PaperFree检测报告简明打印版

相似度：25.13%
编　号：CPQKXGCJIBUS2R4Z
标　题：网上舆情爬取系统的设计与实现
作　者：李林
长　度：20618字符
时　间：2016-05-12 18:38:24
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：2.56%　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：0.52%　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
3. 相似度：0.48%　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
4. 相似度：0.37%　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
5. 相似度：0.37%　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
6. 相似度：0.37%　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
7. 相似度：0.31%　篇名：《基于建构主义的Internet网络成人教学模式的构建》
来源：《科技资讯》　年份：2014　作者：李峰
8. 相似度：0.3%　篇名：《用CSS + DIV重构南阳农校网站》
来源：《商情》　年份：2013　作者：吴延艳
9. 相似度：0.3%　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
10. 相似度：0.23%　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
11. 相似度：0.23%　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
12. 相似度：0.21%　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
13. 相似度：0.2%　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
14. 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
15. 相似度：0.19%　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
16. 相似度：0.19%　篇名：《基于环保电子政务信息资源整合研究》
来源：《中国电子商务》　年份：2014　作者：许聪雄
17. 相似度：0.18%　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
18. 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
19. 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
20. 相似度：0.16%　篇名：《浅析企业进销存管理网站的设计》

来源：《中外企业家》 年份：2013 作者：徐枫
21. 相似度：0.16% 篇名：《网络技术在多媒体教学中的应用》
来源：《剑南文学：经典阅读》 年份：2013 作者：潘东
22. 相似度：0.16% 篇名：《Internet背景下的教师继续教育》
来源：《西北成人教育学报》 年份：2013 作者：秦万祥
23. 相似度：0.16% 篇名：《实验室固定资产管理系统分析与设计》
来源：《中国科技博览》 年份：2014 作者：蓝杨平
24. 相似度：0.14% 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
25. 相似度：0.12% 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
26. 相似度：0.11% 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
27. 相似度：0.11% 篇名：《NET课程辅助实践教学系统的设计与实现》
来源：《科教文汇》 年份：2014 作者：黄静
28. 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
29. 相似度：0.1% 篇名：《面向电网企业的 ERP 访问控制及权限管理研究》
来源：《科技管理研究》 年份：2014 作者：林友谅
30. 相似度：0.09% 篇名：《高校学籍管理系统的设计与实现》
来源：《中国电子商务》 年份：2013 作者：高赫鑫
31. 相似度：0.09% 篇名：《基于知识管理的人力资源管理系统的设计与实现》
来源：《国防交通工程与技术》 年份：2013 作者：魏敏
32. 相似度：0.09% 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
33. 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
34. 相似度：0.09% 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
35. 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：3.24% 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：2.21% 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
3. 相似度：2.14% 标题：《【scrapy】学习Scrapy入门 - 简书》
来源：http://www.jianshu.com/p/a8aad3bf4dc4
4. 相似度：2.06% 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
5. 相似度：1.41% 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
6. 相似度：1.28% 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
7. 相似度：1.13% 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
8. 相似度：0.6% 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
9. 相似度：0.52% 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
10. 相似度：0.28% 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：0.22% 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
12. 相似度：0.1% 标题：《php语言_360百科》
来源：http://baike.so.com/doc/7103846-7326839.html

全文简明报告

网上舆情爬取系统的设计与实现

摘 要

{82% : 随着Internet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,采用B/S设计模式,使用python语言以及Mysql数据库进行开发。系统包括两大部分: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

{86% : With the rapid development of Internet 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. The system for the needs of development, the use of B / S design mode, using python language and Mysql database development. The system consists of two parts: the foreground and background display module crawling module, which shows the front desk module consists of five sections: Posts impressions (the poster, the title and content), graphics display, search articles, sensitive words management and setting up crawling website the URL. 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
表序号 表名称 页码

{100%：随着互联网技术的发展与应用的普及,网络作为信息的载体,已经成为社会大众参与社会生活的一种重要信息渠道。 }{100%：由于互联网是开放的,每个人都可以在网络上发表信息,内容涉及各个方面。 }{100%：小到心情日志,大到国家大事。 }{97%：互联网已成为思想文化信息的集散地,并具有传统媒体无法相比的优势:便捷性、虚拟性、互动性、多元性。 }

{97%：网络舆情热点通常形成迅速,多是人们对于日常生活中的各种问题发表的各种意见,评论,态度,情绪等,随着事件的发展而变化,是反映社会热点的重要载体之一。 }

随着网上舆情的深入发展,网上舆情越来越复杂,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

{ 76%：Python[1]是一种面向对象、解释型计算机程序设计语言,由Guido van Rossum在1989年发明,于1991年发行第一个公开版。 }

Python是最纯粹的自由软件, 源代码和解释器CPython遵循 GPL协议 。

{90% : Python语法很简洁清晰, }其特色之一是强制用空白符(white space)作为语句缩进。

{ 72% : Python具有丰富和强大的库。 }{96% : 它常被称为胶水语言,能够把用其他语言制作的各种模块(尤其是C/C++)很轻松地联结在一起。 }{100% : 常见的一种应用情形是,使用Python快速生成程序的原型(有时甚至是程序的最终界面),然后对其中有特别要求的部分,用更合适的语言改写,比如3D游戏中的图形渲染模块,性能要求特别高,就可以用C/C++重写,而后封装为Python可以调用的扩展类库。 }{100% : 需要注意的是在您使用扩展类库时可能需要考虑平台问题,某些可能不提供跨平台的实现。 }

1.2.2 MySQL数据库简介

{85% : MySQL[2]是一种关系数据库管理系统, }同时也是当今最具影响力的数据库管理系统,作为一个数据库服务器,MySQL的最大优点体现在速度和健壮性两大方面。{ 76% : MySQL之所以称为是关系数据库管理系统, }是因为它并非将所有的数据都不加分类的堆放在一起,而是把数据存放在一些分立的表格中。因此,它在速度同时也提高了存取的灵活性。

对比其他数据库MySQL具有许多高级特征。其特征如下:

1)性能

由于MySQL没有线程进行创建开销,所以MySQL会在以下方面更快一些:

{ 69% : ①DROPE TABLE以及CREATE TABLE。 }

②在不属于一个索引的东西上SELECT。(很容易扫描单个表。)

{ 63% : ③有很少的键和列插入的简单表的插入操作。 }

2)磁盘空间效率

{ 62% : MySQL可以创建占据内存很小的表。 }譬如,MEDIUMINT的长度只有三个字节。在记录繁多的情况下,{ 62% : 每个记录即使是节省一个字节,也显得极其重要。 }

3)稳定性

MySQL具有保存庞大数目的记录、运行速度同类产品中最快、可移植性高以及安装过程简单小巧等一系列优点,这使得一般的中小型的网站都会将MySQL作为数据库。由于业界所称的“LAMP”组合都是开源软件,因此,这种方式可以很廉价的建立一个免费且稳定的网站系统。

MySQL能够支持众多OS,譬如:FreeBSD、Mac以及Windows,{ 71% : 并且为许多编程语言提供API。 }为了保证代码的可移植性,{ 59% : MySQL使用面向对象的编程语言编写, }采用用很多种Editor进行测试。MySQL不但能够支持多线程、提升SQL的查询速度,同时又能优化查询算法。

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,WampServer2.5[这个为什么不引用一下?]

数据库:MySQL

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、整体设计思路以及相关技术等。

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、发帖时间,帖子标题,帖子内容,帖子链接,阅读和回复的数量),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。主要详述了系统的背景、意义、开发工具的选用和介绍、本文的研究内容和主要贡献。

第2章:需求分析。主要介绍了系统的功能需求和相关的性能需求。

{ 74% : 第3章:系统功能模块设计。 }介绍了系统功能模块的划分以及数据库的设计。

第4章:系统实现流程。介绍了系统各个模块,以及相关模块的代码实现。

第5章:关键技术。介绍了网上舆情爬取系统的配置和关键技术。

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[呼应]

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[主谓结构 呼应]

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)用户用例图

{ 59% : 用户用例图表述了用户的操作权限。 }用户可以进行帖子展示、帖子搜索、敏感词管理和URL设置[呼应],用户用例图如图2-1所示。

图2-1 用户用例图

(2)管理员用例图

管理员用例图表示了管理员的操作权限。管理员可以爬取帖子,存储帖子。{ 59% : 管理员用例图如图2.2所示。 }

图2-2管理员用例图

2.2 性能需求

2.2.1 系统的软件环境

ï•数据库服务器。

MySQL+WampServer

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

2.2.2 系统硬件环境

PC机

2.2.3 系统的性能要求

1)安全性要求:本系统具有权限设置。{ 68% : 不同的用户有着不同的权限。 }

2)磁盘容量要求:本网站是基于B/S的架构,所以,在存储容量方面,本网站所占内存容量甚小,系统文件大概需要8M,由于数据库系统所占内存也不是很大,其文件占用的空间也微不足道。但是,本网上舆情爬取系统中的后台管理模块需要一定的内存空间。需要将一个论坛上某个板块所有的帖子全部爬取下来,会占用内存空间。

3)适应性要求:本系统要求功能模块清晰,模块间内聚性强,耦合性弱,能够使用户在很短的时间内熟悉系统的整个操作流程,具有良好的用户体验。

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准确、有效开发项目的前提条件和基础,可行性研究报告能够使开发人员在系统研发的初期就能发现系统在需求方面的不足,如此可以避免耗费人力、物力和财力等许多方面的困难。综上所述,可行性研究及分析在软件开发的整个过程中扮演着极为重要的角色。

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[这几个图为什么这么画??][模仿的邱盈盈的画的,系统业务流程。]所示。

[字太大了]

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)经济的可行性:经济可行性分析当中,最为重要的内容就是效益与成本的对比分析。如果开发一个系统在经济方面就已经不适用,则开发这种系统完全没有必要。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)技术可行性:{ 63% : 本系统主要采用B/S模式设计开发。 }这种架构具有极好的继承性、可扩展性以及开放性,便与系统的开发与维护。此外,本系统采用目前最为流行的scrapy框架已经beautiful soup[4]技术,维护起来方便简单。

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。[呼应,下同]

3.1.1系统功能模块结构图

网上舆情爬取系统前台的功能模块及模块间的关系如图3-1所示。

图3-1系统前台功能模块

网上舆情爬取系统后台的主要功能模块以及模块间的关系如图3-2所示。

图3-2系统后台功能模块

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

在设计管理系统时,数据的管理至关重要。数据库的安全性、稳定性和可恢复性在用户使用系统的过程中起着很大作用。所以,在进行系统开发时,选择合适的数据库极为重要。本系统采用的数据库为 MySQL5.6.17,当研发人员设计系统数据库的时候,首先要对用户的需求以及将来可能会增加的数据需求有着充分的调查和了解。下面将介绍本系统的数据库结构以及创建的表。

3.2.1 实体

数据库中的实体可以指的是人也可以指物。经分析,本网站主要实体有六类: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(jiehemei)

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。

[换个数据多的图]

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[呼应]

设置URL子模块[呼应]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')
[中间断行前面都有注释,下同]
for site in sites:
item = PostItem()
item['title'] = site.css('.s.xst::text').extract_first() # 取出其中的文本,q取出第一个[q是什么?]
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
yield item
```

4.5.2 存储帖子模块[再加注释,代码太长]

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)
{ 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
if flag==0:
```

session.add(admin) #加入数据库

session.commit()

session.close() #提交并关闭

return item

4.6 本章小结

本章详细分析了对各个模块功能。对系统的操作过程进行了详细的讲解,能让用户对本系统有着进一步的理解,便于用户熟练操作此系统。

5 关键技术

5.1系统开发模式

网络程序开发模式[5]有两种:C/S[6]以及B/S[7],{ 60% : 也叫做客户机/服务器以及浏览器/服务器模式, }这两种模式用于不同的场合,各有千秋。{ 63% : 本系统采用的是B/S模式。 }

B/S模式是一个具有三层结构的技术体系:第一层中,客户机扮演者整个系统和用户间接口的角色。浏览器将使HTML[8]代码转化为具有交互功能的Web页面,能允许用户在Web页面的申请表里输入信息发送至后台;在第二层中,Web服务器会启动相关程序处理用户请求,生成HTML代码。第三层中,MySQL负责协调WEB所发的SQL请求来管理数据库。B/S模式结构如图5-1所示:

图5-1 B/S模式结构图

B/S模式的优点主要有:

1)开发容易,共享性、交互性强。

2)能随时进行查询、浏览等操作。

3)提供了更为安全的存取模式。

4)维护简单。

5)能够降低网络通信量

6)对于相同任务,采用C/S模式比B/S模式速度更快。

综上所述,本数据库精品课程管理系统采用B/S模式设计实现

5.2页面布局DIV+CSS

{ 70% : 本系统采用DIV+CSS[9]布局页面。 }

{ 76% : DIV+CSS是“Web标准”中常用术语之一。 }{ 70% : CSS[10](层叠样式表单)英文缩写是Cascading Style Sheets, }它是表示HTML等文件式样的语言。通俗来讲,{ 70% : DIV是用来搭建网站结构的,而CSS是用来创建网站的表现(包括样式、美化等), }这样便于使HTML代码更为简洁,同时也便于日后网站的维护。

本系统使用DIV+CSS布局的优势主要体现在形式内容相分离,缩减了页面代码,灵活的控制页面布局,{ 59% : 不仅提高系统的易用性及扩展性, }同时降低了网站改版成本。

DIV+CSS的优势主要体现在一下几个方面:

1)代码简洁,能提升Web的浏览速度。

2)内容和形式相分离。

3)能加快搜索引擎的搜索效率

4)易于改版和维护

5.3 jQuery和Ajax技术

5.3.1 jQuery技术

本爬取系统网站在开发过程中使用了jQuery[11]技术,{ 73% : jQuery是一个优秀的javascript类库。 }jQuery文档说明十分齐全,提供很多完美的插件,能使用户设计页面的代码与内容相分离。它的使用极大地简化了系统开发人员工作,具有完美的用户体验。

jQuery具有的重要特性如下:

1)对Ajax进行了改进,引入很多Ajax[12]以及JSON[13]处理方面的更新。

2)更易于设置函数(sett function),为所有的对象新增很多易用的设置函数。

3)重写大部分早起函数,{ 59% : 常用方法的性能有了大幅度的提升。 }

5.3.2Ajax技术

本爬取系统网站在开发过程中也使用了Ajax技术,{ 60% : Ajax的英文简写为Asynchronous JavaScript+XML[14], }该技术为用户提供了自然的浏览体验的同时还提供了和服务器进行异步通信的功能,{ 67% : 能够让用户从无尽的请求/响应中解脱出来。 }{ 59% : Ajax并不是一种技术。 }他是几种技术的综合。{ 62% : Ajax包含的技术如下: }

{ 78% : 1)使用XMLHttpRequest进行异步数据检索。 }

2)使用CSS标准和XHTML。

3)使用XLST和XML的数据操作和交换。

4)使用文档对象模型的交互和动态显示。

5)将他们绑在一起的Javascript[15]。

5.4 scrapy框架

{93% : Scrapy[16]是一个为了爬取网站数据,提取结构性数据而编写的应用框架。 }{98% : 可以应用在包括数据挖掘,信息处理或存储历史数据等一系列的程序中。 }

{100% : 所谓网络爬虫,就是一个在网上到处或定向抓取数据的程序,当然,这种说法不够专业,更专业的描述就是,抓取特定网站网页的HTML数据。 }{100% : 抓取网页的一般方法是,定义一个入口页面,然后一般一个页面会有其他页面的URL,于是从当前页面获取到这些URL加入到爬虫的抓取队列中,然后进入到新页面后再递归的进行上述的操作,其实说来就跟深度遍历或广度遍历一样。 }

{100% : Scrapy 使用 Twisted这个异步网络库来处理网络通讯,架构清晰,并且包含了各种中间件接口,可以灵活的完成各种需求。 }scrapy框架结构图5-2如下。

图5-2 scrapy框架结构图

5.5 flask框架

{92% : Flask[17]是一个基于Python的微型的web开发框架。 }{100% : 虽然Flask是微框架,不过我们并不需要像别的微框架建议的那样把所有代码都写到单文件中。 }{100% : 毕竟微框架真正的含义是简单和短小。 }{ 61% : Flask 依赖两个外部库:Jinja2[18]模板引擎Werkzeug WSGI [19]工具集。 }{100% : Flask遵循“约定优于配置”以及合理的默认值原则。 }

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

{93% : SQLAlchemy[20]是Python编程语言下的一款开源软件。 }{100% : 提供了SQL工具包及对象关系映射(ORM)工具,使用MIT许可证发行。 }

{100% : SQLAlchemy “采用简单的Python语言,为高效和高性能的数据库访问设计,实现了完整的企业级持久模型”。 }{100% : SQLAlchemy的理念是,SQL数据库的量级和性能重要于对象集合;而对象集合的抽象又重要于表和行。 }{100% : 因此,SQLAlchmey采用了类似于Java里Hibernate的数据映射模型,而不是其他ORM框架采用的Active Record模型。 }{100% : 不过,Elixir和declarative等可选插件可以让用户使用声明语法。 }

{ 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,我自学网页制作技术,并按照Web标准,严格设计页面。采用div+css设计页面,不仅能加快网页的访问速度还能提高网页的兼容性。同时,采用jquery和ajax使得设计的网页更加清晰、美观。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极其复杂的过程。本此设计极大的提高了我的逻辑思维以及动手能力,这次的开发经历让我不仅学会使用一门新的编程语言,还让我领略到数据库设计对于系统开发的重要意义。

本系统的特色工作有:

1)系统功能模块清晰,操作简单。

2)前台展示多维度,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,便于系统后期的维护、更新以及功能扩展。

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。

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[另起一页]

[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. JSP em dash a research signal processor in josephson technology[C]. IBM Journal of Research and Development,Vol24, No2,1980:222-235.

[8]Tsui,Frank F. JSP em dash a research signal processor in josephson technology[C]. IBM Journal of Research and Development,Vol24, 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

毕业设计体会

本次的课题设计,全面提高了我在系统研发过程中的动手能力,与此同时我的编程能力也有了极大的提高。

历时两个月,我的毕业设计终于接近尾声。毕业设计不仅能够巩固我本科期间的所学知识,而且也是对自己科研能力的一种锻练。毕业设计不同于本科期间的课程设计,它涉及到的知识面更为宽泛,开发过程更为繁杂。本次的设计,让我明白大学期间所学习的理论知识只有上升到动手操作的层面,才会使学习这一过程变得更有意义。这段时间以来,我深刻体会到了知识积累是一个漫长而复杂的过程。学习技巧、研发能力的提高不是并非一蹴而就。

在做设计的过程中,我上网搜索了大量的资料,同时也观看了相关系统开发的教学视频。因之前没有研发整个系统的开发经验,因此刚上手时有些手足无措。通过向老师请教等以及与同学们的交流,使我学到了大量的专业知识。虽然设计过程中也遇到了许多问题譬如python乱码问题,scrapy框架安装和建立等问题,也曾失落、沮丧过,但是在指导老师的悉心指导以及同学们的热情帮助下,我克服了大大小小的困难,一步步坚持走到最后。在完成毕业设计的过程中,我和 同学互相交流,取长补短,家一起商量、解决开发过程中遇到的问题,积极思考并采纳不同的意见,不仅使同学之间的相互关系更为紧密,也让我拥有更好的理解、运用知识的能力。

在整个毕业设计的过程中,我学到了许多课堂上不曾设计的东西,同时也让我对自己的动手能力更有信心。本系统在界面设计方面还存在些许不足,很多功能模块还有待开发,但是在此过程中积累的知识才是我获得的最为宝贵的财富,这将是我受益终身。

致谢

本论文在董永权老师的悉心指导下完成。从论文的选题、写作以及成稿,董永权老师给与我极大地帮助。在系统开发前期,老师对我的系统需求分析提出了很多宝贵的建议,在后期,老师仔细审查并修改了我的论文,导师丰富的科研实践经验以及严谨的治学态度给与我极大的鼓励。导师严谨的治学态度以及对计算机科学研究事业孜孜不倦的追求精神,给了我无穷的启发与思考。在此,我要向董永权老师致以最诚挚的谢意。

感谢江苏师范大学智慧教育学院的全体老师,衷心的感谢他们能为我们提供良好的学习环境以及周到、细致的学习计划。

通过本次论文的撰写,我系统的学习了系统开发方面的前言理论知识并获得了宝贵的开发经验,这对于我今后攻读硕士研究生亦或是工作来说,无疑是个不可多得的锻炼机会。本网站在细节方面可能存在些许欠缺,论文中涉及到的一些技术介绍可能还存在不足,恳请老师们指正。