统各个功能模块,并且对模块的核心代码进行展示。

第5章:关键技术。介绍了网上舆情爬取系统的核心技术以及配置。

第6章:总结和展望。

1.4 本章小结

本系统主要介绍了系统的研究背景及意义、开发工具及语言和研究内容及主要贡献。

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

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

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)

用户用例图表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)

管理员用例图表示了一个管理员的操作权限。管理员可以爬取帖子,存储帖子。{ 65% : 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

计算机

2.2.3 系统的性能要求

1)磁盘容量要求:由于本爬取系统是基于B/S的架构,因此在存储容量方面,本网站所占用的内存容量很小,系统文件大概需要9M,数据库系统系统文件占用空间很有限。但是,本网上舆情爬取系统中的后台爬取模块需要一定内存空间,需要将一个论坛上某个板块所有的帖子全部爬取下来,会占用一定的内存空间。

2)适应性要求:本系统要求功能模块清晰,功能明确。模块与模块之间耦合性弱,内聚性强,能够使用户在极短的时间内熟悉网上舆情爬取系统的整个操作流程,用户体验良好。

2.3 可行性分析

可行性分析指的是在现有的组织环境下,分析一下该爬取系统的开发工作是否已经具备了必要的工作条件及资源。本爬取系统不仅是对于网上舆情的爬取,对于其他爬取系统类似于电影或者课后习题等网站的爬取具有很强的参加价值。

2.3.1概述

网上舆情爬取系统的可行性研究是系统的设计与开发人员进行开发的前提条件和基础,可行性研究能够使开发人员在系统研发的初期敏感的发现系统在需求方面的不足,这样可以避免耗费大量财力,物力和人力等诸多困难。综上所述,可行性研究及分析在软件开发的整个软件开发过程的中有着举足轻重的地位。

2.3.2 系统业务流程调查

本系统的业务流程大致可以分为两部分。一个部分是搜索帖子。用户根据需求进行搜索。另一部分是添加敏感词,同样,用户可以根据需求添加敏感词。还有一部分是设置URL。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:{ 73%：本系统主要采用B/S架构进行开发。 }这种架构具有较为良好的开放性、可扩展性和共享性,便与系统的开发和维护。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。

2)经济的可行性:成本效益分析是经济可行性分析中最为重要的内容。如果开发一个系统在经济方面就不适用,则完全没有必要开发这个系统。此网上舆情爬取系统开发成本低廉,因此在经济方面完全可行。

2.4本章小结

本章主要介绍了该网站的需求,包括软硬件环境,系统的性能需求、功能需求以及可行性分析和调查。

3 系统总体功能模块设计

3.1 系统功能模块的划分

网上舆情爬取系统有四个模块:帖子展示、图表展示、搜索帖子、敏感词管理和URL设置。

3.1.1系统功能模块结构图

网上舆情爬取系统前台展示功能模块和模块之间的关系如图3-1所示。

图3-1系统前台展示功能模块

网上舆情爬取系统后台的爬取功能模块和模块之间的关系如图3-2所示。

图3-2系统后台爬取功能模块

3.1.2系统的功能模块描述

1)爬取帖子

爬取帖子主要分为爬取帖子的发帖人、发帖时间、帖子标题、阅读和回复数量、帖子链接以及该帖子内容,同时该系统添加了循环抓取下一页的功能。

爬取帖子模块图如图3-3所示。

图3-3爬取帖子模块

2)存储帖子

存储帖子模块主要包括去除重复帖子和存储有效帖子两个功能。

存储帖子功能图3-4所示。

图3-4存储帖子管理

3.2 数据库设计

在设计爬取系统时,数据库至关重要。数据库的稳定性、安全性和可恢复性在用户使用系统的过程中起着非常大的作用。因此,在进行系统研发时,选择合适的数据库极为重要。本系统采用的数据库是 MySQL5.6.17,当研发人员设计系统数据库之前,首先要对用户的需求以及将来可能会增加的数据需求进行充分的调查,并且有较为深入的思考。下面将介绍本系统的数据库结构以及主要的表(TABLE)。

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,本爬取系统的实体主要有两类:帖子和敏感词。本网上舆情爬取系统的实体属性如图3-5和3-6所示。

图3-5“帖子”属性描述图

图3-6“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。通过对功能的描述,数据库的分析以及各模块功能主要完成的任务等进行分析,能够使用户对本网站有着初步的理解,便于用户对系统的阶段。

4 系统实现过程

4.1 浏览帖子子模块

浏览帖子子模块分帖子展示和图表展示,这样可以从不同的维度展示爬取的帖子,以此来洞察舆情。

4.1.1 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^

4.1.2 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

4.2 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^

4.2.1 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)
```

4.2.2 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)
```

4.3 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)))
```

4.4 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())
)
```

4.5 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。

4.5.1 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:设置URL、爬取帖子相关信息、爬取下一页。

1)设置URL的功能函数,代码如下所示:

```
{88% : def __init__(self): }
# 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
```

4.5.2 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
{ 64% : 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
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

4.6 本章小结

{ 61% : 本章详对各个功能模块进行了较为详细的分析。 }同时对系统的操作过程进行了深入浅出的讲解,能让用户对本系统有深入的理解,便于用户熟练操作此系统。

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],也被称为浏览器/服务器和客户机/服务器模式,两种模式对于不同的场合,应用是截然不同的。{96% : 本系统采用B/S开发模式。 }

B/S模式是一种具有三层结构的技术系统:第一层,客户机和在整个系统参与者角色的用户界面之间。{ 59% : 浏览器将使HTML[8]的代码与交互网页, }并且可以允许用户在输入信息的应用形式的网页被发送到后台;在第二层中,将启动级联的Web服务器来处理用户请求,将生成的HTML代码。第三层时,MySQL负责协调颁发WEB管理数据库SQL请求。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,交互性、共享性强。

2)能随时进行浏览、查询等操作。

3)提供了更为安全、便捷的存取模式。

4)维护简单,维护成本低。

5)能够降低网络通信量

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页面布局DIV+CSS

本系统采用了DIV+CSS[9]进行页面布局。

{ 75% : DIV+CSS是“Web标准”最为常用的术语之一。 }CSS[10](层叠样式表)是层叠样式表的缩写,这是一种风格的语言如HTML文件。通俗来讲,DIV是用来建立网站结构和CSS来创建网站的性能(包括风格,美化等),很容易使HTML代码更简洁,同时也为日后的维护的网站。

使用DIV+CSS布局的优势主要体现在内容和形式相分离,减少了页面的代码,页面布局的控制更加灵活,{ 70% : 不仅提高了易用性,系统的可维护行和可扩展性, }大大降低成本的网站。

{ 67% : DIV+CSS的优点主要体现在以下几个方面: }

1)代码简洁易懂,提升了Web的浏览速度。

2)形式和内容相分离。

3)能很大程度上提高搜索引擎的搜索效率

4)易于维护

5.3 jQuery和Ajax技术

5.3.1 jQuery技术

本爬取系统网站使用了jQuery[11]技术,{89% : jQuery是一个优秀的JavaScript库。 }jQuery的文档非常齐全,提供了许多优美的插件,采用代码和内容相分离来设计网页。它的出现一定程度上解放了系统的开发者,提供了完美的用户体验工作。

jQuery具有的重要特性如下:

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

2)更易于设置函数,为所有的对象新增很多易用的设置函数。

3)重写了大部分函数,函数的性能有了较大幅度的提升。

5.3.2Ajax技术

本爬取系统网站在发展过程中还采用了Ajax技术,{ 62% : Ajax的英文缩写为Asynchronous JavaScript+XML[14], }该技术給用于提供了更为自然的体验,并且能够实现用户服务器的异步通信功能,从而可以无尽的自由请求/响应。{ 70% : Ajax不是一种技术。 }他集成了多种技术。{ 62% : Ajax包含的技术如下: }

{81% : 1)使用XMLHttpRequest进行异步数据搜索。 }

{ 69% : 2)使用XHTML和CSS标准。 }

3)使用XML和XLST的数据操作和交换。

{ 75% : 4)使用DOM对象的交互和动态显示。 }

5)将他们绑定在的Javascript[15]上。

5.4 scrapy框架

{ 71% : Scrapy[16]是抓取网站的数据,提取结构化数据的应用框架。 }{ 77% : 可应用于包括数据挖掘,信息处理和历史数据的存储等一系列程序。 }

所谓的网络爬虫,或在互联网上其他地方是一个方向性的数据采集过程中,{ 64% : 当然,这种说法是不是很专业,更专业介绍,按照特定的要求抓取HTML页面。 }首先定义一个爬取的起始URL,也就是start_urls,那么一般是这个网页内部会有很多URL,通过这些URL会连接到很多其他页面,{ 59% : 所以他从当前页面的URL连接, }便可以获取这些爬虫抓取队列,然后进入一个新的页面,然后递归上面的操作,其实只是说的深度还是广度遍历遍历相同。

{ 78% : Scrapy使用Twisted异步网络库来处理通信网络,框架清晰,并拥有很多种中间件接口,可以灵活地完成各种需求。 }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]是一个很小的Python的Web开发框架。 }{ 62% : 尽管Flask框架是一个很轻量级的框架, }但是其功能不容小觑。{ 69% : 毕竟,微架构的真正含义是简单而短。 }{ 64% : Flask依赖于两个外部库:WERKZEUG WSGI[19]工具集和Jinja2[18]模板引擎。 }{97% : Flask遵循“约定优于配置”以及合理的默认值的原则。 }

前台系统配置在display.py里,代码如下。

```
app = Flask(__name__)
@app.route('/') #根路径
def hello_world():
items = db.getItem()
words = db.getWord()
return render_template('index.html',
items = items,
words = words,
data = str(db.getCount_byMonth()))
@app.route('/search', methods=['POST'])
def Search():
content = request.form['search_content']
items = db.searchTitle(content)
words = db.getWord()
```

```
return render_template('index.html',
items = items,
words = words,
data = str(db.getCount_byMonth_byContent(content)))
@app.route('/search/^word^', methods=['POST','GET'])
def Search_Word(word):
items = db.searchTitle(word)
words = db.getWord()
return render_template('index.html',
items = items,
words = words,
data = str(db.getCount_byMonth_byContent(word)))
if __name__ == '__main__':
app.run(debug=True)
```

5.6 SQLAlchemy

{ 70% : SQLAlchemy的[20]是Python语言的一个开放源代码软件。 }{ 61% : SQLAlchemy提供了SQL工具包和对象关系映射(ORM)工具,SQLAlchemy的发行使用了MIT的许可证。 }

{89% : SQLAlchemy采用简单的Python语言,它为高效和高性能的数据库访问而设计,实现了较为完整的企业级持久性模型。 }{93% : SQL数据库的量级和性能着重于对象集合;而对象集合的抽象又重要于表和行。 }{84% : 因此,SQLAlchmey类似Java,Hibernate通过数据映射模型,而不是其他对象关系映射(ORM)所采用的Active Record模型。 }{84% : 然而,Elixir和declarative等其他可选插件允许用户使用声明的语法。 }

{ 76% : SQLAlchemy的代码如下。 }

{100% : def process_item(self, item, spider): }

```
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)
#创建Session类型
DBSession = sessionmaker(bind=engine)
#创建session对象
session = DBSession()
admin = Post(postMan = item['postMan'],
firstTime = item['firstTime'],
title = item['title'],
content = item['content'],
readCount = item['readCount'],
replyCount = item['replyCount'],
link = item['link'])
session.add(admin)
session.commit()
session.close()
```

return item

5.7 本章小结

本章节重点介绍了本系统使用的核心技术,分别为B/S模式、DIV+CSS、jQuiry+Ajax、scrapy框架、flask框架和SQLAlchemy软件,同时也对系统使用技术的概念和应用做了详细的说明,以便更好地理解与运用本系统。

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,我自学了一些前端技术,严格按照Web标准设计页面。采用div+css进行页面设计,不仅能加快网页的访问速度,还能提高网页的兼容性。同时,本系统采用jQuery和Ajax技术使得设计的网页更加美观、清晰。系统的后台部分,主要实现的功能是爬取帖子,存储帖子,其中爬取帖子我是用了scrapy框架和beautiful soup技术,存储帖子我使用了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,我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我的动手能力和逻辑思维能力,这次开发经历让我不仅学会使用Python语言,还让我领略到数据库设计对于系统开发的重大意义。

本系统的特色工作有:

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

{ 61% : 本课题研发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

本网上舆情爬取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,董永权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,我要向董永权导师致以最诚挚的谢意。董永权导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,我要向董永权老师真诚地感谢。

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。

本网上舆情爬取系统一定有很多的不足,恳请各位老师指正。