博物馆新闻采集分析子系统设计报告

**目录**

1. 导言...................................................................................................................3

1.1 系统设计背景...........................................................................................3

1.2系统设计前提...........................................................................................3

1.3系统设计目的...........................................................................................3

1. 体系结构..........................................................................................................3
   1. 系统运行环境...........................................................................................3
   2. 系统体系结构...........................................................................................4
2. 数据库定义......................................................................................................4
   1. 概述.............................................................................................................4
   2. 数据项定义................................................................................................5
   3. 接口定义.....................................................................................................5
3. 模块设计...........................................................................................................6
   1. 模块设计.....................................................................................................6
   2. 流程图..........................................................................................................8
4. 项目结构............................................................................................................9
5. 实现方法............................................................................................................9

6.1数据获取.....................................................................................................9

6.2数据处理....................................................................................................10

6.3数据分析....................................................................................................10

6.4服务端控制...............................................................................................10

6.5数据入库更新..........................................................................................10

**1. 导言**

**1.1系统设计背景**

随着社会的繁荣、经济的发展和科学技术的进步,博物馆新闻宣传面临前所未有的发展机遇和挑战。博物馆已经成为思想意识形态领域的一个重要阵地,也成为中国传播政治思想、弘扬民族传统文化的重要场所。本子系统着重在对博物馆的新闻进行采集筛选分析，来使人们对各大博物馆发生的事件有更好的了解认识。

**1.2 系统设计前提**

项目的目标是开发一款博物馆新闻分析子系统，同时组员们获得系统的软件工程项目训练。发布的产品是子系统的源代码、技术文档和用户手册等。主要工作是需求分析、系统设计、开发测试，关键里程碑分别为需求规格说明书的发布，系统设计说明书的发布，和系统的交付。小组成员需对系统进行了较为详细的需求分析及具备一定的系统设计理论知识。

**1.3 系统设计目的**

通过程序能从主要的新闻网站上爬取博物馆相关的新闻信息，进行加工处理。主要包括以下功能：

数据获取：爬取主要的新闻网站中的博物馆相关新闻（如百度新闻）。可以支持爬取指定时间范围内的新闻，如1年内的新闻，半年内的新闻等。

数据加工：对于爬取的信息进行过滤和加工，抽取需要的内容。例如，抽取新闻的发布时间、新闻的标题、新闻的内容、新闻涉及的博物馆等。

数据分析：对于加工好的新闻，分析是正面新闻还是负面信息。可采用已有的可直接调用的服务和代码实现。

数据定制服务：可以根据指定的某一家博物馆，获取该博物馆的指定时间的新闻，并进行加工分析，得到该博物馆指定时间内的主要新闻，正面新闻和负面新闻。

同时帮助开发人员进行系统的实现部署，并与其他子系统达成一致。

**2. 体系结构**

**2.1 系统运行环境**

- 操作系统

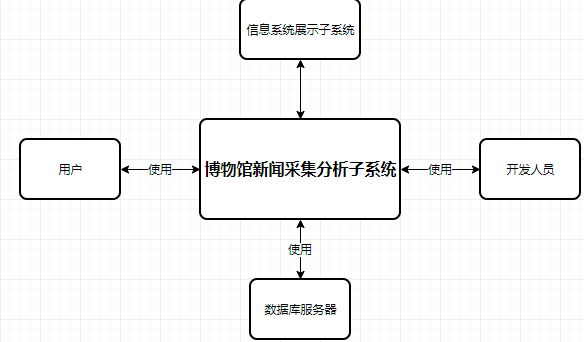
- Win10/Macos/Linux

- 数据库

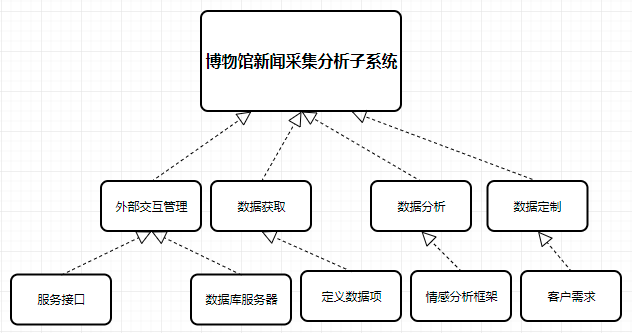
- MySQL 或其他数据库操作软件

- 浏览器

- Microsoft Edge/Google Chrome



**2.2 系统体系结构**



**3. 数据库定义**

**3.1 概述**

该子系统的数据来源是各大正规新闻网站有关博物馆的新闻。爬取内容包括新闻标题，发布时间，新闻正文，涉及到的博物馆以及新闻的分析。数据库的概念模型设计用来反映现实世界中的实体、属性和它们之间的关系等的原始数据形式。逻辑结构设计是将各局部的E-R图进行分解、合并后重新组织起来形成数据库全局逻辑结构，包括所确定的关键字和属性、重新确定的记录结构、所建立的各个数据之间的相互关系。而信息存储结构的设计在系统的设计中至关重要，要考虑到数据冗余、系统执行效率、信息控制以及维护等方面的要求。信息的管理离不开数据库的支持，我们采用mysql数据库管理系统。 数据库的物理设计主要是对数据在内存中的安排，包括对索引区、缓冲区的设计；对使用的外存设备及外存空间的组织，包括索引区、数据块的组织与划分；设置访问数据的方式方法。

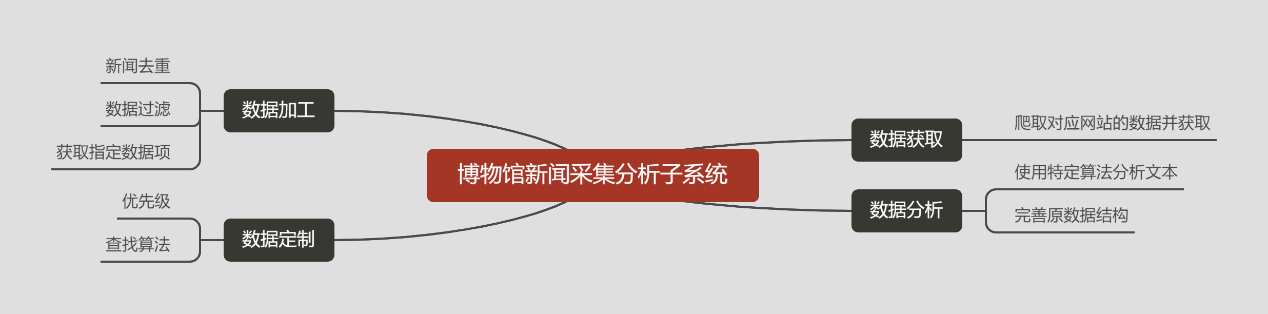
**3.2 数据项定义**

表名newsall

| **序号** | **字段** | **类型** | **关键值** | **描述** |
| --- | --- | --- | --- | --- |
| 1 | id | int(11) | 是 | 序号 |
| 2 | museum | varchar(255) | 否 | 博物馆名称 |
| 3 | title2 | varchar(255) | 否 | 新闻标题 |
| 4 | author | varchar(255) | 否 | 新闻作者 |
| 5 | time | varchar(255) | 否 | 新闻时间 |
| 6 | content | text | 否 | 新闻内容 |
| 7 | url | text | 否 | 新闻url |
| 8 | emotions | double | 否 | 新闻正负 |

**4. 模块设计**

**4.1 模块设计**

****

1. **数据获取**

用例描述：子系统对新闻数据进行爬取。爬取主要的新闻网站中的博物馆相关新闻。可以支持爬取指定时间范围内的新闻，如1年内的新闻，半年内的新闻等。

执行者：子系统

前置条件：用户登陆并进入系统

后置条件：子系统对爬取的信息进行加工

基本路径：用户登入到此系统内，根据其他子系统提供的接口，查看自己所需的新闻。

1. **数据加工**

用例描述：对爬取的信息进行加工。子系统对于爬取的信息进行过滤和加工，抽取需要的内容。例如，抽取新闻的发布时间、新闻的标题、新闻的内容、新闻涉及的博物馆等。

执行者：博物馆新闻分析子系统

前置条件：子系统已经对数据爬取完成

后置条件：信息加工好后，系统对信息进行分析

基本路径：1.对爬取的信息进行过滤和加工。2. 从初步加工的信息中抽取所需的内容。

1. **数据分析**

用例描述：对加工好的新闻，分析是否是正面信息。对于加工好的新闻，分析是正面新闻还是负面信息，并对其进行一个百分制的打分。

执行者：博物馆新闻分析子系统

前置条件：对爬取的信息已加工完成

后置条件：分析结束后，用户可获得所需新闻

基本路径：1. 采用已有的可直接调用的服务和代码实现。2. 分析是正面新闻还是负面新闻

1. **数据定制服务**

用例描述：根据用户需求对指定的博物馆，获取该博物馆指定时间的新闻，并进行加工和分析。可以根据指定的某一家博物馆及用户的指令，获取该博物馆的指定时间的新闻，并进行加工分析，得到该博物馆指定时间内的主要新闻，正面新闻和负面新闻。

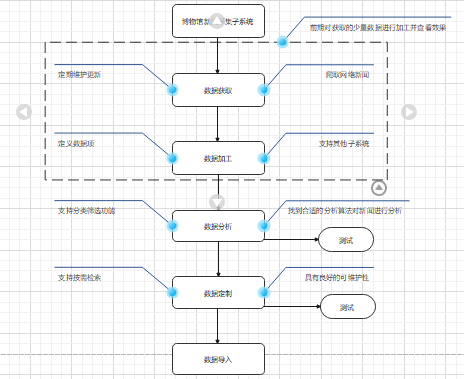
执行者：用户

前置条件：用户已选择指定的博物馆和时间

后置条件：用户指定博物馆和时间后，获得相应的新闻信息

基本路径：1. 指定某一个博物馆，指定获该博物馆新闻的时间。2. 得到该博物馆指定时间内的主要新闻，正面新闻和负面新闻。

**4.2流程图**



**5. 项目结构**

博物馆新闻获取分析子系统

·文档

·项目管理计划

·需求规格说明

·设计报告

·开发周记

·测试报告

·用户使用手册

·代码

·数据获取与处理

·新闻标题情感分析

·服务端代码控制

·数据

·远程数据库中博物馆新闻表的定义与操作

1. **实现方法**
   1. **数据获取**

在Python爬虫中，我们利用 xpath 快速地定位 HTML\XML 响应中的特定元素以及获取节点的信息，并且通常情况下会比使用正则表达式提取更简单而且更高效。首先是通过response.xpath实现百度新闻的数据获取与搜索引擎传参，接下来使用contest.xpath从html中提取所有正文内容。同时定义了scrapy中spider的各种进程请求响应以及时了解到爬取的异常。提取出全部相关新闻页的链接与标签，再爬取提取出来的链接对应的界面，分析源码后得到新闻的标题、发布时间、内容和封面链接。

* 1. **数据处理**

piplines中将爬取的数据整理并写入数据库中。其中processitem（）为pipelines中的主要处理方法，默认会自动调用。之后将存在缺失字段的新闻丢弃（如无时间标注等），并且对于掺杂在正文间的无用信息（例如记者信息、新闻来源等）通过人工排查形式和正则表达式进行删除。

* 1. **新闻情感分析**

调用Python类库snownlp，逐条分析新闻标题情感。由于该模块内并没有直接涵盖新闻类分析，则需要调用tensorflow库预先进行训练集训练。这里选择创造一些常见的新闻标题作为正面训练集和负面训练集。为了防止过拟合现象，训练集总量不超过全部数据的百分之三十。最后通过分数来体现该新闻的正面/负面程度。

* 1. **服务端控制**

实现工具 Pycharm专业版。

实现步骤如下：

⑴ 通过pycharm的内置功能tool-deployment-configuration实现远程连接服务器ip，配置端口等相关属性。在服务器文件夹中建立对应的项目文件，以便文件上传。

⑵ 通过update功能上传代码。

**6.5 数据入库更新**

在Python环境下调用pymysql以直接连接远程数据库。为了避免产生主键重复的错误（数据库中主键是新闻编号），首先需要确定数据表中现有的主键的最大值，作为本次数据更新入库的起始新闻编号。之后则遍历字典值内爬取且处理过的新闻，并对其执行pymysql内置的查询语句，判断这条新闻的标题是否存在，如果不存在并且该数据没有字段缺失，就存储到定义的临时变量里，否则继续遍历下一条新闻。最后将数据导入数据库中的新闻表。