知识图谱构建子系统

完成人员：
岳孟涵（组长）、张佳琦、刘佳煜、王艺诺、吴昊函、韩明辰

1 引言

1.1标识

知识图谱构建子系统是通过数据爬取从指定网站爬取文物信息，将信息分析处理、保存转化构建mysql数据库；同时对现有数据的缺失项进行补充，完善mysql数据库；并对完善后的数据库建模保存到Neo4j图数据库中，用于后续功能开发；最终构建云数据库，发布成链接开放数据，可供访问使用的数据功能子系统。

1.2系统概述

该子系统爬取海外博物馆网站的中国文物的名字、文物图片、年代、 介绍等信息，按照规定格式保存下载的数据，并将爬取的数据转化为三元组形式。同时补充现有数据，如书画作家信息等。将建模好的数据三元组数据保存到Neo4j 图数据库中，发布成链接开放数据，用于关系图谱、时间轴等知识图谱可视化、问答等功能开发;全部数据(用户数据和文物数据等)均保存到云数据库(mysql)中。

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。

1.3文档概述

该文档用于向用户介绍该数据库的作用与使用方法，描述数据库建立的具体情况，并说明其在整个元件系统中的作用。

2 系统统综综述

2.1 系统简介

整个系统主要分为四大部分：

（一） 数据获取（Data Acquisition）

数据获取的来源为从指定网站爬取文物信息，指定网站为五家海外博物馆网站，通过代码抓取这些海外博物馆网站的中国文物信息，然后将信息按照名字、年代、作者、介绍等进行分析处理，保存数据为统一表头的表格形式，对于缺失的项进行人工补充，具体为从其他来源进行爬取或人工补充，补充信息缺失项，完善数据库。将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，将网络上公开的爬取数据解析为结构化数据，以此为主体进行数据库的构建。

(二）信息抽取（Information Extraction）

信息抽取是从非结构化文本中自动抽取有意义信息的技术。信息抽取的目的是将这些大量的非结构化文本数据转换为结构化数据，使其可以更轻松地被计算机处理和分析。

信息抽取的过程通常包括三个步骤：命名实体识别（Named Entity Recognition, NER）、关系抽取（Relation Extraction）和事件提取（Event Extraction）。在命名实体识别阶段，系统会识别文本中的实体（例如人名、组织名、地点、时间、数字等），并标记它们的类别。接下来，在关系抽取阶段，系统会读取已经识别的实体并确定它们之间的关系。在事件提取阶段，系统会基于识别的实体和关系提取出文本中描述的事件和行动。

（三）知识融合（Knowledge Fusion）

将来自不同数据源的信息进行集成和组合，以生成新的、更丰富、更准确的知识或信息。知识融合的目标是利用多样化的信息来填补知识空白，并生成更高质量、更精确的知识。

知识融合可以分为两种主要类型：基于内容的融合和基于实例的融合。在基于内容的融合中，从不同数据源中提取出的信息被组合在一起，形成更全面、更丰富的知识。基于实例的融合则基于相似性和重叠性，将不同数据源中的实例进行归并和合并。

（四）知识加工（Knowledge Processing）

海量数据经过信息抽取和知识融合之后，需要通过知识加工进一步处理以形成结构化的、网络化的知识体系。知识加工主要分为三个方面的内容：本体构建、知识推理和质量评估。

本体构建是指将经过处理后的知识分析进行描述和抽象，形成一个针对特定领域的本体。本体可以理解为一种知识模型，在该领域内定义相关的概念、关系和属性等。通过本体构建，不同数据源的知识可以进行统一、标准化的表达，方便后续的知识推理和应用。

知识推理是指利用本体和规则，对知识进行逻辑推理和推断，生成新的知识表达形式。知识推理可以基于本体和规则库进行推理，从而生成更为深入和广泛的知识。质量评估是指对经过加工处理后的知识进行评估，以确保知识库的质量和准确性。质量评估可以通过计算各种指标、采用人工审核等方式进行。

通过本体构建、知识推理和质量评估等过程，可以构造一个结构化、网络化的知识体系，从而使得知识库更具有系统性和规范性。

2.2 系统应用

在博物馆数据收集过程中，需要将所收集到的文物信息进行汇总和储存，以方便对所有文物的名字、年代、介绍和图片等信息进行管理和查询。通过数据可视化技术，可以将收集到的文物信息生成知识图谱，以清晰、直观的方式展示文物之间的关系和联系，方便用户进行查看和使用。此外，对储存的数据信息需要进行管理和维护，确保文物信息的准确性和完整性，以便于对外提供相应的搜索服务和数据接口。

2.3 系统环境

博物馆建立完成后，可以通过掌上博物馆 App 对接该数据库，以基于存储的文物数据进行搜索和展示。同时，可以将知识问答子系统和海外文物知识子系统与博物馆数据库进行对接，为问答系统和海外文物知识子系统提供相关的文物信息和数据。这样，用户可以通过这些系统方便地浏览、搜索、了解和分享博物馆的文化和历史，以及与其他用户交流和互动。为了保证系统的顺利运作，需要对接口进行规范和安全性进行加强，以确保数据的完整性和安全性。

2.4 保密性和私密性

对于一些设计国家机密受法律保护的文物，其信息不予对外展示；若收藏文物的博物馆有相应的要求，文物的具体信息可以适当隐藏，以保障文物的安全。

2.5 帮助和问题报告

若用户在使用过程中遇到问题，可以通过评论或相关选项与创作团队取得联系，相关技术人员会根据问题及时做出调整与反馈。

3 软件使用指南

3.1 数据样例

部分数据样例

id	Name	Date	Type	Credit Line	Accession N	Dynasty	ALL_type	Material
1	Teapot	undated	Ceramic Teapot	Gift of Dr. and I	1995.98	undated	Porcelain	Porcelain
2	All Victorious Guanyin B	1147-117	Copper Sculpture	Museum purch	1941.83	Song Dyna	Metalwork	Copper
3	2000.5.5	2000	Woodblock print	Museum purch	2019.26	Modern Tir	Wood Art	Paper
4	A Divine Gift of One Hur	undated	Scroll Painting	Museum purch	2002.2	undated	Wood Art	Silk
5	A Hermitage	18th centu	Hanging Scroll Paint	Gift of Mr. S. M.	1964.71.16	Neoteric Ti	Painting	Paper
6	A Mandarin's Autumn H	19th centu	Silk Hat	Gift of Mrs. Har	1945.16.a	Neoteric Ti	(Null)	Silk
7	A Picture of Ancient Tre	1623	Scroll Painting	Gift of Mr. S. M.	1964.71.10	Ming Dyna	(Null)	Paper
8	A Smile in the Dream	1961	Calligraphy Album	Gift of Ambass	1995.1.14	Modern Tir	Calligraphy	Paper
9	Album of Eight Landsca	1822	Painting Folio	Gift of Dr. and I	1976.136.1-	Neoteric Ti	Artwork	Silk
10	Album of Eight Landsca	1822	Painting Folio	Gift of Dr. and I	1976.136.1	Neoteric Ti	Artwork	Silk
11	Album of Eight Landsca	1822	Painting Folio	Gift of Dr. and I	1976.136.2	Neoteric Ti	Artwork	Silk
12	Album of Eight Landsca	1822	Painting Folio	Gift of Dr. and I	1976.136.3	Neoteric Ti	Artwork	Silk
13	Album of Eight Landsca	1822	Painting Folio	Gift of Dr. and I	1976.136.4	Neoteric Ti	Artwork	Silk
14	Album of Eight Landsca	1822	Painting Folio	Gift of Dr. and I	1976.136.5	Neoteric Ti	Artwork	Silk
15	Album of Eight Landsca	1822	Painting Folio	Gift of Dr. and I	1976.136.6	Neoteric Ti	Artwork	Silk
16	Album of Eight Landsca	1822	Painting Folio	Gift of Dr. and I	1976.136.7	Neoteric Ti	Artwork	Silk
17	Album of Eight Landsca	1822	Painting Folio	Gift of Dr. and I	1976.136.8	Neoteric Ti	Artwork	Silk
18	Album of Ten Paintings	mid 18th (	Painting album	Gift of Dr. Regi	1975.49.1-1	Neoteric Ti	Artwork	leaf
19	Album of Ten Paintings	mid 18th (	Page of Album	Gift of Dr. Regi	1975.49.1	Neoteric Ti	Artwork	leaf
20	Album of Ten Paintings	mid 18th (	Page of Album	Gift of Dr. Regi	1975.49.10	Neoteric Ti	Artwork	leaf
21	Album of Ten Paintings	mid 18th (	Page of Album	Gift of Dr. Regi	1975.49.2	Neoteric Ti	Artwork	leaf
22	Album of Ten Paintings	mid 18th (	Page of Album	Gift of Dr. Regi	1975.49.3	Neoteric Ti	Artwork	leaf

数据库的结构设计

名	类型	长度	小数点	不是 null	虚拟	键	注释
id	varchar	255	0	☐	☐		
Name	text	0	0	☐	☐		
Date	text	0	0	☐	☐		
Artist/maker	text	0	0	☐	☐		
Type	text	0	0	☐	☐		
Credit Line	text	0	0	☐	☐		
Accession Number	text	0	0	☐	☐		
State/Province	text	0	0	☐	☐		
Dynasty	text	0	0	☐	☐		
ALL_type	text	0	0	☐	☐		
Material	text	0	0	☐	☐		

3.2 错误故障处理和恢复

3.2.1 消息

评论区的消息回复会带有提示。

3.2.2 处理

用户发现问题并将问题上传反馈后，我们会在第一时间进行 BUG 的修复与处理。

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。

3.3 软件维护

定期进行软件的压力测试，保障系统在人流量高峰期不会崩溃与瘫痪，同时定期核实数据库中的信息，发现错误时及时更正修改。