毕业设计中期报告

昆山杰普软件科技有限公司 填表日期：2017年4月17日

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 班 级 | 软件1301 | 课题类型 | ☑设计类 □调研类 □实验类 □其他 | |
| 学生姓名 | 杨泽鑫 | 课题名称 | 网络爬虫的设计与实现 | |
| 基地  指导老师 | 李楠 | | 校内指导老师 | 方昀 |
| 毕业设计进度 | 代码完成进度：共设想有三到四块程序：  第一部分：百度百科词条页面爬取，完成程度：基本完成 后期有时间可以增加数据库的部分；  第二部分：花瓣网图片的抓取 基本完成 后期有时间可以增加数据库的部分；  第三部分：qq空间说说的抓取 基本完成 后期可添加有关数据库的操作、利用采集到的数据进行简单的数据分析的操作。  第四部分：采集新闻网站的新闻并进行简单数据分析的操作，完成了单页面新闻采集部分，批量采集和数据分析还未完成。根据后期的前三个的完成程度看是否需要完成这个模块。  论文完成进度：  课题的背景、意义，所使用的的技术、算法已经添加到论文中  百度百科词条的抓取也已经添加到论文之中  还有部分课题的实现过程未添加到论文中  论文的格式还未统一  现在存在的主要问题：  程序中缺少和数据库交互的部分  程序中缺少数据分析的部分  论文中还缺少部分已实现程序的分析 | | | |
| 主要内容 | 论述了网络爬虫的发展历史  论述了网络爬虫所应用到的具体技术  以实例方式剖析了一个网络爬虫的基本架构  探索如何抓取动态网页的方法  进行简单数据分析 | | | |
| 主要成果 | 百度百科词条爬虫完成的功能是可以从一个任意指定的词条页面出发，爬取一系列相关的词条页面URL、词条标题以及词条简介。爬取的数量由用户指定。  花瓣网图片爬虫完成的功能是可以从一个给定的花瓣网的搜索页面出发，自动下载指定数量的原图，因为花瓣网上搜索界面的图片不是一次全部加载出来的，而是通过用户在浏览时下拉刷新完成的，所以为了模仿用户的操作，采用了requests库的gets方法，重置ajax请求，成功实现加载目标。  qq空间说说爬虫完成的功能是给出一个qq空间的说说页面后，可以自动抓取该用户发表的全部说说的日期和文字内容。如果该页面需要登陆查看，程序中也已经预设了自动登录功能。与花瓣网类似，qq空间也是采用的动态加载机制，但每一页能够显示的也是有限制的，所以，我们更换了一种方式来模拟请求，利用第三方类库selenium来模拟一个真实的浏览器来进行自动化操作很有意思。 | | | |
| 基地指导教师意见 | **合格（ ） 不合格（ ）** | | | |
| 基地指导教师签字： | | | |