PaperTime检测报告简明打印版

相似度：30.42%
编　号：VSQMPCMO3UQ6W69V
标　题：王浩毕业设计
作　者：-
长　度：11234字符
时　间：2018-05-09 19:00:13
比对库：本地库（学术期刊、学位论文、会议论文）；PaperTime云论文库；互联网

本地库相似资源（学术期刊、学位论文、会议论文）
1. 相似度：1.15%　篇名：《Web用户界面研究》
来源：《电脑编程技巧与维护》　年份：2013　作者：阮丽红
2. 相似度：0.73%　篇名：《一种基于用Python网络编程的P2P实现》
来源：《电脑编程技巧与维护》　年份：2013　作者：郝亚超
3. 相似度：0.38%　篇名：《Redis数据库在微博系统中的实践》
来源：《厦门城市职业学院学报》　年份：2012　作者：唐诚
4. 相似度：0.29%　篇名：《试论动态开发语言Python研究》
来源：《电脑编程技巧与维护》　年份：2014　作者：卫启哲
5. 相似度：0.26%　篇名：《图书馆光盘管理系统》
来源：《中国电子商务》　年份：2013　作者：贺铭德
6. 相似度：0.25%　篇名：《基于Android智能手机系统平台的新闻接收客户端的设计与实现》
来源：《计算机与现代化》　年份：2012　作者：阙锋
7. 相似度：0.23%　篇名：《浅论NOSQL数据库与关系数据库》
来源：《中国电子商务》　年份：2017　作者：鲍飞字
8. 相似度：0.21%　篇名：《软件测试在软件开发过程中的应用研究》
来源：《硅谷》　年份：2016　作者：焦胜男
9. 相似度：0.19%　篇名：《浅谈Session对象在用户登录页面中的应用》
来源：《电子技术与软件工程》　年份：2015　作者：周广深
10. 相似度：0.18%　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
11. 相似度：0.16%　篇名：《从UEditor谈Web编辑器技术》
来源：《程序员》　年份：2013　作者：战毅
12. 相似度：0.16%　篇名：《基于ueditor富文本编辑器的实验报告在线编辑系统设计》
来源：《考试周刊》　年份：2013　作者：吕景美
13. 相似度：0.15%　篇名：《高校精品课程网站管理与设计》
来源：《中国科教创新导刊》　年份：2014　作者：吴长忠
14. 相似度：0.14%　篇名：《海量数据采集系统的研究》
来源：《沈阳工业大学硕士论文》　年份：2009　作者：于潇宇
15. 相似度：0.12%　篇名：《关于软件测试课程的初步探究》
来源：《信息技术与信息化》　年份：2013　作者：张亿军
16. 相似度：0.12%　篇名：《高职院校图书馆特色数据库系统开发与实施》
来源：《阜阳职业技术学院学报》　年份：2013　作者：韦建国
17. 相似度：0.10%　篇名：《大学生社会实践活动案例分析报告》
来源：《中国校外教育：下旬》　年份：2017　作者：蒙铁
18. 相似度：0.09%　篇名：《关于电子商务网站建设中的安全问题探究》
来源：《计算机光盘软件与应用》　年份：2016　作者：张艳秋

PaperTime云论文库(知网，万方，维普，百度文库等镜像）
1. 相似度：1.09%　标题：《定向网络爬虫 开题报告_百度文库》
来源：http://wenku.baidu.com/view/1a1ccf6a6529647d272852f2.html
2. 相似度：0.99%　标题：《第3章 测试用例设计 - 道客巴巴》

来源：http://www.doc88.com/p-973196687909.html
3. 相似度：0.26% 标题：《Python编程语言的教程_百度文库》
来源：https://wenku.baidu.com/view/f841badaa58da0116c17496b.html
4. 相似度：0.11% 标题：《主题搜索引擎的关键技术研究与实现_CNKI学问》
来源：http://xuewen.cnki.net/CMFD-2010163724.nh.html

互联网相似资源(博客，百科，论坛，新闻等）

1. 相似度：1.64% 标题：《Apache Maven 概述_w3cschool》
来源：https://www.w3cschool.cn/maven/u7oe1ht0.html
2. 相似度：1.64% 标题：《蜗牛学院-在线课堂-课程：maven工具讲解》
来源：http://www.woniuxy.com/course/40
3. 相似度：1.63% 标题：《测试用例实例- Test Dancer的个人空间- 51Testing软件测试网...》
来源：http://www.51testing.com/html/42/131542-19547.html
4. 相似度：1.54% 标题：《软件工程专业毕业论文--网络爬虫设计与实现+开题报告.doc》
来源：https://max.book118.com/html/2013/0322/3477103.shtm
5. 相似度：1.35% 标题：《Python 爬虫初探(一)-布布扣-bubuko.com》
来源：http://www.bubuko.com/infodetail-1960644.html
6. 相似度：1.12% 标题：《从JAVA传数据到Python 脚本用Jython的解决方案 - CSDN博客》
来源：https://blog.csdn.net/u010088415/article/details/73647892
7. 相似度：1.06% 标题：《爬虫爬取新闻(一) - 水月灵心 - 博客园》
来源：https://www.cnblogs.com/Hebe/p/5180045.html
8. 相似度：1.06% 标题：《爬虫爬取新闻(二)》
来源：http://www.mamicode.com/info-detail-1201989.html
9. 相似度：0.97% 标题：《网络爬虫设计与实现-毕业论文.docx》
来源：https://max.book118.com/html/2017/0804/125872746.shtm
10. 相似度：0.89% 标题：《爬虫技术和爬虫需求现状和展望 - CSDN博客》
来源：https://blog.csdn.net/zhongshanb/article/details/46503489
11. 相似度：0.80% 标题：《intelliJ IDEA全称是什么?_百度知道》
来源：https://zhidao.baidu.com/question/1756197118458974988.html
12. 相似度：0.80% 标题：《IDEA|IntelliJ IDEA(Java开发工具)下载 12.1.4 官方稳定版-新云...》
来源：https://www.newasp.net/soft/16081.html
13. 相似度：0.74% 标题：《Git版本控制管理(第2版) (豆瓣)》
来源：https://book.douban.com/subject/26341974/
14. 相似度：0.74% 标题：《Git 全部文档 OPEN开源文档》
来源：http://www.open-open.com/doc/list/282?o=p
15. 相似度：0.62% 标题：《Mybatis- 基础知识 - CSDN博客》
来源：https://blog.csdn.net/ganquanzhong/article/details/80193559
16. 相似度：0.59% 标题：《python3下中文编码问题 - 博客频道 - CSDN.NET》
来源：https://blog.csdn.net/gdsfga/article/details/69258680?locationNum=14&fps=1
17. 相似度：0.56% 标题：《In recent years, with the rapid development of the ...》
来源：http://tool.xdf.cn/juku/1076265.html
18. 相似度：0.53% 标题：《什么是304页面,出现304页面该怎么办_百度知道》
来源：https://zhidao.baidu.com/question/500705890662672004.html
19. 相似度：0.47% 标题：《Jquery怎么将一个object对象转换成json字符串_百度知道》
来源：https://zhidao.baidu.com/question/1757543876424859228.html
20. 相似度：0.46% 标题：《Tomcat 6.0下载|Apache Tomcat 6 V6.0.53 - 快乐无极软件园》
来源：http://www.oyksoft.com/soft/5890.html
21. 相似度：0.44% 标题：《SSM三大框架整合详细教程(Spring+SpringMVC+MyBatis)(转..._博客园》
来源：https://www.cnblogs.com/Joetao/articles/4544572.html
22. 相似度：0.43% 标题：《Spring+SpringMVc+Mybatis实现数据库查询 - Yrion - 博客园》
来源：https://www.cnblogs.com/wyq178/p/6815373.html
23. 相似度：0.43% 标题：《at the same time the computer at any time storage ...》
来源：http://blog.sina.com.cn/s/blog_c0bc3bb20101eema.html
24. 相似度：0.42% 标题：《开源爬虫 ── 专业、强大的万维网资源定向抓取、爬抓工具》
来源：http://www.mamicode.com/info-detail-1095515.html

25. 相似度：0.41% 标题：《从Larbin看互联网爬虫设计 - CSDN博客》
来源：https://blog.csdn.net/zhz_2V/article/details/1642931
26. 相似度：0.37% 标题：《java调用python方法的库jython介绍及使用实例- 阿里云》
来源：https://www.aliyun.com/jiaocheng/459422.html
27. 相似度：0.37% 标题：《Python的Java实现 - Jython2.5 a1 发布 - 惠然如是的日志 - 网易博客》
来源：http://blog.163.com/jimmy_5328/blog/static/628649200951103416651/
28. 相似度：0.37% 标题：《due to the limitation of traditional web search engine,_360英文》
来源：http://en.so.com/s?q=due+to+the+limitation+of+traditional+web+search+engine%2C
29. 相似度：0.34% 标题：《WEB前端教程(HTML/CSS/JavaScript/表单验证/JQuery/Bootstrap)_... 》
来源：https://ke.qq.com/course/237980
30. 相似度：0.29% 标题：《手把手Maven搭建SpringMVC+Spring+MyBatis框架(超级详细版) - ...》
来源：http://wosyingjun.iteye.com/blog/2249750
31. 相似度：0.29% 标题：《log4j日志管理系统简单使用说明 - open java project(转载) - CSDN博... 》
来源：https://blog.csdn.net/carefree31441/article/details/4679075
32. 相似度：0.29% 标题：《maven - 蓝蝶飞扬的个人页面 》
来源：https://my.oschina.net/xiaoq6427/blog/529723
33. 相似度：0.28% 标题：《mysql 中的变量可以存放数组吗_百度知道》
来源：https://zhidao.baidu.com/question/1963911155075225140.html
34. 相似度：0.27% 标题：《数据库持久层框架iBatis、myBatis、Hibernate对比 - CSDN博客》
来源：https://blog.csdn.net/nicolas_huan/article/details/67632852
35. 相似度：0.25% 标题：《求一个建议的新闻管理系统,能实现新闻的添加删除修改等..._百度知道》
来源：https://zhidao.baidu.com/question/529966928.html
36. 相似度：0.24% 标题：《java是属于什么语言?》_百度知道》
来源：https://zhidao.baidu.com/question/589338198.html
37. 相似度：0.24% 标题：《Mybatis与Hibernate的详细对比 - 哼哼哈哈二将 - 博客园》
来源：https://www.cnblogs.com/snowbook/p/5410697.html
38. 相似度：0.24% 标题：《MyBatis和Hibernate的比较 - CSDN博客》
来源：https://blog.csdn.net/ylm670444337/article/details/50302207
39. 相似度：0.23% 标题：《springMVC工作原理 - CSDN博客》
来源：https://blog.csdn.net/chenleixing/article/details/47143813
40. 相似度：0.21% 标题：《【javascript城市选择器】javascript城市选择器插件源..._CSDN博客》
来源：https://blog.csdn.net/zhou1988217/article/details/7470347
41. 相似度：0.21% 标题：《js实现的切片效果图片切换幻灯片特效源码 下载-脚本之家》
来源：http://www.jb51.net/jiaoben/216753.html
42. 相似度：0.21% 标题：《Java调用Python并传递参数(爬虫8) - CSDN博客》
来源：https://blog.csdn.net/XiaoFengLiuYu/article/details/70211109
43. 相似度：0.20% 标题：《IntelliJ IDEA 如何设置快捷键及快捷键风格 - CSDN博客》
来源：https://blog.csdn.net/xyw591238/article/details/51742852
44. 相似度：0.19% 标题：《redis 缓存技术与memcache的最大区别-布布扣-bubuko.com》
来源：http://www.bubuko.com/infodetail-237702.html
45. 相似度：0.18% 标题：《13个Python GUI库_搜狐科技_搜狐网》
来源：http://www.sohu.com/a/231007796_494939
46. 相似度：0.17% 标题：《Java是解释型还是编译型?_百度知道》
来源：https://zhidao.baidu.com/question/1964378202710559020.html
47. 相似度：0.16% 标题：《jquery ajax异步获取数据 怎么么做?有点简单例子么?_百度知道》
来源：https://zhidao.baidu.com/question/1819688468761039308.html
48. 相似度：0.16% 标题：《使用IntelliJ IDEA编写自己的第一个java程序 - CSDN博客》
来源：https://blog.csdn.net/nextljp/article/details/77949597
49. 相似度：0.16% 标题：《redis做缓存,怎么更新里面的数据_百度知道》
来源：https://zhidao.baidu.com/question/373188803899151924.html
50. 相似度：0.15% 标题：《log4j使用详解--创建自己的日志系统 - CSDN博客》
来源：https://blog.csdn.net/sain615/article/details/6578928
51. 相似度：0.14% 标题：《教师主页系统使用手册学习.doc》
来源：https://max.book118.com/html/2017/0710/121610291.shtm
52. 相似度：0.14% 标题：《html中使用jQuery和css,js的方法_百度知道》

来源：https://zhidao.baidu.com/question/552487593132225132.html
53. 相似度：0.14% 标题：《Jython:无缝地结合了Java类与Python的新贵_Jython_Python_java_课...》
来源：http://www.kokojia.com/article/11793.html
54. 相似度：0.14% 标题：《MyBatis和Hibernate相比,优势在哪里? - 知乎》
来源：https://www.zhihu.com/question/21104468
55. 相似度：0.12% 标题：《Linux Operational Notes - Dig All Possible》
来源：http://www.z-dig.com/
56. 相似度：0.11% 标题：《成都党员e家》
来源：http://www.cddyjy.com/website/contents/7/24462.html
57. 相似度：0.11% 标题：《B/S结构简介 - CSDN博客》
来源：https://blog.csdn.net/mxq007/article/details/1710164
58. 相似度：0.09% 标题：《JSP新闻网站系统论文.doc文档全文免费阅读、在线看》
来源：https://max.book118.com/html/2016/0110/33019068.shtm
59. 相似度：0.09% 标题：《Nutch、heritrix、crawler4j优缺点 - zzm - ITeye博客》
来源：http://m635674608.iteye.com/blog/2220454

全文简明报告

基于python网络爬虫的实时新闻系统

Real Time News System Based on Python Crawler

基于python网络爬虫的实时新闻系统

{ 63% : 【摘要】近年来网络技术蓬勃发展， }网络信息量爆发式的剧增，与此同时，计算机硬件技术不断进步，运行速度越来越快，网络的带宽和信息容量也是直线增长。各种数据挖掘，数据分析技术层出不穷。由于传统网络搜索引擎的局限性，在特定场景下的网络数据抓取变得越发重要。

为了解决用户对于特定场景下的网络信息检索需求，专业性的网络爬虫技术应运而生，为特定的用户提供专业级的网络信息检索服务。本文以爬取新浪新闻，网易新闻等门户新闻网站为例，对开发python网络爬虫进行深入研究。

【关键词】网络爬虫；python；数据分析；搜索引擎

Real Time News System Based on Python Crawler

{ 67% : 【Abstract】With the rapid development of network technology in recent years, } the explosion of network information has increased dramatically.{ 57% : At the same time, the computer hardware technology is progressing, } the speed is speeding up, and the bandwidth and information capacity of the network is also increasing in a straight line. All kinds of data mining and data analysis technologies emerge in endlessly.{ 59% : Due to the limitation of traditional web search engine, } data crawling in specific scenarios becomes more and more important.

In order to solve the user ‘s demand for network information retrieval in a specific scene, professional web crawler technology emerges as the times require, providing professional network information retrieval services for specific users. This paper takes Sina News, NetEase news and other portal news websites as an example to further explore Python web crawler.

【Key words】Web crawler; Python; Data analysis

1 引言

1.1研究背景

{91% : 互联网是一个巨大的非结构化数据库，将数据有效的检索并呈现出来有着巨大的应用前景。 }{96% : 搜索引擎作为一个辅助人们检索网络信息的工具成为用户访问万维网的入口和指南。 }{ 71% : 但是，这些通用的搜索引擎在一定程度上有一定的局限性。 }不同领域不同背景的用户在使用网络搜索引擎的时候有一定的局限性。{ 60% : 不同背景和领域的用户对于网络信息的检索往往有不同的需求， }因此通过网络搜索引擎活获得的信息有比较大的冗余。

因此，用户普遍渴望获取符合她们个性化需求的检索工具。{ 55% : 为了解决用户的对于信息检索的需求，参照与网络爬虫的成功的模式，对网络爬虫进行深入研究，从而使网络爬虫的产品拥有更贴合用户需求的产品，提供满足用户对于网络信息检索的个性化网络爬虫应用。 }

1.2研究现状

{ 58% : 对于网络爬虫的研究开始于上世纪90年代， }伴随网络搜索引擎的兴起，网络爬虫因运。{ 67% : 网络爬虫时网络搜索引擎的主体部分， }网络爬虫技术也是日趋成熟。{ 60% : 开源社区里比较成熟的爬虫项目有Nutch, Heritrix, }{ 58% : Crawler4j等。而生 }

{99% : 数据时代，对大数据的分析应当成为一个行业，数据拥有者应该开放数据的分析接口，让数据的价值释放，而爬虫开发者，很多时候是数据分析者（最起码是个数据清洗和筛选者）。 }为不同的用户提供符合不同需求的网络爬虫应用，是网络爬虫发展的趋势。

1.3研究内容

本文以爬取新浪新闻，网易新闻等门户新闻网站为例，对开发python网络爬虫进行深入研究。

{ 59% : 本爬虫系统采用python语言作为脚本， }后端使用JakartaEE等技术。{ 58% : Python语言系统类库十分强大，语法简单，并且拥有高效率多样化高层的数据结构， }代码结构比较清晰易懂，而且是简单的面相对象编程，切实现功能比较容易。{ 61% : 爬虫通过线程池及任务队列实现了并发爬去网络资源， }对爬取到的网络信息进行过滤，模式匹配，文本扫描，以获取对用户有用的信息。

通过脚本可以对爬取的网页进行敏感文件扫描，日志分析。{ 74% : 脚本程序可以利用工具规则库对于用户请求的web日志进行安全性分析，提取出日志中的xss跨站脚本攻击， }{97% : 提取出日志中的sql注入攻击。 }通过脚本发现网站中的安全性漏洞。

1.4创新点

本作品的创新点在于将传统的python爬虫程序与java相结合，{ 63% : 利用java定时任务技术调用python爬虫脚本， }定时更新缓存中的新闻信息。{ 55% : mysql关系性数据库做为持久化存储。 }打破传统新闻系统的

1.5应用前景

当今世界，万维网称为亿万信息的载体，而网络爬虫的作用是不停的爬取网络上的信息。如何利用有用的信息成为了一种挑战。{ 76% : 传统的搜索引擎作为一个辅助人们检索信息的工具成为网民用户涉足万维网的门户。 }面对搜索引擎对于不同用户不同需求的局限性，{ 69% : 定向爬取网络资源的网络爬虫应运而生。 }网络爬虫的健壮和功能强大，是网络搜索引擎功能强大的支撑。

1.6本文的组织结构

本文主要对于目前传统搜索引擎中所存在的功能问题进行了分析，在此基础上出了一套基于python的实时全解决方案，对 其设计进行了详细介绍，着重介绍了该方案中对于系统安全通信问题的解决方 式，之后 出了该方案的一种简单实现，最后分析了该方案的创新性并进行了总结与展望。

全文结构如下:

第一章介绍了本项目的研发背景、研究现状、研究内容等信息

{ 55% : 第二章对整个系统的需求分析进行了详细介绍。 }

第三章介绍了系统系统结构设计，对整个新闻系统的总体规划，技术选型与实现，{ 59% : 技术难点攻破等信息进行了详细介绍。 }

第4章对整个系统的测试与调试进行了分析

第五章介绍了本系统制作总结，前景展望等信息。

2 需求分析

2.1 可行性分析

当今时代网络信息爆发，而传统的新闻机构需要耗费巨大的人力物力。在本系统中，采用当今比较流行的爬虫技术，定时爬取指定新闻门户网站的新闻信息，并及时存库，定时刷新前端页面，达到实时新闻同步的功能，节省了比较大的人力物力，有巨大的市场潜力。在后期的功能优化中，实现pc端，移动端的结合，使新闻服务更加细化，提高用户体验水平。因此本系统拥有极其广阔的前景和使用价值。

2.2 功能概述

本作品的创新点在于将传统的python爬虫程序与java相结合，{ 63% : 利用java定时任务技术调用python爬虫脚本， }定时更新缓存中的新闻信息。Mysql关系性数据库做为持久化存储。用户打开前端页面，定时刷新爬虫爬取的最新的新闻信息，并提供超链接进行门户网站详情页面的跳转，实现实时新闻系统的功能。

2.3 运行环境

本系统的运行环境为centOS7.4服务，搭载mysql5.7关系型数据库及redis4.0.9非关系行数据库缓存，tomcat8，java1.8版本。

2.4 硬件要求

中心服务环境：

CPU为双核2G内存以上

CentOS7.4，64位以上（初步为阿里云centOS服务器）

带宽10M以上（视并发量而定）

云盘40GSSD，后期数据增加，可购买对象存储服务

2.5 运行要求

基于阿里云CentOS7.4服务

tomcat8.0：{ 62%：tomcat服务器是apache基于java语言研发的一个开源web服务项目，是apache基金会jakarta项目中的一个核心项目， }同时支持Servlet4.0和JSP2.0规范：。

Mysql5.7数据库：Mysql数据库是Oracle公司开发的一款开源关系型数据库，使用结构化查询语言进行操作。

{ 61%：Redis数据库：Redis是VMvare主持开发的一款开源非关系型数据库，是一个key-value存储系统， }使用c语言编写，单线程，内存型数据库，支持主从结构，{ 56%：每秒可以实现110000次set操作，81000get操作， }性能极其优异。

2.6 爬虫模块需求概要

爬虫模主要功能是定时爬取新闻门户网站的新闻信息，并进行存储。{ 71%：块使用python语言编写， }实时新闻将更新redis缓存，{ 56%：并进行mysql持久化存储。 }

2.7 新闻管理系统模块需求概要

主要功能是对爬虫模块爬取的新闻进行增删改查操作。使用JakartaEE实现，{ 55%：前端技术使用jquery，html，css， }bootstrap等技术，后端mvc框架使用springmvc，{ 73%：数据库持久层框架使用mybatis。 }后端java使用quarz实现定时任务，{ 57%：定时调用python爬虫模块， }{ 59%：进行数据更新，定时更新redis缓存中的新闻数据， }对用户提供最新最快的新闻信息。

在后端新闻管理系统中，为管理员提供新闻信息的后期修改功能，包括信息的增删改查，隐藏及显示，以及新闻的优先级操作，便于处理新闻信息的显示顺序等功能。同时包含基本的用户管理，管理员操作日志记录，用户登录注册等功能。

2.8 用户界面模块需求概要

用户界面模块主要功能是对爬虫模块爬取的最新新闻进行浏览，也是整个系统展现给用户的部分。使用jquery，html，css，javascript等技术进行实现。对后端的传递回来的新闻进行实时页面更新，并对用户提供新闻查阅。

2.9 用户文档

新闻管理系统：

提供对于新闻信息的增删改查以及优先级修改功能，对于新闻爬虫模块爬取的新闻信息，后台管理员可以进行爬取数据的二次干预操作。

3 系统结构设计

3.1 系统总体规划

3.1.1系统总体规划

{ 56%：本系统实现了一个在线新闻web系统， }基本实现了python定时爬取门户网站的新闻，进行持久化存储，不断更新新闻。整个系统为B/S结构，也就是我们熟知的浏览器/服务器结构。

3.1.2开发语言

{ 69%：java：是一门面向对象的编译型计算机语言。 }{ 65%：拥有面向对象，泛型编程，跨平台等特性。广泛用于企业级应用的开发。 }

{ 76% : python：是一门动态类型的高级编程语言， }是一门以通用型编程语言，同时也是一门解释型语言，运行速度次于java等编译型语言。

3.1.3数据库

Mysql：{ 57% : Oracle维护的开源关系型数据库。 }拥有性能高，成本低，可靠性好的特点。底层为多线程实现，可以支持多处理器，同时支持主从架构。具有极高的安全系统以及及其细化的权限管理机制。

3.1.3 基本框架介绍

SpringMVC：{80% : springmvc是基于springframework的衍生产品， }{ 67% : Spring框架提供了构建javaWeb应用程序的mvc架构， }{ 55% : 更好的与springframework相融合， }从而摒弃了struts以及struts2在配置上的一些缺点，是现在较为主流的mvc框架。

Mybatis：{83% : Mybatis的前身是apache公司的一个开源项目ibatis， }在2010年，{ 69% : 迁移到了Google code，2013年迁移至github并改名为Mybatis。 }{ 55% : 同为持久层框架，相比于hibernate， }{ 66% : mybatis封装更为基础，容易上手， }SQL的操作性也较高，因此，{96% : mybatis可以进行更为细致的SQL优化， }{ 61% : 而hibernate的查询优化策略较为狭隘。 }综上所述，{ 63% : 本系统的持久层框架技术选型为Mybatis。 }

Jquery：{ 56% : jquery是一个简易的优秀的javascript仓库，正如它的宗旨：write less，do more。 }{ 68% : 本身倡导利用最少的代码去做更多的事情。 }{ 73% : 源码全部采用原生javascript编写，使用简单，}响应时间快。在DOM操作，事件处理，动画设计和ajax处理上有极其高的优势。而且引用简单，资源占用少，并且支持各种主流浏览器。深受前后端开发者的喜爱。

3.1.4集成开发环境（IDE）

{ 59% : 本系统的开发工具是IntelliJ IDEA，简称为IDEA。 }是java程序的开发工具。{ 63% : IDEA是业内公认的最好用的java开发工具。 }{100% : 尤其在智能代码助手、代码自动提示、重构、J2EE支持、Ant、JUnit、CVS整合、代码审查、 创新的GUI设计等方面的功能可以说是超常的。 }

3.1.5辅助开发工具

1.maven

{100% : Maven是一个项目管理和整合的工具。 }{100% : Maven为开发者提供了一套完整的构建生命周期框架。 }{100% : 开发团队基本不用花多少时间就能自动完成工程的基础构建配置，因为Maven使用了一个标准的目录结构和一个默认的构建生命周期。 }{100% : 在创建报告、检查、构建和测试自动配置时，Maven可以让开发者的工作变得更简单。 }

2.Git

{81% : Git 是一款免费、开源的分布式版本控制系统，最早由 Linus Torvalds 创建， 用于管理 Linux 内核开发，现已成为最为主流的分布式版本控制工具。 }

3.2 概要设计

3.2.1 系统总体功能模块图

图3.1系统功能模块图

3.2.2 系统的业务流程

3.2系统结构概述

基于python网络爬虫的实时新闻系统主要包括三部分：

python网络爬虫模块，其中python网络爬虫模块又由爬虫调度端、网页下载器、网页解析器、URL管理器等模块组成。

java后端管理系统模块，web展示模块。Java后端管理系统使用springmvc，mysql，mybatis，git版本控制等技术。

{ 61% : 用户前端展示模块使用html，css，javascript，jquery，bootstrap等技术。 }

3.2.1 python爬虫模块实现

{ 56% : 本系统使用模块化的设计进行开发， }python爬虫模块实现了物联网信息的抓取存库等功能。一个完整的爬虫架构通常含有网页解析器、URL管理器、网页管理器、爬虫调度器等模块组成。

爬虫调度器使用java后端实现，java采用quartz定时任务框架，使用Jython工具进行python爬虫脚本的分时调用，爬取最新的新闻数据。

{80% : URL管理器的作用是对已经获得的URL和将要爬取的URL进行管理， }在本系统中，url管理器的功能是对新闻页面中的各条新闻践行数据的抓取，url的分析与保存。

网页下载器的作用是对URL管理器中的URL进行识别，将其制定的URL网页代码进行下载并存储，并将其未被下载过的送入URL管理器中，类似于一个深度遍历，读取一条新闻之后，进入新闻页面，进行链接分析，已经下载的页面不需要继续下载，没有下载的URL送回URL管理器进行处理。

```
# 爬虫核心模块
def Spider ( url ) :
i = 0
print “downloading “, url
myPage = requests.get ( url ) .content.decode ( “gbk “ )
# myPage = urllib2.urlopen ( url ) .read ( ) .decode ( “gbk “ )
myPageResults = Page_Info ( myPage )
save_path = u “网易新闻抓取 “
filename = str ( i ) + “_ “+u “新闻排行榜 “
StringListSave ( save_path, filename, myPageResults )
i += 1
for item, url in myPageResults:
print “downloading “, url
new_page = requests.get ( url ) .content.decode ( “gbk “ )
# new_page = urllib2.urlopen ( url ) .read ( ) .decode ( “gbk “ )
newPageResults = New_Page_Info ( new_page )
filename = str ( i ) + “_ “+item
StringListSave ( save_path, filename, newPageResults )
i += 1
```

网页解析器使用lxml进行对网页下载器下载的网页代码进行解析，抽离出自己有用的数据，在网页下载器下载好代码之后，对整个html文档进行处理，过滤出有用的信息。

```
# 网页解析器模块
{96% : def Page_Info ( myPage ) : }
‘ ‘ ‘Regex ‘ ‘ ‘
mypage_Info = re.findall ( r ‘^div class= “titleBar “ id= “.*? “^^h2^ ( .*? ) ^/h2^^div
class= “more “^^a href= “ ( .*? ) “^.*?^/a^^/div^^/div^ ‘, myPage, re.S )
return mypage_Info
{97% : def New_Page_Info ( new_page ) : }
dom = etree.HTML ( new_page )
new_items = dom.xpath ( ‘//tr/td/a/text ( ) ‘ )
new_urls = dom.xpath ( ‘//tr/td/a/@href ‘ )
assert ( len ( new_items ) == len ( new_urls ) )
return zip ( new_items, new_urls )
```

网页管理器的作用主要是对处理好的数据进行存储。本系统中使用redis进行缓存最新新闻信息，{ 55% : 使用mysql进行持久化存储。 }

```
# 爬取的文件，分类型存入Mysql持久化存储
{98% : def StringListSave ( save_path, filename, slist ) : }
if not os.path.exists ( save_path ) :
os.makedirs ( save_path )
path = save_path+ "/ "+filename+ ".txt "
{94% : with open ( path, "w+ " ) as fp: }
for s in slist:
{89% : fp.write ( "%s\t\t%s\n " % ( s[0]. }encode ( "utf8 " ) , s[1].encode ( "utf8 " ) ) )
```

爬虫脚本实现详情见附录。

3.2.2 web后端模块概述

后端管理系统主要是通过JakartaEE技术实现，mvc框架使用springmvc，{ 70% : 持久层框架使用mybatis实现， }java定时任务组件使用quartz框架实现。{ 71% : 关系型数据库使用mysql， }java后端调用pyhton爬虫组件使用Jython实现。

系统采用quatz定时任务，定时调用python爬虫脚本进行分时爬取门户网站的新闻信息，存入数据库，按照热度进行新闻的优先级处理，并更新redis缓存信息。{ 64% : 通过java实现一个后台新闻管理系统， }实现新闻动态管理、用户管理、新闻模块管理等功能。在爬虫模块爬取了新闻信息之后，对新闻进行分类分优先级存储的时候，后端管理员可以对程序整理好的数据进行二次修改，例如增删改查和基于优先级修改重排序等。

3.2.2.1 登录模块

新闻发布管理平台，登录模块，采用简洁明了的呈现形式编写登录页面，用户密码通过MD5加密验证，后端用户名密码非空验证，图像验证码验证等逻辑实现。{ 67% : 整个请求通过jquery ajax异步发起， }{ 57% : 用户登录信息全部填写正确，则将用户登录信息存入session， }后重定向进入后端管理界面，session进行权限控制。若登录不成功，系统将重定向到登录页面，并添加提示信息。

3.2.2.2 管理员核心模块

管理员核心管理模块主要分为新闻管理和用户管理两个模块。Title栏目中，{ 59% : 分为logo，新闻管理，新闻发布，用户管理， }设置，用户操作等模块

1.新闻管理

首先是新闻发布模块，点击进入之后，是新闻分条件查看。分为单审核新闻，已审核新闻，全部新闻。同时还包含根据新闻id和关键字进行新闻检索等功能。新闻列表为分页展现。

其次是新闻管理模块，新闻管理模块第一栏是新闻条件检索模块，分别是新闻类别查找，新闻ID查找，新闻关键字查找。并可以随意组合。{ 57% : 管理员可以对新闻进行新增和删除的批量操作， }对于每条新闻可以进行编辑和删除操作。在所有新闻详细信息的处理中，{ 69% : 都使用ueditor富文本编辑器进行实现， }确保文件显示格式与新闻上传格式相同，提高了用户体验。

新闻发布模块中，{ 56% : 前端页面传入新闻标题，新闻类别，新闻内容等内容， }后端接收之后，进行非空验证后存入数据库。存储新闻信息的同时，按照登陆存储的session信息进行上传用户信息的填充，更新时间的添加等，存入数据库。

新闻上传代码如下：

2.用户管理

首先是账号管理，在账号管理中，角色分为编辑人员和测试人员。{ 62% : 账号状态分为正常和停用两种状态。 }账号分为手机号，邮箱，姓名三种表现形式。同时可以通过关键词查找用户帐号信息。管理员可以新增数据和批量删除用户帐号数据，并可以对用户数据进行更新。角色管理中，管理员可以编辑和删除用户角色数据。进行权限划分，编辑人员与测试人员之间的角色切换。栏目管理中，{ 61% : 管理员可以对新闻的栏目类别进行管理， }设置不同的优先级和栏目名称，便于新闻的编辑上传等操作。

角色添加代码如下：

3.设置模块

主要涵盖菜单管理模块，设置新闻菜单的树状结构。对于单个新闻模块的操作，包括菜单名称，上级菜单，

URL，会否显示，备注等信息编辑修改。检索功能包含新闻菜单的类别关键词等信息进行查找。新闻模块可以单个编辑，新增及批量删除等。

查询新闻列表代码如下：

4.账号管理

账号管理中，可以对用户密码数据进行修改操作。{81%：用户输入一边原密码，两遍新密码。 }后端模块接收到用户提交的表单数据，进行密码验证。成功则更新用户数据，失败则返回失败信息。在用户前端表单提交之初也会对于用户提交的数据进行初步校验，非空验证以及两次密码输入是否一致等问题。

3.2.3用户前端展示模块概述

前端展示页面为用户停工新闻查看，分类型查看以及检索功能。每条新闻提供跳转到新闻详情页面的超链接，新闻按照发布时间以及访问热度进行排序查看。新闻简略信息查看会展示新闻的抓取时间，{ 56%：新闻标题名称，新闻的来源等信息。 }

3.2.4 系统日志与异常管理体系

{ 68%：系统日志体系使用log4j实现，log4j也是apache公司的一个开源项目。 }{ 62%：也是一个基于java 的日志记录工具。 }通过使用log4j，我们可以痛殴配置决定日志的输出方式以及输出位置。同时可以定义日志的级别，从而有利于在不同的环境（生产或者测试环境）中，控制不同的日志输出量，节省服务资源。{ 59%：整个系统使用同一个日志体系，有利于系统的规划。 }

系统异常处理体系，整个系统在编写初期极力避免异常出现。在可能出现并发等情况的地方使用try...cache语句，捕获异常并打印到log中。定义log级别为error或者warn，并及时反馈给开发者或者运维人员。

3.2.5 系统难点与解决方案

制定了需求分析之后，{ 61%：在开发过程中，不免一些技术过程上的问题。 }

{ 55%：在编写完python爬虫脚本之后， }关于爬虫调度器方面，一时不知如何下手。Python本身也含有sched等定时任务实现，但是python代码完全与java后端模块之后，会产生新闻无法无缝同步的问题，当时也进行了redis存储分布式锁解决这一问题，效果并不是很理想。综合多方面考虑，最后的爬虫调度器的技术选型为java后端调用python爬虫脚本架，使用jython框架。Jython是一个python语言在java程序中的一个实现。{80%：但Jython不像Python或其他任何高级语言，它提供了对其实现语言的一切存取。 }{91%：所以Jython不仅为我们提供了Python的库，同时也提供了所有的Java类。 }{ 67%：这使Jython有一个巨大的资源库。 }同时使用java的quartz定时任务框架进行分时任务调度，爬虫获取的信息以http请求的方式，请求java后端系统的接口，代码复用性也有所提高。

3.3 数据库设计

3.3.1 数据库表的设计

数据库使用mysql，具体数据库表设计如下

表3.1 新闻菜单表

字段 数据类型 注释 类型

4 系统测试与调试

{ 68%：软件系统测试是软件开发过程中保证软件质量的至关重要的部分， }在软件开发过程周期中起着极为重要的作用。软件系统测试的目的是发现系统中存在的设计缺陷与程序BUG。{ 55%：在软件测试过程中，设计全面正确的测试用例， }实现路径覆盖与语句覆盖的模块化测试，才能尽最大可能的实现软件系统的正确运行。

4.1 系统测试与调试的目的

软件系统测试的目的就是发现系统中的问题并进行修正和改进从而提高软件系统产品的质量。测试和调试中，不仅是要在软件系统中找到bug并修正，在遇到问题并解决问题的过程中，使我们积累经验，在后期的软件开发工作中，极力避开这些会出现的问题，做出更优秀的产品。

4.2 系统软件测试的重要性

本实时新闻系统的开发重心是后端java新闻管理系统，同时也是系统软件测试的重心。Java后端代码逻辑复杂，代码实现上却非常类似，对开发人员的逻辑思考能力有着一定的要求。同时要综合考虑资源分配与抗压能力等等因素。因此，对于java后端代码的测试尤为重要。后端是整个新闻系统的核心，保证后端服务的高可用，稳定持续运行，{ 60%：才能保证整个系统的正常运行。 }

4.3 测试用例示例

登录模块测试用例设计

测试用例ID 场景 测试步骤 预期结果 备注

TC1 初始页面显示 从用例入口处进入 页面元素完整，显示与详细设计一致

TC2 用户名录入 - 验证 输入已存在的用户：admin 输入成功

{97% : TC3 用户名 - 容错性验证 输入：aaaaabbbbbccccccddddddeeeee 输入到蓝色显示的字符时， }系统拒绝输入 输入字段超限

TC4 密码 - 密码录入 输入与用户名相关联的数据：test 输入成功

TC5 系统登录 - 成功 TC2，TC4，单击登录按钮 登录系统成功

{85% : TC6 系统登录 - 用户名、密码校验 没有输入用户名、密码，单击登录按钮 系统登录失败，并提示：请检查用户名和密码的输入是否正确 }

{ 70% : TC7 系统登录 - 密码校验 输入用户名，没有输入密码，单击登录按钮 系统登录失败，并提示：需要输入密码 }

{84% : TC8 系统登录 - 密码有效性校验 输入用户名，输入密码与用户名不一致，单击登录按钮 系统登录失败，并提示：错误的密码 }

{88% : TC9 系统登录 - 输入有效性校验 输入不存在的用户名、密码，单击登录按钮 系统登录失败， }并提示：用户名不存在

TC10 系统登录—安全校验 连续3次未成功 系统提示：您没有使用该系统的权限，请与管理员联系！

5 总结与展望

在这个新闻系统的设计开发过程中，{ 58% : 我遇到了一些困难，也学到了很多。 }在阿里云centos上面搭建了一个完整的测试开发环境，页面设计与ps，后端各种工具的使用，把实习工作中学到的东西尽可能多的运用的这个项目的制作中。一个优秀的产品，不光要有贴近于用户的全面的需求分析，还要有前后端全面严谨的技术实现。这个项目在学习中，加深了对python的认识，学到了很多。在以后的工作生活中，我也会直面困难，奋力前行。

参考文献

[1] 杨天翔.从“房屋银行”现象看房地产中介的创新[J].中国房地产，2002（01）：20-22

（小四号宋体， 1.25行距，英文为新罗马，小四）

致谢

实时新闻系统

Python爬虫模块

Java

cms管理系统

用户管理模块

功能模块管理

常用工具

系统管理

新闻模块管理

45

- 45-