目录

[1 引言 2](#_Toc483164254)

[1.1 编写目的 2](#_Toc483164255)

[1.2 命名规范 2](#_Toc483164256)

[1.3 术语定义 2](#_Toc483164257)

[1.4 参考资料 2](#_Toc483164258)

[1.5 相关文档 2](#_Toc483164259)

[1.6 版本信息更新 2](#_Toc483164260)

[2 总体设计 3](#_Toc483164261)

[2.1 硬件运行环境 3](#_Toc483164262)

[2.2 软件运行环境 3](#_Toc483164263)

[2.3 开发环境 3](#_Toc483164264)

[2.4 子系统清单 4](#_Toc483164265)

[2.5 功能模块清单 4](#_Toc483164266)

[3 数据库设计 4](#_Toc483164267)

[3.1 数据库中表名列表 4](#_Toc483164268)

[3.2 数据库表之间的关系 5](#_Toc483164269)

[3.3 数据库表的详细清单 5](#_Toc483164270)

[3.4 数据库的视图设计 5](#_Toc483164271)

[4 功能模块设计 5](#_Toc483164272)

[5 存储过程设计 5](#_Toc483164273)

[6 接口设计 5](#_Toc483164274)

[6.1 用户接口 5](#_Toc483164275)

[6.2 外部接口 5](#_Toc483164276)

[6.3 内部接口 6](#_Toc483164277)

[7 角色授权设计 6](#_Toc483164278)

[8 系统错误处理 6](#_Toc483164279)

[8.1 出错信息 6](#_Toc483164280)

[8.2 故障预防与补救 6](#_Toc483164281)

[8.3 系统维护设计 6](#_Toc483164282)

[9 测试计划 7](#_Toc483164283)

# 1 引言

## 1.1 编写目的

完成“爬取大数据有关新闻的网站”的网站设计需求分析的基础上，我们需要编写本文档，来文档化对系统的设计。并且利用本文档以及相应的规格说明书实现“爬取与大数据有关的新闻”网站。

## 1.2 命名规范

数据库表名，字段名统一用大写的英文字母开头。

函数参数变量统一用“\_”开头的小写英文命名。其他变量用大写英文字母开头

所有函数用大写英文字母开头。

## 1.3 术语定义

（1）大数据：指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

（2）框架：是一个由框架网页所定义的浏览器初见窗区域，它通过框架网页实现。框架网页是一种特别的HTML网页，它可将浏览器视窗分为不同的框架，而每一个框架则可显示一个不同网页。

（3）网络爬虫：是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本

## 1.4 参考资料

[1]”爬取与大数据有关的新闻网站“的《需求规格说明书》

[2]“爬取与大数据有关的新闻网站“的《软件开发说明书》

[3]吕云翔.软件工程实用教程.北京：清华大学出版社，2015（2017.2重印）.

[4]软件设计说明书编写指南

## 1.5 相关文档

[1]”爬取与大数据有关的新闻网站“的《需求规格说明书》

[2]“爬取与大数据有关的新闻网站“的《软件开发说明书》

## 1.6 版本信息更新

版本更新信息记录如表1所示。

表 1 版本更新记录

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 版本号 | 创建者 | 创建日期 | 维护者 | 维护日期 | 维护纪要 |
| V1.0 | 颜正 | 2017-4-19 |  |  |  |
| V1.1 |  |  | 邓经纬 | 2017-4-23 | 更新术语 |
| V1.2 |  |  | 谢仰哲 | 2017-4-25 | 更新接口信息 |
| V1.3 |  |  | 韩良喆 | 2017-4-26 | 更新模块信息 |
| V1.4 |  |  | 杨培钦 | 2017-4-29 | 更新错误处理信息 |

# 2 总体设计

## 2.1 硬件运行环境

1. 服务器

处理器型号：AMD或Intel 2.0GHz以上

内存容量：1GB以上

网络配置：100M网卡

1. Web浏览PC

处理器型号：AMD或Intel 1.6GHz以上

内存要求：256MB以上

网络配置：100M网卡

## 2.2 软件运行环境

1. 服务器

操作系统：Linux CentOs 6(64位)

运行环境：Python3.5, php5.7

1. 客户机

操作系统无限制，有网络、浏览器即可。

## 2.3 开发环境

1. 硬件环境

本系统采用PC开发，配置如下。

处理器型号：AMD/Intel 1.6GHz及以上。

内存剩余空间：512MB及以上。

外存剩余空间：1GB及以上。

网络配置：100M网卡、串口。

1. 软件环境

1）服务器

操作系统：Windows10及Linux CentOs 6(64位)。

数据库：MySql5.7。

Web服务器：apache2.4。

2)客户机

操作系统等均无限制，有浏览器、网络即可。

## 2.4 子系统清单

子系统清单如表2所示

表2 子系统清单

|  |  |  |
| --- | --- | --- |
| 子系统编号 | 子系统名称 | 子系统功能简述 |
| SS1 | 浏览网页 | 1. 浏览摘要的内容 2. 听取摘要的语音 3. 跳转至原文网站 |
| SS2 | 数据管理 | 1. 将网页内容进行摘要 2. 将摘要转换为语音 3. 提取网页的链接 |

## 2.5 功能模块清单

功能模块清单如表3所示。

表3 功能模块清单

|  |  |  |
| --- | --- | --- |
| 模块编号 | 名称 | 模块功能描述 |
| SS1-1 | 浏览摘要 | 浏览最新新闻的摘要内容 |
| SS1-2 | 听取语音 | 听取摘要转换的中文语音内容 |
| SS1-3 | 跳转原文 | 通过原文链接跳转至原文网页 |
| SS2-1 | 进行摘要 | 将新闻原文进行摘要 |
| SS2-2 | 获取语音 | 将新闻摘要转化为中文语音 |
| SS2-3 | 获取原文链接 | 获取新闻原文的网页链接 |

# 3 数据库设计

## 3.1 数据库中表名列表

数据库中表名列表格式如表4所示。

表4 数据库中表名列表

|  |  |  |
| --- | --- | --- |
| 编号 | 表名 | 表功能说明 |
| 1 | News | 储存需要的信息 |

## 3.2 数据库表之间的关系

数据库中的表之间的关系如图1所示。



图1

## 数据库表的详细清单

News表如表5所示。

表5 News表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 英文字段名 | 中文字段名 | 数据类型 | 是否允许为空 | 主键/外键 |
| 1 | Id | 身份标识码 | Int(11) | 否 | 主键 |
| 2 | Title | 题目 | Varchar(80) | 是 |  |
| 3 | Abstract | 摘要 | Varchar(400) | 是 |  |
| 4 | Uptime | 更新时间 | Varchar(40) | 否 |  |
| 5 | Link | 原文链接 | Varchar(100) | 是 |  |
| 6 | Audio | 是否更新音频 | Int(11) | 是 |  |

# 4 功能模块设计

此项目后端共涉及7个模块，分别是ConnectToDatabase,py, GetTime.py, initNews,py, MyCrawler,py, MyTTS,py, UpdateAudio,py, updateNews,py

1. 连接数据库模块：ConnectToDatabase.py

参数：无

调用背景：需要向数据库读取数据

调用描述：使用pymysql库连接数据库，返回一个数据库的链接。

1. 获取当前时间模块：GetTime.py

参数：无

调用背景：需要当前新闻的更新时间

调用描述：返回当前一个格式化的时间

1. 爬虫主体模块：MyCrawler,py

参数：源网页链接

调用背景：实现主要功能

调用描述：使用requests库爬取网页的源代码，并使用beautifulsoup库解析HTML文档，获取新闻的标题和摘要。具体方式为输入原网页的链接，返回一个储存序号标题和摘要的字典。

1. 文字转语音模块: MyTTS,py

参数：需要转为语音的文字、储存的文件名

调用背景：实现文字转语音接口

调用描述：调用来自台湾高等研究院的文字转语音接口，输入文字和文件名，返回一个音频的下载链接并将音频下载到服务器本地。

1. 自动生成语音模块: UpdateAudio,py

参数：无

调用背景：爬取新闻标题和摘要后需要将它们转化为语音

调用描述：接收爬虫服务器模块主体返回的数据，将数据传给文字转语音模块，生成每一条新闻的音频。

1. 初始化数据模块: initNews,py

初始化数据

1. 自动更新数据模块: updateNews,py

参数：无

调用背景：主程序入口

调用描述：程序的主入口。获取新闻数据和数据库链接，将数据存入数据库并生成数据的语音。最终将各个功能模块得到的数据利用PHP和MySql按照设计的格式显示在网页相应的位置。

# 5 接口设计

## 5.1 用户接口

用php5.6搭建，提供清晰简洁易用美观的用户界面。将爬取的新闻简明扼要的展现在用户眼前。

## 5.2 外部接口

数据存储：本系统设计新闻的各个要素的数据存储和处理。通过LNMP（Linux+ Nginx +MySql + PHP）和MySQL 5.6.3进行交互。其中有PHP 5.6.3版本和Python 3.4版本。

## 5.3 内部接口

本系统以数据为中心，网站各模块通过PHP和数据库服务器进行交互。当页面跳转时，通过session及query string传递参数。

# 6 角色授权设计

角色授权如表 所示。

表 角色授权

|  |  |  |  |
| --- | --- | --- | --- |
| 模块编号 | 模块 | 用户 | 管理员 |
| SS1-1 | 浏览摘要 | 1 | 1 |
| SS1-2 | 听取语音 | 1 | 1 |
| SS1-3 | 跳转原文 | 1 | 1 |
| SS2-1 | 进行摘要 | 0 | 1 |
| SS2-2 | 获取语音 | 0 | 1 |
| SS2-3 | 获取原文链接 | 0 | 1 |

# 系统错误处理

## 7.1 出错信息

1. 定期对所爬取新闻的原网站进行检索和检查，确保原新闻能正常阅读。
2. 当爬取的原新闻被删除时，及时更新网站并加以说明。
3. 对语音朗读功能可能产生的异常进行捕获，按统一的方式对语音进行修改。

## 7.2 故障预防与补救

1. 对已经爬取的新闻应该备份，防止数据库重要信息的丢失。
2. 做到对多个网站的新闻爬取，可以减少原网站崩溃等不确定因素带来的损失。
3. 对数据库中的数据信息做到加密保护，减少黑客攻击的危害。

## 7.3 系统维护设计

1. 编码时尽量提高模块内部的内聚，减少模块间的耦合。采用模块化和分层的思想。使系统的逻辑清晰，网站显示的内容简要。增强可读性和可维护性。
2. 在编码时注意函数和变量的命名方式，同时在适当的地方给予必要的注释和分析。方便之后的设计和维护。

# 8 测试计划

测试计划作为本系统的测试指导，测试成员安计划和规定进行测试。本网站的测试将从单元测试，集成测试和系统测试三个方面来设计