编　　号：
审定成绩：_____

重庆邮电大学
毕业设计（论文）

设计（论文）题目：校园网内搜索引擎搜索器的分析与设计

学 院 名 称：经济管理学院

学 生 姓 名：张

专　　业：信息管理与信息系统

班　　级：0310701

学　　号：07050122

指 导 教 师：王

答辩组负责人：

填表时间：2011年6月

重庆邮电大学教务处制

摘　要

ABSTRACT

目　录

第一章　课题概述

第一节 课题背景

第二节 发展现状

第三节 课题成果概要

第二章　开发环境介绍

第一节 开发工具与环境

一、 编译器：gcc

二、 调试器：gdb

三、 解释器：python

四、 编辑器：vim

五、 操作系统：GNU/linux

六、 数据库系统：mysql

七、 其它

第二节 开发语言与技术

一、 c

二、 python

三、 c语言内调用python技术

第三章　搜索器需求说明

第一节 任务说明

一、 目标

1. 对当前流行的全文搜索引擎进行认真的研究，以了解其主要工作流程，方便后期进行系统的设计与开发工作。
2. 高度重视系统的分析与设计工作，保证系统有较高的可扩展性和安全性。
3. 系统使用C语言、Python语言等多种语言合作开发，在保证系统可高效、稳定运行的同时要充分发挥不同类型语言的优势，另外，还就注意处理好不同语言间的数据交互问题。
4. 搜索器应该包含网络爬虫和网页链接提取的功能，可实现自动化地下载网页、提取链接的工作。

二、 运行环境

操作系统 Linux2.6
数据库 mysql5.5
Python python2.7
CPU intel-i386

三、 条件与限制

1. 开发、测试与运行环境均处于校园网内，网络环境相对简单，并且网速较快，不会成为系统的限制因素。
2. 有较多的分析与设计案例可以参考。
3. 开发与测试的主机硬件条件（CPU频率、内存容量、硬盘容量等）较差，难以满足系统高负荷运行的需要。
4. 开发者本身对开发语言等并不足够熟悉。

第二节 数据描述

第三节 功能需求

搜索器包含网络爬虫与链接提取的功能。从初始链接开始，在配置文件

和数据库中相关信息的辅助与限制下，可以自动地进行网页下载以及从网页源码中提取链接的工作。

一、 网络爬虫

1. 可从初始链接开始自动到校园网中抓取网页。
2. 能够识别链接的权重，并先下载权重较高的网页。
3. 可以处理重复的网页，以及链接相同但内容不同的网页

二、 网页链接提取

1. 不断地对新出现的网页进行分析，提取其中的链接。
2. 可提取以不同形式出现的链接，如html标签¡a¿中出现的链接、javascript重定向的链接等。
3. 具有计算链接权重的能力。

第四章　搜索器系统设计

第一节 总体设计

第二节 详细设计

第五章　搜索器开发与测试

第一节 前期准备工作

第二节 开发进度

第三节 测试计划

第四节 测试分析报告

第六章　系统试运行

第一节 运行环境

第二节 运行情况

一、 系统稳定性

系统运行N天，出现严重错误X次（还是不出现的好）。

系统最大连续无故障运行M小时，产生的数据正常，

二、 系统执行效率

搜索器平均运行N小时便可得到校园网内约60％的网页数据。

对于web1，web2等校内访问量较大（重要性高）的网站的更新，搜索器可以在约M分钟内发现，并完成新页面的下载。

第七章　结论与前景

第一节 课题研究结论

第二节 搜索引擎研究与开发前景

参考文献

[1] He, short latex, tshp, 1

致　谢

谢谢！

第八章　附录

一、 英文原文

二、 中文翻译

三、 主要源代码