大雅相似度分析

论文标题： 宋少忠
检测日期: 2018年04月23日
学号: 毕业论文终稿查重版陈巍瑜
正文字符数： 28892
正文字数： 21367
检测范围： 大雅全文库

一、总体结论

文献相似度15.4%	重复字符数4448	最密集相似处	密集相似处	非密集相似处	前部相似度	中部相似度	尾部相似度
		3	17	7	12	7	8

二、相似片段分布

三、典型相似文献

相似图书

作者	题名	出处	相似度
李志义	Web信息检索系统的设计及应用	北京：清华大学出版社，2011.10	2.29%
赵伟	Azure WebSites权威指南 微软云计算Web平台开发实战详解	北京：清华大学出版社，2015.06	1.14%
白英彩;薛质;王豪行;王思伟;陈骁;石燕华;董静翔;钱向阳	英汉计算机通信辞典	上海：上海交通大学出版社，2016.04	0.79%
梁海宏	连接时代 未来网络化商业模式解密	北京：清华大学出版社，2014.03	0.76%
于富强	ASP.NET 4.0 Web网站开发实用教程	北京：北京大学出版社，2012.10	0.67%
丰艳;姜媛媛;李海涛	iLike职场 大学生就业指导 .NET方向	北京：电子工业出版社，2012.04	0.66%
陈营辉;赵海波	PHP+Ajax完全自学手册	北京：机械工业出版社，2009.01	0.63%
蒋罗生	电子商务网站设计与维护	北京：中国电力出版社，2009.09	0.61%
孙光宇;张玲玲	Android物联网开发从入门到实战	北京：清华大学出版社，2015.07	0.61%
达特森	响应式设计、改造与优化	北京：电子工业出版社，2015.09	0.54%
查尔斯·亚瑟（CharlesArthur）;余淼	数字战争 苹果、谷歌与微软的商业较量	北京：中信出版社，2013.07	0.51%
孙祥飞	新闻传播学热点专题80讲 2015	北京：人民日报出版社，2014.10	0.5%
老男孩	跟老男孩学Linux运维 Web集群实战	北京：机械工业出版社，2016.04	0.5%
闫小坤;周涛	微信公众平台开发基础与实战	北京：机械工业出版社，2014.09	0.49%
闫小坤;周涛	微信公众平台应用开发从入门到精通	北京：清华大学出版社，2015	0.49%
周辉	互联网信息监测系统研究	北京：知识产权出版社，2015.09	0.47%
穆虹	市场管理法	北京：中国政法大学出版社，2015.09	0.46%
咎辉	SEO实战密码 60天网站流量提高20倍	北京：电子工业出版社，2011.03	0.45%
咎辉	SEO实战密码 60天网站流量提高20倍 第2版	北京：电子工业出版社，2012.05	0.45%
袁毅	电子商务概论	北京：机械工业出版社，2013.05	0.45%
程虹	网络营销	北京：北京大学出版社，2013.03	0.45%
商玮	电子商务网站设计与建设	北京：人民邮电出版社，2011.07	0.45%
孙庆磊;曲秋晨	淘金微营销 微时代，你必须知道的6个成功法则	北京：北京联合出版公司，2014.03	0.45%

佟力强	首都互联网发展报告 2013	北京：人民出版社，2013.11	0.44%
宋德富;司爱侠	计算机专业英语教程 第4版	北京：高等教育出版社，2013.09	0.44%
黄慧芳;王琳	PHP+MySQL项目开发权威指南	北京：中国铁道出版社，2013.08	0.43%
罗森林	信息安全与对抗实践基础	北京：电子工业出版社，2015.04	0.42%
传智播客高教产品研发部	PHP网站开发实例教程	北京：人民邮电出版社，2015.09	0.42%
程光;杨望	网络安全实验教程	北京：北京交通大学出版社，2013.01	0.42%
梁春燕	Internet主题搜索引擎设计与研究	北京：中国水利水电出版社，2012.03	0.42%
刘兵;俞勇	世界著名计算机教材精选 Web数据挖掘 第2版	北京：清华大学出版社，2013.01	0.42%
齐伟	跟老齐学Python 从入门到精通	北京：电子工业出版社，2016.03	0.41%
修思禹	什么时候出发都不晚 创业路上的老男孩	北京：中国商业出版社，2014.02	0.35%
冯晓霞	大学计算机文化基础 Windows 98、Office 2000版	杭州：浙江科学技术出版社，2002.08	0.23%
李俊民;许波	Visual Basic轻松入门	北京：人民邮电出版社，2009.04	0.22%
王长青	Android智能穿戴设备开发指南	北京：人民邮电出版社，2015.05	0.22%
	中外电器 1997年合订本	《中外电器》杂志社，1998.03	0.21%
传智播客高教产品研发部	Java Web程序开发入门	北京：清华大学出版社，2015.02	0.2%
冯相忠	计算机文化基础	北京：中国铁道出版社，2004.09	0.2%
高禹;冯相忠	大学计算机基础	北京：中国铁道出版社，2005.08	0.2%
Microsoft Corporation;北京博彦科技发展有限责任公司	全面掌握电子商务开发技术 Business to Consumer	北京：清华大学出版社，2000.10	0.2%
张健	Web应用系统开发 PHP	北京：中国铁道出版社，2011.08	0.2%
顾沈明	计算机基础 第3版	北京：清华大学出版社，2014.07	0.2%
詹德隆	商业智能与云计算	北京：人民邮电出版社，2015.03	0.2%
任宏萍	面向对象程序设计	武汉：华中科技大学出版社，2010.10	0.19%
佩兰	时尚家庭电脑应用	上海：上海科学技术出版社，2003.11	0.19%
陶再平;吕侃徽	局域网组建与管理	北京：高等教育出版社，2008.08	0.19%
胡伏湘;龙超;党伟华	计算机网络技术教程 基础理论与实践	北京：清华大学出版社，2015.08	0.19%
蔡希尧	信息系统的发展与创新	西安：西安电子科技大学出版社，2009.12	0.19%
叶乃文;王丹	Java语言程序设计教程	北京：机械工业出版社，2010.01	0.19%

相似报纸

作者	题名	出处	相似度
	360诉百度索赔4个亿	法制晚报，2013.10.17	0.65%
	360诉百度	西安晚报，2013.10.18	0.57%
	什么是robots协议	新快报，2012.11.02	0.55%
	百度胜360标的1亿判赔70万	青年时报，2014.08.08	0.55%
徐晓风	搜索“手持身份证”出来一堆真人照？别慌，大多是示例照片，但你要注意自己的信息安全	金乡新闻，2016.09.26	0.51%
陈庆麟	“3B大战”一审宣判	新快报，2014.08.08	0.51%
	大众点评诉百度为争20万亿市场	长江商报，2016.04.18	0.51%
叶丹	奇虎赔70万可抓取百度内容	南方日报，2014.08.08	0.51%
徐然	大众点评网诉百度不正当竞争 索赔9000万	21世纪经济报道，2016.04.13	0.51%

宋宁华;王治国	大众点评起诉百度索赔9000万	新民晚报美国版 , 2016.05.27	0.51%
	搜索“手持身份证照”出来一堆真人照？是真的！	城市商报 , 2016.09.08	0.51%
赵霞	Robots协议之争：互联网豪门恩怨谁惹的祸	中华工商时报 , 2014.09.12	0.51%
孙思娅	百度诉360索赔亿元案开庭	京华时报 , 2013.10.17	0.51%
	搜索“手持身份证”出来一堆真人照？	承德晚报 , 2016.09.08	0.51%
	风头正劲 遭遇侵权之诉	南岛晚报 , 2014.07.17	0.51%
徐晓风	搜索“手持身份证”出来一堆真人照？	扬子晚报 , 2016.09.06	0.51%
	百度诉360一审获赔70万元	京华时报 , 2014.08.08	0.51%
	3B大战起诉与反起诉：1亿元 4亿元	重庆商报 , 2013.10.17	0.51%
	360被判赔70万元	武汉晚报 , 2014.08.08	0.51%
	搜索“手持身份证”出来一堆真人照？	湛江晚报 , 2016.09.06	0.51%
	3B搜索大战引发Robots协议之争	法制晚报 , 2013.10.25	0.51%
	“3B大战”硝烟暂灭	济南日报 , 2012.11.13	0.47%
辛华	“3B搜索大战”和解收场	中国消费者报 , 2012.11.07	0.47%
	中国12家搜索企业昨日签署自律公约	长江商报 , 2012.11.02	0.46%
	惊心！我们的信息正在“裸奔”	兰江导报 , 2016.09.14	0.46%
卢燕;王治国	“百度美食”复制他站介绍和评论	青年报 , 2016.05.27	0.46%
	惊心！我们的信息正在“裸奔”	宿迁晚报 , 2016.09.14	0.46%
舒锐	还互联网行业以规则世界	今日云和 , 2014.08.11	0.46%
舒锐	机器人协议案，一份多赢的判决	人民法院报 , 2014.08.12	0.46%
	机器人协议案一份多赢的判决	现代金报 , 2014.08.11	0.46%
黄;鑫	互联网商业环境需要多方维护	经济日报 , 2013.10.16	0.46%
	方兴东：滥用Robots协议将致互联网大乱	劳动报 , 2012.09.11	0.46%
薛松	国家版权局指360侵权	广州日报 , 2012.12.31	0.46%
千量	百度垄断or360不正当竞争？	人民邮电报 , 2012.11.30	0.46%
舒锐	机器人协议案，一份多赢的判决	人民政协报 , 2014.08.12	0.46%
	工信部牵头协调“3百大战”和解收场	成都商报 , 2012.11.02	0.46%
舒锐	还互联网行业以规则世界	人民邮电报 , 2014.08.25	0.46%
徐晓风	搜索“手持身份证”出来一堆真人照	宜宾晚报 , 2016.09.07	0.46%
	百度获赔70万元	贵州商报 , 2014.08.08	0.46%
	我们的信息正在“裸奔”	柳州晚报 , 2016.09.14	0.46%
	惊心！我们的信息正在“裸奔”	北京晚报 , 2016.09.13	0.46%
	法院判360赔偿百度70万元	南京晨报 , 2014.08.11	0.43%
	百度被判不正当竞争赔323万	解放日报 , 2016.05.27	0.43%
	“3Q”诉讼大战上演终极篇	北京日报 , 2013.11.27	0.43%
	微微一搜很惊人	羊城地铁报 , 2016.09.06	0.42%
	360被指违反国际通行规范 百度索赔1亿元	中国贸易报 , 2013.10.24	0.41%
	大众点评诉百度等两公司不正当竞争案一审宣判	人民法院报 , 2016.05.27	0.4%
孙政华	360违反Robots协议法律无可奈何	法治周末 , 2012.09.13	0.4%
	奇虎360搅局搜索市场	中国科学报 , 2012.09.13	0.4%
	与Robots协议有关的著作权问题	中国知识产权报 , 2013.10.25	0.4%

相似期刊

作者	题名	出处	相似度
李志义	网络爬虫的优化策略探略	现代情报 , 2011 , 第10期	1.83%

忻建;范建中	一种虚拟执行蜘蛛的设计与实现	电脑与电信 , 2009 , 第6期	1.31%
黄宏博;冯温迪;王思远	一种基于局域网的分布式搜索引擎设计与实现	软件导刊 , 2015 , 第3期	0.67%
陈涛;叶荣华	基于Spring Boot和MongoDB的数据持久化框架研究	电脑与电信 , 2016 , 第C1期	0.65%
陈骏;谭庆平;谭雄	ASP.NET AJAX在博客网站中的应用	微计算机信息 , 2008 , 第3期	0.64%
	二、定向链接的基本含义及技术原理	科技与法律 , 2016 , 第2期	0.55%
张玉亮;哈斯	蒙古文网络文本识别与采集方法	内蒙古师范大学学报(哲学社会科学汉文版) , 2016 , 第4期	0.51%
马晓明	视频聚合平台的直接侵权认定探究	电子知识产权 , 2016 , 第4期	0.51%
冯寅杰	张一鸣最长的一个月	环球市场信息导报 , 2014 , 第30期	0.5%
张玲玲	定向链接网络服务提供者侵犯著作权责任问题研究	科技与法律 , 2016 , 第2期	0.5%
	专利	电子知识产权 , 2014 , 第1期	0.46%
蔡舜	美国网页存档调查及启示	图书馆理论与实践 , 2016 , 第2期	0.44%
李雨石	大众点评诉百度“搭便车”	光彩 , 2016 , 第7期	0.4%
杨丽彬;李海林;张飞波	大数据环境下的管理信息系统发展研究	大数据 , 2016 , 第1期	0.2%
马凯航;高永明;吴止锾;李磊	大数据时代数据管理技术研究综述	软件 , 2015 , 第10期	0.2%
张光墨	基于JavaServer~(TM) Faces和DAO模式的大型设备采购系统	计算机与信息技术 , 2007 , 第6期	0.19%
丛佩丽	试析以Windows Server 2012为平台架设局域网服务器	无线互联科技 , 2016 , 第15期	0.19%
刘贤	详解 Web 服务基础	计算机与网络 , 2017 , 第11期	0.15%
银强	对计算机嵌入式实时操作系统的研究及分析	价值工程 , 2010 , 第36期	0.15%
宁永军	新形势下思想政治工作面临的挑战及对策研究	企业文化 , 2016 , 第32期	0.15%
陈耀东;吴阳波	基于嵌入式Web服务器的远程控制系统	微计算机信息杂志 , 2007 , 第32期	0.15%
李淑芝	Web网页制作探讨	计算机与现代化 , 1998 , 第6期	0.15%
陈运贵	地方高校服务农村文化建设的路径选择	社科纵横 , 2012 , 第7期	0.1%
陈志兵;贝来债	分布式操作系统（II）	抗恶劣环境计算机 , 1990 , 第3期	0.1%
张骏;史振华;白丽晗	基于.NET的Web结构挖掘技术研究及应用	电脑编程技巧与维护 , 2009 , 第4期	0.1%
Alberto Faro;金传升	用于计算机成网的以多微机为基础的结构	通信技术 , 1986 , 第1期	0.1%
路辉;李筠;孔令军;刘伟	连云港市开展“农超”对接的对策与建议	上海蔬菜 , 2016 , 第1期	0.08%
陶正娟	关于高职院校考风建设的几点思考	林区教学 , 2016 , 第9期	0.05%
徐一化	党报报业集团如何进一步提高舆论引导能力	新闻窗 , 2009 , 第2期	0.05%
	常州市国营企业试行厂长负责制的暂行规定	经济管理 , 1984 , 第12期	0.05%
刘会丽	双黄连的药理作用及临床应用研究	医药前沿 , 2016 , 第6期	0.05%
朱炜婧	浅谈创意面料在现代家居空间中的应用	大众文艺 , 2016 , 第11期	0.05%
于帅;冉佳	章沈强诠释星光农机成功基因	农业机械 , 2014 , 第11期	0.05%
董风莉	建筑工程造价拦标价的探析	城市建设 , 2011 , 第11期	0.05%
王堃	高校资产经营公司内部审计工作中存在的问题与对策	齐鲁珠坛 , 2011 , 第1期	0.05%
金宁;黄奕裔	南航高速风洞的回顾与展望	中国空气动力学回顾与发展 , 1996 , 第1期	0.05%
彭志宇	石油天然气资金筹措与发展战略的探索	石油经济 , 1999 , 第2期	0.05%
周炜	西藏语言政策的变迁	西北民族研究 , 2002 , 第3期	0.04%
顾建跃	刍议铁道高职院校校园文化的培育与建构	科技致富向导 第22期	0.04%
黄庭标;冯尚克	小儿病毒性脑炎60例临床分析	医学文选 , 2000 , 第4期	0.04%
夏立平;许嘉	美国对香港回归中国的政策及其对中美关系的影响	世界经济与政治 , 1997 , 第4期	0.04%

钟波;肖智;李勇;张志恒	一种基于遗传算法的数据预处理组合方法	西南师范大学学报(自然科学版) , 2002 , 第4期	0.04%
许林	厦门市服务外包业发展现状和对策研究	发展研究 , 2008 , 第8期	0.04%
涂建军;张明举;何劲耘	宜宾市经济发展综合优势分析及战略选择探讨	西南师范大学学报(自然科学版) , 2002 , 第4期	0.04%
张兴海;樊秋菊	新形势下加强知识分子思想政治工作的几个着眼点	复印报刊资料(思想政治教育) , 1993 , 第10期	0.04%
喜蕾	元代高丽贡女与蒙古族以外的其他民族通婚状况考述	西北民族研究 , 2002 , 第3期	0.04%
袁天凤;邱道持	四川省区域经济发展水平差异性分析	西南师范大学学报(自然科学版) , 2002 , 第4期	0.04%
吴宝山;秦良才;付鸿雁	论中国邮政服务“三农”的对策	邮政研究 , 2005 , 第4期	0.04%
莫安达	试论政府在外资中小企业增长方式转变中的作用	中国人力资源开发 , 1997 , 第7期	0.04%
莫安达	试论政府在餐资中小企业增长方式转变中的作用	东莞理工学院学报 , 1997 , 第1期	0.04%

其他网络文档

作者	题名	相似度
王鹏	AJAX技术在WEB应用中的研究与实现	0.76%
蔡晓庆	Ajax技术在企业电能管理系统中的应用研究	0.75%
彭轲	基于浏览器服务的网络爬虫的设计与实现	0.73%
刘哲	基于AJAX的富士通EBISS销售管理系统	0.71%
丁娜	基于AJAX的WEB2.0技术研究	0.71%
李蕾	农业专业搜索引擎个性化服务研究与实现	0.71%
任鸿	基于异构网络的知识挖掘与服务关键技术研究	0.7%
	MongoDB实战	0.67%
孙鹤	基于ASP.NET的企业综合信息系统的研究与实现	0.64%
罗兵	支持AJAX的互联网搜索引擎爬虫设计与实现	0.6%
	How UniqueIsYour Web Browser?	0.59%
袁宇丽	基于HTML网页的Web信息提取研究	0.57%

四、典型相似内容对比

1

当前位置: 0% **本段(页)重复比例:** 13.42%

*第1章绪论1.1研究现状:随着大数据时代的到来,互联网对人类生*活的影响越来越大,已成为人类获取信息的主要来源之一。互联网为用户带来大量数据的同时,也带来了问题。如何及时获得有效信息已成为研究的重点。搜索引擎根据预定的策略发现并从互联网获取数据,并将其存储在本地;它执行诸如去噪,提取和生成数据索引等处理,并最终为用户提供信息检索服务并向用户的系统显示相关信息。爬虫(Crawler)是搜索引擎架构中最低级别的模块。它使用某种策略从互联网上获取数据,预处理数据,然后将处理后的数据提交给其他搜索引擎模块。数据的质量和数量直接影响用户的体验。然而,随着互联*网数据在大数据时代的爆*发式增长,爬虫的抓取数据已经不能满足实际应用的需求。

2

当前位置: 1.3% **本段(页)重复比例:** 4.75%

主要从硬件和软件方面考虑解决这个问题:首先,升级爬虫硬件,使用更好的硬件设备,但成本不高,且不易扩展;其次,使用分布式方法来提高爬虫的并行处理能力,但这种方法会增加爬虫系统设计的复杂性。目前,大多数大型爬虫系统采用分布式方式,但仍不*能满足用户的实际需求。其次,爬*虫系统还需要解决网页动态变化导致本地副本过期的问题。网页可能随时更改。在某些情况下,爬行动物系统必须及时发现和更新本地网页。然而,互联网海洋中的网页数量庞大且分布广泛,爬行系统可能需要几周甚至更长的时间才能更新。本地图书馆的网页副本不太新。因此,采集速度快,网页更新及时的高可靠性爬虫系统不仅为搜索引擎提供基础数据,而且为数据分析和挖掘提供基础数据,从而获取信息和知识。

3

当前位置: 2.6% **本段(页)重复比例:** 37.66%

*1.2研究的目的与意义本课题*的研究目的在于使用基于Python的开源技术,结合其他网络编程知识,实现具有强大开发能力的定制网络爬虫原型。1.*基于Soapy-Redis框架,自定义爬网规则。2.*选择适当的数据库进行数据存储。3.使用多线程并发*结构来提高操作效率。4.因为*是本科阶段,*无论是缺乏学习时间还*是难度比较大。因此,本课题主要研究高效鲁*棒爬行系统,并将爬行数据用于非常简单的数据分析。*这项研究的意义:1.深入学习Pyth*on和Soapy-Red*is开源框架,自己动手实现网络爬升的良好扩展性的爬虫原型将对我们在学习新技术*和拓宽视野方面发挥积极作用。2.*尽管只实施了一个原型程序,但探索简单易行的想法是非常正确的。

4

当前位置: 3.9% **本段(页)重复比例:** 31.03%

针对不同用户的特定需求提供特定功能正是当今程序*开发领域的热门话题。简单和*容易改变爬行动物将是一个很好的做法。1.3 论文结构与限定:本*论文分八章节进行阐述:第一章,绪论。首先介绍本文的研究背景网络*爬虫的产生与发展以及在数据分析等邻域中数据获取的手段,*接着是本课题的研究目*的和意义。*最后给出了论文的研究内容及论文结构安排。第二章,分别*介绍了网络爬虫系统的基本知识和相关技术。在这一章介绍爬虫的分类及作用,随后介绍网络爬虫涉及到的网络协议,最后介绍爬虫搜索策略。第三章,*爬虫总体架构设计、使用的框架*。本章主要是爬虫系统*的需求与架构设计进行了*详述的设计与实现。首先介绍爬虫系统的业务需求,性能需求和其他需求,然后完成爬虫系统使用的Scrapy框架,再来是分别介绍两个NOSQL数据库,一*个MongoDB是放在*磁盘层作为获取的数据的持久化操作,*另一个Redis内存*数据库放在内存中一方面进行查重,另一方面进行爬虫的调度。

5

当前位置: 9.09% **本段(页)重复比例:** 42.41%

图2.1比特币市场股票图只不过要达到这种水平,那么需要爬取到特别优质的数据集.而比特币市场比起股票来说是一个很简单的市场,但是这也能证明爬虫在该方面的运第三类为各种论文以及文*章提供有力的数据支撑。比如,我们*要找出全国气温最低的地方并直观的表示出来,那么我们仅仅需要爬取一下全国的气象信息,然后通过类似于D3.js这*种可视化JavaScript库就可*以得到一张柱状图.如2.2图所示,这样,全国最低气温就*能直观的显示出来。同时也可以用过各种气象论文中论点的有力支撑.图2.2全*国最低*气温排名2.2HTTP协议2.2.1 Http协议定义HTTP,超文本传输协议。HTTP是基于“ request和respond ”模*式的无状态应用层协议。

6

当前位置: 11.69% **本段(页)重复比例:** 65.25%

*Cookie是一种*可以弥补无状态HTTP缺乏的机制。会议开始之前,基本上所有的网站都*使用cookie来跟踪会话。Cookie不能跨域使用。*会话是服务器记录客户端状态的机制。使用比cookies更简单,服务器存储压力也*相应增加。Cookie数据存储在客户端的浏览器中,session数据放在服务器上;Cookie不是很安全,别人可以分析CO*OKIE存放在本地和COOKIE的欺骗行为,*考虑到安全性应该使用session;如果session将在一段时间内大量的保存在服务器上。当访问量增加时,它将大量占用服务器的CPU资源。考虑提高服务器性能,应使用cookie;客户端的总cookie大小也*是有限的(基本4k)。

7

当前位置: 12.99% **本段(页)重复比例:** 62.53%

*Firefox和Safari允许cookie高*达4097字节,包括名称,值和等号。*每个域的cookie都是有限的,每个域名的Firefox限制为50个cookie。2.3Rebots协议2.3.1网络爬虫引发的问题:网络爬*虫法律风险:*Web服务器上的数据所有权*,归网站所属者所有,网络爬虫获取数据可能带来法律风险隐私Web爬虫的泄漏:Web爬虫可能突破访问控制并获得受保护的数据以*揭示用户个人隐私,并且造*成大规模隐私数据的泄露。Web服务器默认接收人员访问。受写作水平和目的限制。Web爬虫将为Web服务器带来巨大的资源浪费,造成服务器的负担。

8

当前位置: 14.29% **本段(页)重复比例:** 46.27%

作用:网站管理员制定规则告知网络爬虫哪些页面可以抓取,哪些不行。*形式:在网站根目录下的robots.txt文件./robots.txt。/网络爬虫:自动或人工识别robots.txt中的规定,再进行内容爬取。约束性:Robots协议仅*仅是建议但非约束性,网络爬虫可*以不遵守robots,但是存在一定的法律风险。1.)为什么需要Robots协议对*爬虫来说网站非常被动,只有老老实实被抓取的份。所以,对于网站的管理者来说,就存在这样的需求:某些路径下是个人隐私或者网站管理使用,*不想被搜索引擎抓取,小网站*使用的是公用的虚拟主机,流量有限或者需要付费,希望搜索引擎抓的温柔点;某些网页是动态生成的,没有直接的链接指向,但是希望内*容被搜索引擎抓取和索引。*

9

当前位置: 22.08% **本段(页)重复比例:** 41.78%

*从互联网结构的角度来看,网页通过各种超链接相互连接,形成一个相互关联的大而复杂的有向图。因此,根据深度优先的原则,网络爬虫通常可能多*次抓取多个网页,这意味着重复抓取同一个网页,引发*去重问题。如何规避这类问题已成为纠正深度优先策略的首要任务*。实际上,通常建立爬行路径*优化算法来简化网络爬行器的行走路径,并且需要根据具体情况确保适当*的穿越深度。图2.8 DFS遍历分析:深度优先搜索算法:枝叶上所有节点不保留,占用较少的空间;存在回溯操作,使用栈这种数据结构实现(即,存在堆叠操作和堆叠操作),并且操作速度较慢。广度优先搜索算法:所有节点都保留,但是占用大量空间;没有回溯操作(即无堆叠或堆叠操作),操作速度快。

10

当前位置: 23.38% **本段(页)重复比例:** 15.76%

通常,深度优先搜索方法不保留所有节点。展开的节点从数据库中弹出。通过这种方式,数据库中存储的节点数通常是深度值,因此占用的空间较小。因此,当搜索树中有许多节点,而其他方法发生内存溢出时,深度优先搜索是一种有效的解决方案。广度优先搜索算法通常需要存储所有生成的节点,并且占用的存储空间比深度优先搜索大得多。因此,在程序设计中,必须考虑溢出和节省空间。然而,广度优先搜索方法通常使用回溯操作,即pop和push操作,所以运行速度比深度优先搜索更快。总结一下:所以我采取了广度优先的搜索抓取策略,以避免爬取网站过程中出现深度优先陷入环路不断循环,无法访问其他网站,我们的设计就失败。2.*5非关系型数据库2.5.1 NOSQLNOSQL数据库其实就是与关系型数据库相对应的非关系型数据库,NoSQL用于存*储大型数据或者文档、文件、照片视频等等。

11

当前位置: 25.97% **本段(页)重复比例:** 60.71%

*2)NoSQL的优点优*点:高可扩展性分布式计算成本低架构灵活性,半结构化数据没有复杂的关系2.*5.2Mongodb简介MongoDB是一种面向文档的基于分布式文件存储的数据库,由C++*撰写而成,以此来解决应用程序开发社区中的大量现实问题,并与200*7年10月,MongoDB由10gen团队所发展。MongoDB是间于*SQL与NOSQL之间的。功能最丰富的NOSQL,它更像一个*关系数据库。MongoDB将数据存储为文档,数据结构由键值对组成。MongoDB文档与JSON对象相似。*字段值可以包含其他文档,数组*和文档数组。MongoDB的应用场景:(1)表结构不清晰,数据变大MongoDB是一个非*结构化文档数据库。

12

当前位置: 35.06% **本段(页)重复比例:** 62.50%

*原子:Redis的所有*操作都是原子操作,意味着成功执行或失败。单个操作是原子操作。多个操作还支持事务,原子性,由MULTI和EXEC指令包装。丰富的功能--*Redis还支持发布/订阅,通知,密钥到期等。3.3 Scrapy框架3.3.1Scrapy框架简介Scrapy是为抓取网站数据和提取结构化数据而*编写的应用程序框架。因此可以应用于一系列的包括数据挖掘,信息处理或存储历史数据*等等领域。Scrapy是一个用Python开发的开源*Web爬虫框架,可用于快速*爬取网站并从页面中有效提取结构化数据。Scrapy可广泛用于数据挖掘,监控和自动化测试,提供各种类型的爬虫基类,如BaseSpider,SitemapSpider等。*

13

当前位置: 46.75% **本段(页)重复比例:** 9.32%

项目管道默认使用json。pipelines.py这*用于实现分布式处理(在此处*设置为独立分发)。它将项目存储在redi*s中进行分布式处理。另外,可以发现,*写入管道是一样的,这里的代码的执行不同于分析文章,因为这里需要读取配置,所以使用from_crawler()函数。Cheduler.py该扩展是调度程序自己的调度程序的替代方法(在SCHEDULER变量的设置中指定)。这是用于实现爬虫分布式调度的扩展。它使用的数据结构来自队列中实现的数据结构。Spider spider.py这个蜘蛛设计读取url从redis抓取,然后执行抓取。如果在爬网期间返回更多网址,请继续操作,直到完成所有请求。然后继续从redis中读取url,循环这个过程。

14

当前位置: 49.35% **本段(页)重复比例:** 65.83%

*图4.1 PC端移动*端比较想从wap爬取,但不知道怎么做。例如用PC端浏*览器打开http://weibo.com 在登录*的时候会自动跳回m域名网站,*甚至用requests打开网页时会返*回403错误。这是因为网站服务器会根据你的浏览器表头判断你是从哪个平台发送的请求,识别到PC端的请求会给你作相应处理。因此只需要修改浏览器*表头就是http报文头(User-Agent)即可。如*果是爬虫程序,只需要带上旧版手机

15

当前位置: 51.95% **本段(页)重复比例:** 7.59%

4.2 爬虫伪装4.2.1User-Agent伪装User Agent中文称为用户代理,缩写为UA。http 报文head一个特殊的字符串头,它使服务器能够识别客户*端使用的操作系统和版本,CPU类型,浏览器和版*本,浏览器渲染引擎,浏览器语言,浏览器插件等信息。那么,设置User Agent有什么用?实际上,简而言之,User-Agent是诸如客户端浏览器等应用程序所使用的特殊网络协议。当用户使用浏览器(邮件客户端/搜索引擎蜘蛛)发出HTTP请求时,它就会被发送到服务器。识别用户正在使用什么浏览器(邮件客户端/搜索引擎蜘蛛)来访问。由于它是一种人为规定的协议,因此无论浏览器如何,都可以更改默认的UA。

16

当前位置: 54.55% **本段(页)重复比例:** 51.41%

这两个数据代表链接。提取页面连接的过程总结如下:1确定文件类型是否为*txt /html如果不是*,则跳过,如果要继续分析。*2使用正则表达式匹配方法读取文件以找到标签<a href=>,<area href =>,<base href =>,<frame src =>,<img src =>,<body background =>,<appletcode =>等,记*录下URL。3以预定格式使录制的URL保持一致并完整。准备下一个URL分析。4.*3.2正则表达式简介HTTP协议本*质上就是HTML的传输,所以当我们编写爬虫时会遇到很多HTML页面,我们需要找出我们需要什么,如何检索我们需要的内容是一个关键,这里可能需

17

当前位置: 57.14% **本段(页)重复比例:** 100.00%

简单来说就是通过Hash函数存储网络爬虫的遍历轨迹,并规定某一Web页被遍历过,则在哈希表中的相应槽位填充1,否则填充0。在具体实现过程中,哈希函数起到至关重要的作用,目前一般使用MD5 ()函数,将网页文件的地址即URL字符串转换为128位散列值。MD5就是将任意长度的消息转换成128位固定长度的消息摘要的函数,显然,MD5 ()函数产生的值很大,为2个不同的数,需要的内存空间巨大。因此,在实际处理中还要将MD5()函数的值进行模运算映射到哈希表中。其公式可设为:MD5 (URL)MOD N其中,URL为抓取的地址,N为存储哈希表的位长。通过该式的转换,可使输入的URL地址被映射到大小为N的哈希表的某个位上,以便确定其地址是否被抓取过。

18

当前位置: 58.44% **本段(页)重复比例:** 68.95%

*为了解决重复搜集网页的问题,可以定义两个数据库:“未爬行的URL库”和“已爬行的URL库”。“未爬行的URL库”存储待访问队列的URL“已爬行的U1u库”存储已遍历过的URL。对于已访问过的、未访问过的URL利用MD5(URL)函数分别作MD5摘要,以获取其惟一标识,并建立两个集合。新解析出的URL,首先根据已经访问过URL集合判断是否已抓取过.如没有被抓取.则放入“未爬行的URL数据库”中.否则放入“已爬行的URI库”*4.4.2 Redis去重当数据量不大时,可以直接将其存储在内存中进行重复数据删除。例如,python可以使用set()来删除数据。重复删除需要保留的数据时,使用redis设置的数据结构。

19

当前位置: 67.53% **本段(页)重复比例:** 67.96%

总而言之,封禁类中的如何判断是否是爬虫是反爬系统的核心关键所在,确定了是爬虫之后,那么就不光是封禁,可能会有各种手段惩罚或者戏弄你.比如后面讲的投毒,无限循环,伪装404页面等待.....4.5.*2 对抗AJAX技术在Web用户界面对用户交互和响应灵敏方面却投入不足。用户在强大的业务逻辑背后,还忍受着“提交-响应-等待-刷新”的同步运行机制,不管在页面呈现上的变化多么小,都需要耐心地等待服务器将整个页面重新发送给客户端。为了构建更为动态和响应更灵敏的Web应用程序,实现浏览器和服务器的异步并行处理,减轻服务器端负担,提出了AJAX(Asynchronous JavaScript and XML)这一新的概念。*

20

当前位置: 68.83% **本段(页)重复比例:** 72.09%

*与90年代的Web开发不同,AJAX不以静态页面方式更新整个页面。对于AJAX,Web应用程序应该由表单组成,其中每个表单都是一个较小的AJAX应用程序,每个页面都使用JavaScript开发的AJAX组件。这些组件使用XMLHttpRequest对象以异步I/O方式与服务器进行通信。从服务器获取所需数据后,DOM API更新页面内容。如图4.6 AJAX的工作原理AJAX增强人机交互,*但同时它也对Web爬虫的爬取数据带来了很大的挑战,因为许多表单是通过向服务器异步I/O发送请求获得的,但是传统爬网程序仅仅能分析静态*数据比如HTML页面中的超链接*。显然,传统的爬取方*式来使用AJAX技术在网站中抓取网页是不够的,或者说*无用的因为捕获的信息不完整。

21

当前位置: 74.03% **本段(页)重复比例:** 44.04%

但是!后两者不管是云打码还是图像识别出错率很高,云打码还不支持移动版微博拖动,同时图像识别又太难对于现在的我来说使用,反正没有几个,并且不是搞得多高大上就越好所以用最笨的方法是最见效的同时也是成功率100%的!运行launch.py启动爬虫,中途会要求输入验证码,查看项目路径下新*生成的aa.png,输入验*证码回车,如图4.9,即可顺利成功。如图,4.9 打验证码4.

22

当前位置: 75.32% **本段(页)重复比例:** 75.41%

*服务器可以设置或读取Cookies中包含信息,借此维护用户跟服务器会话中的状态。从cookies的定义可以看出,cookies也是可*以作为一个验证用户身份的工具,*所以可以通过cookies来区别*机器和人,所以有一种反爬的策略,*就是通过cookies,微博的反爬虫机制就是基于cookies,所以同一个cookies可以重复请求,而同一个IP不带cookies却是不*能重复请求,会封IP,但是微博做的还是不够,就是只*需要一个cookies,但是微博的cookies所保存的时间短一点,就需要一个cookies池了,定期加入cookies,这样爬*取的难度就会增大很多,再厉害一点,就*是IP和cookies一起识别*。

23

当前位置: 76.62% **本段(页)重复比例:** 60.96%

4.6.2构建Cookies池登录之前,您不想抓取页面。然后我们可以使用Urllib2库来保存我们的登录cookie,然后再抓取另外一个页面来达到重写中间件的目的,详情请*看官方文档:http://scrapychs.readthedocs.io/zh_CN/latest/topics/downloadermiddleware.html#scrapy.contrib.downloadermiddleware.DownloaderMiddlewar*e*在项目中新建一个middleware*s.py的文件(如果你使用*的新版本的Scrapy,在*新建的时候会有这么一个文件,直接用就好了)首先导User Ag*entMiddlewar*e 毕竟我们要重写它啊第一行:定义了一个类UserA*gentmiddleware继承自UserAgentMiddleware第二行:定义了函数process_request(request,spider)为什么定义这个函数,因为Scrapy每*一个request通过中间件都会调用这个方法。

24

当前位置: 80.52% **本段(页)重复比例:** 2.47%

第5章数据采集与数据分析5.1数据模型5.1.1数据库概念设计概念模型是对信息世界进行建模。有很多方法来表达信息世界。最常用的是实体连接模型,即E-R模型。这是描述现实世界概念模型的E-R图。下面的E-R图,很清楚而简单地描述了该系统的*实体与其属性之间的关系。图5.1 用*户关系表(relationships)图5.2 用户发表的微博(Tweets)图5.3 用户个人信息(information)图5.4 数据库概念模型E-R图5.1.2数据库逻辑设计下面将E-R图转换为关系模型。用户发表的微博(_id ,ID ,Content ,PubTime ,Co_oridinates ,Tools ,Like ,Comment ,Transfer)用户关系表(_id,Host2,Host1)用户个人信息(_id,NickName,Gender,Province,City,BriefIntroduction,Birthday,Num_Tweets:,Num_Follows,Num_Fans,SexOrientation,Sentiment,VIPlevel,Authentication,URL)5.1.3数据表设计本爬虫系统数据库中总共设计了3张数据表,分别为具体结构设计如下:表5.1 用户关系表(relationships)表5.2 用户个人信息(information)表5.3 用户发表的微博(Tweets)图5.5 information:图5.6 relationships图5.7 Tweets5.1.4 Redis搭建下载Redis,我的电脑是win10,所以最好下载64位版本,在运行中输入CMD,然后将目录指向解压缩的Redis目录。

25

当前位置: 81.82% **本段(页)重复比例:** 67.76%

安装过程中出错测试错误:Redis启动错误[9844]02 Dec 13:57:40.682#创建服务器TCP侦听套接字*:6379:bind:Unk*nownerror解决*方案如下输入以下命令以成功连接1.*redis-cli.exe2.关机3.退出4.redis-server.exe redis.windows.conf启动命令:Redis-server.exe redis.windows.conf,图5.8显示启动成功。图5.8 redis安装成功1.2对Redis-cli并进行测试CMD下打开redis-cli.exe,*输入如下命令的测试案例>Set user*info chenwei*yuOK>get user*info*“ chenweiyu ”测试成功图5.9 测试redis成功windows安装redis服务CMD的Redis目录下:命令:s>redis-server --service-install redis.windows-service.conf --loglevel verbose显示成功:图5.10 安装redis服务成功5.1*.5 Mongodb的搭建从MongoDB官方网站下载并安装mongodb3.4 64位创建数据目录MongoDB将数据目录存储在db目录中。

26

当前位置: 84.42% **本段(页)重复比例:** 67.38%

该文件必须具有syst*emLog.path参数集*,包括一些其他配置选项。在C:m*ongodb mongod.*cfg中创建配置文件,其中指定了syst*emLog.path和*storage.dbPath。具体配置如下:systemLog ::目标:文件:*文件路径:c:data log mongod.log:c:data log mongod。日志存储::dbPath:c:data db:c:data db安装Mongodb:运行mongod.exe,使用*--install选项安装服务,并使用*--config选项指*定先前创建的配置文件。“ C:mon*godb bin mongod.exe* ”--config “ C:m*ongodb mongod.cfg* ”--install -配置“ C:m

27

当前位置: 92.21% **本段(页)重复比例:** 22.57%

起初,通用的Web Crawler被用于互联网的所有领域,因为它能够整合信息并使用户能够进行全面的搜索。*尽管如此,普通爬虫的缺点也变得更加明显。例如,内容陈旧*,召回率低,信息冗余,分布不均匀。无法满足特定用户和人群的搜索需求,*因此在这种情况下,基*于网*络爬虫的搜索引擎的主题*就诞生了。它的运行效率直*接关系到下一代搜索引擎的性能.....(1)设*计并实施对微博信息的抓取系统。通过模拟登录的方法,解决了身份认证问题,解决了验证码问题,隐藏了伪装的微博信息移动网络问题,爬虫采用了广度优先搜索的想法。结合Web爬虫,实现了BeautifulSoup,Scrapy框架,正则表达式和多线程并发,Cookie池技术,各种用户信息和微博信息的收集。