知识图谱构建子系统

1.引言

1.1编写目的

本项目为知识图谱构建子系统，书写此文档是为了确定用户对本系统的真正需求，确定一份完整、准确、清晰、具体的系统要求及设计方案，从而有效地帮助开发人员实现此系统的各个模块和各项功能，也让用户对此系统有更全面的了解。

知识图谱构建子系统将构建出整体的数据库信息，用于掌上博物馆、海外文物知识服务子系统、后台管理子系统等子系统的信息调取。明确各个子系统需求，从而确定出一份准确的数据库设计方案，有效地帮助开发人员实现彼此子系统的模块功能的信息调用

1.2背景

海外中国文物信息数据库的建立，能够较为全面地掌握海外中国文物信息，结合国内文物信息，能够形成相对完整的我国历史文物信息，可较好的研究和反映中国历史，为未来文物追索、文物征集和文物研究保护提供有力的协助。 知识图谱构建子系统就是着手于海外文物信息的爬取和处理，录入到数据库中形成完善的信息来源，通过图数据库的方式更好地展现信息，搭建云数据库提供云端渠道，以供各方面方便使用。

1.3定义

web crawler：是一种自动化程序或脚本，自动地遍历、采集和存储互联网上的信息，以便后续的处理和分析

Neo4j：高性能的 NoSQL 图形数据库，将结构化数据存储在网络上而不是表中，是一个高性能的图引擎，该引擎具有成熟数据库的所有特性

SQL语言：用于数据库操纵的标准语言

1.4参考资料

（1）CIDOC-CRM参考文档：https://cidoc-crm.org/html/cidoc_crm_v7.1.1_with_translations.html#E1

（2）建模工具Karma：https://github.com/usc-isi-i2/Web-Karma

（3）Virtuoso数据库，用来管理三元组:https://github.com/openlink/virtuoso-opensource

（4）基于知识图谱的玉米病虫害问答系统研究

2.任务概述

2.1 目标

（1）数据爬取

从指定网站爬取文物信息，将信息按照名字、年代、作者、介绍等进行分析处理，保存数据后转化为三元组形式，以所列信息为表头构建mysql数据表格。

（2）数据补充

对现有数据的缺失项进行补充，具体为从其他来源进行爬取或人工补充，补充信息缺失项，完善mysql数据库。

（3）图数据库构建

学习如何构建图数据库，将建模好的三元组数据保存到Neo4j图数据库中，用于知识图谱可视化、问答等后续功能开发。

（4）云数据库构建

发布成链接开放数据，可供访问使用，将全部数据（用户数据和文物数据等）需要保存到mysql数据库中。

2.2 用户特点

使用海外文物信息的知识图谱数据的人员,用知识与知识之间的联系进行知识图谱可视化，利用爬虫实时更新知识图谱的数据,半自动化构建知识图谱。

3.总体设计

3.1 功能

（1）数据建模

将爬取的数据清洗,剔除不必要的信息,得到规范的爬取信息。将每个文物的相关数据处理成实体,抽取关系,构建成三元组形式，向知识图谱里添加新实体和关系

（2）数据支持

为其它子系统的建设提供文物数据包括文物名称，文物年限，作者等详细情况，便于其他系统功能的实现

（3）数据可视化显示

通过图数据库将文物知识图谱可视化，通过各种不同的角度对文物进行展示，便于全面理解所包含文物的各种信息

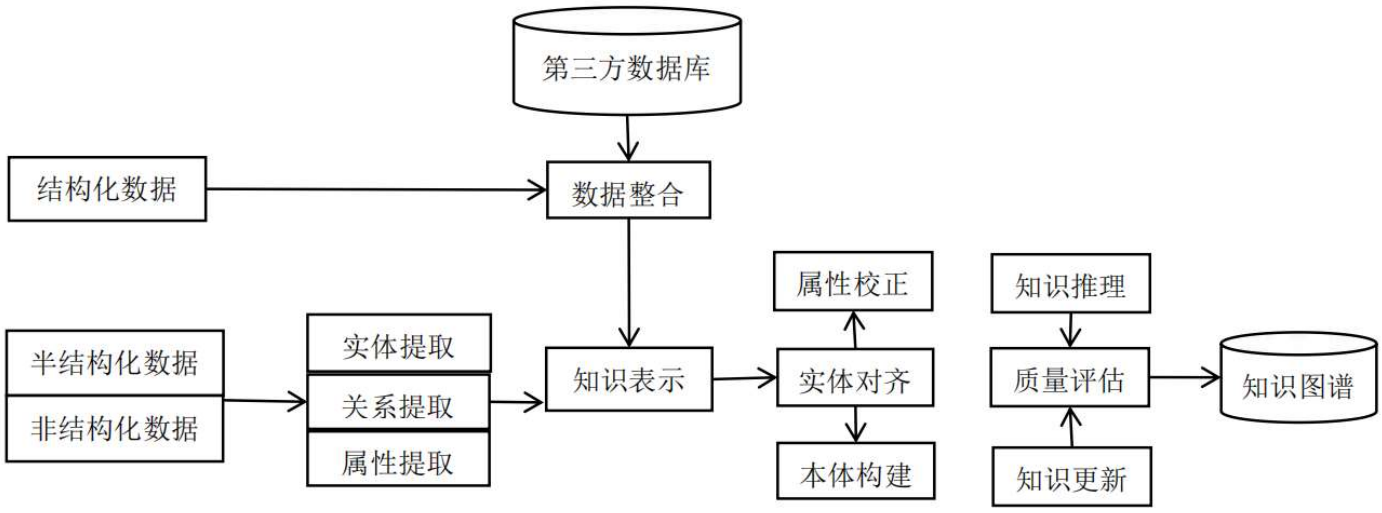
（4）数据上传和保存

定时对数据进行更新检查并将数据上传至云平台，保证数据完整与正确

3.2 运行环境

CPU：尽量使用多核CPU
RAM：建议1G及以上

3.3系统操作流程图



4.接口设计

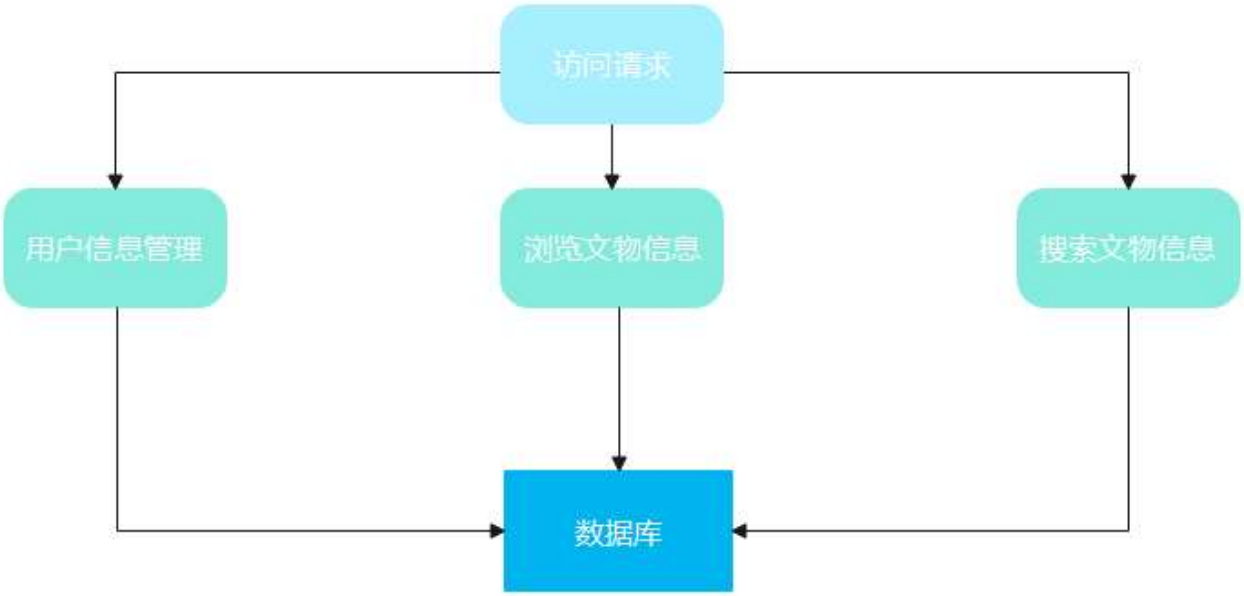
4.1外部接口

数据库采用mysql语句编写，部署在腾讯云端。可通过前端访问程序访问数据库。

4.2内部接口

服务器接收到访问请求时，首先判断该请求的合法性，然后根据不同的操作请求进入相应的操作模块，调用相应的模块处理请求，随后将对应数据发送给客户端。

后台数据库操作模块



5.数据结构设计

5.1 逻辑结构设计要点

文物：

- 文物ID
- 文物名称
- 博物馆名称
- 文物时代
- 出土地
- 文物规格
- 文物图片
- 捐赠信息
- 文物材质
- 文物故事
- 用途

用户：

- 用户ID
- 用户名
- 密码

- 性别
- 年龄
- 地址
- 类型

5.2物理结构设计要点

文物存储：

序号	名称	作者	時代	出土・国	備考	撮影部位	撮影日	数量	图片
1527	蓋弓帽	(Null)	戦国時代 _ 前4-前3c	中国出土	(Null)	なり	2003-03-03	6本	https://image.tnm.jp/imag
1528	蓋弓帽	(Null)	前漢時代 _ 前2-前1c	中国出土	(Null)	正面	2003-03-03	2本	https://image.tnm.jp/imag
1529	くつわ	(Null)	夏家店上層文化 _ 前9-前7c	推定　内蒙古東部出土中国	(Null)	集合	2000-03-21	3面	https://image.tnm.jp/imag
1530	行書五言律詩軸	王鐸	明時代 _ 永暦元年(1647)	(Null)	(Null)	全図	2021-10-13	1幅	https://image.tnm.jp/imag
1531	行書「槐安」軸	呉昌碩	中華民国15年(1926)	(Null)	高奇萄次郎（槐安）筆、手	俯瞰1	2020-10-13	1幅	https://image.tnm.jp/imag
1532	青磁瑞獣	(Null)	唐時代 _ 8 c	中国山西省天龍山石窟第14	(Null)	正面	2013-10-03	1個	https://image.tnm.jp/imag
1533	四神四獣鏡	(Null)	古墳時代 _ 4c	群馬県高崎市柴崎町鷺沢 鷺	〔中国製・正始元年（240）	鏡面	2010-03-23	1面	https://image.tnm.jp/imag
1534	十六羅漢図軸(第八尊者)	金大受	南宋時代 _ 12c	(Null)	(Null)	全図	2022-01-12	1幅	https://image.tnm.jp/imag
1535	十六羅漢図軸 (第十六尊者)	金大受	南宋時代 _ 12c	(Null)	(Null)	裱嵌表	2022-01-12	1幅	https://image.tnm.jp/imag
1536	十六羅漢図軸(第十五尊者)	金大受	南宋時代 _ 12c	(Null)	(Null)	裱嵌表	2022-01-12	1幅	https://image.tnm.jp/imag
1537	十六羅漢図軸(第十三尊者)	金大受	南宋時代 _ 12c	(Null)	(Null)	全図	2022-01-12	1幅	https://image.tnm.jp/imag
1538	十六羅漢図軸(第十一尊者)	金大受	南宋時代 _ 12c	(Null)	(Null)	全図	2022-01-12	1幅	https://image.tnm.jp/imag
1539	十六羅漢図軸(第九尊者)	金大受	南宋時代 _ 12c	(Null)	(Null)	裱嵌表	2022-01-12	1幅	https://image.tnm.jp/imag

用户管理：

userno	username	pwd	usersex	userage	address	type	review_permission
1	1234	awdawd	男	18	北京化工大学	superadr	允许
2	123	21412412	待选	(Null)	(Null)	user	允许
3	32423	34243223	待选	(Null)	(Null)	user	允许

6.系统出错设计

6.1出错信息

系统输出信息的形式	含义	处理方法
数据库无法连接	由于网络堵塞繁忙，数据库软件繁忙，连接数据库配置不正确等因素导致数据库无法连接	等待连接、修复网络、更改数据库连接方法等。
数据库存储信息错误	将爬取的信息导入数据库中时，可能由于文档形式或数据库表格创建不统一导致数据库存储信息错误。	对文本文件中确实的信息进行检查，同时确保导入数据库中时不会丢失或插入错误信息。
数据库信息不同步	数据存储在关系型数据库和图数据库中，修改其中一个数据库的信息，但没有更新到另一数据库，会导致数据不同。	在对数据库进行更改时，同步更改另一数据库中的信息。