王宏, 朱海庭. 基于知识图谱的数字文物平台的研究与实现[J]. 智能计算机与应用,2024,14(10):170-175. DOI:10.20169/j.issn.2095-2163.241024

基于知识图谱的数字文物平台的研究与实现

王 宏, 朱海庭
(西安石油大学 计算机学院, 西安 710065)

摘 要: 在数字化时代背景下,文物的保护与传承面临着新的挑战与机遇,本文探讨并实现一种基于知识图谱的数字文物平台,以促进文物信息的整合、展示与利用。通过采集的丰富的文物数据,包括名称、简介、图片以及分类标签,以知识图谱的形式将这些数据进行组织与整合,存储于图数据库中,从而实现对文物数据的高效管理。其次,借助前端技术实现知识图谱的直观的可视化展示,用户可以通过交互式界面深入了解文物之间的关联、历史背景及分类情况。同时,基于知识图谱在平台中引入智能问答功能,为用户提供全面、准确的解答。本研究不仅为数字文物平台的构建提供了一种新的方法和技术支持,同时也为文物领域的数字化转型与应用提供了有益的参考与借鉴。

关键词: 数字文物; 图数据库; 知识图谱; 可视化; 智能问答
中图分类号: TP391.7 **文献标志码**: A **文章编号**: 2095-2163(2024)10-0170-06

Research and implementation of digital cultural relics platform based on knowledge graph

WANG Hong, ZHU Haiting
(School of Computer Science, Xi′an Shiyou University, Xi′an 710065, China)

Abstract: Under the background of digital age, the protection and inheritance of cultural relics face new challenges and opportunities. This article aims to explore and implement a digital cultural relics platform based on knowledge graphs to promote the integration, display, and utilization of cultural relics information. By collecting rich cultural relics data, including names, introductions, images, and classification labels, these data are organized and integrated in the form of a knowledge graph, stored in a graph database, thereby achieving efficient management of cultural relics data. Secondly, with the help of front-end technology, intuitive visualization of knowledge graphs can be achieved, allowing users to gain a deeper understanding of the relationships, historical backgrounds, and classification of cultural relics through an interactive interface. At the same time, intelligent question and answer functions are introduced into the platform based on knowledge graphs, providing users with comprehensive and accurate answers. This study not only provides a new method and technical support for the construction of digital cultural relics platforms, but also provides useful reference and inspiration for the digital transformation and application in the field of cultural relics.

Key words: digital cultural relics; graph database; knowledge graph; visualization; intelligent question answering

0 引 言

文物是人类文明的历史见证,承载着丰富的文化内涵和历史价值,通过构建文物展示系统,将网络与文物连接起来,使文物数据得以数字化、网络化则具有重要的现实意义[1]。数字化文物平台的建设将给文物的展示与传播带来诸多优势[2]。首先,可以实现文物信息的高度集中和整合,用户能够更加便捷地获取到文物的详细信息。其次,数字化展示也让文物跨越时间和空间的限制,实现广范围的共享与传播。然而,随着数字化时代的不断进步,为适应科技发展的新模式和新方法,文物的保护、传承与展示技术也亟需更新。传统的文物展示系统通常采用静态的图片陈列和文本数据的形式,这种方式虽然能够呈现文物的基本信息,但往往缺乏交互性和个性化体验,且难以应对信息化、智能化的需求。同时,传统的展览和研究手段也逐渐显得滞后。因此,构建一种高效、智能的数字文物平台迫在眉睫。

作者简介: 王 宏(1968-),男,副教授,硕士生导师,主要研究方向:人工智能应用,政务信息化。
通讯作者: 朱海庭(2000-),男,硕士研究生,主要研究方向:人工智能应用,大数据应用技术。Email:1627490229@qq.com

收稿日期: 2023-06-15

知识图谱可以清晰地呈现现实中实体的关系与属性[3]，本文致力于研究与实现一种基于知识图谱的数字文物平台，通过图形结构的形式存储文物实体及其之间的关系，为公众、研究者和文物管理者提供一个全新的、可交互的文物管理和展示平台[4]。通过从数据库中获取文物数据，并使用 Python 语言对文本数据进行清洗和修正，利用函数库将数据写入到图数据库中，构建起文物知识图谱[5]。基于这一知识图谱，并借助融合图形库的前端技术，实现知识图谱的图形可视化展示。此外，为进一步提升数字文物平台的智能化水平，通过使用算法和规则方法，为平台增添智能问答功能，使用户能够获得实时的问题解答[6]，从而更加深入地了解和通晓文物知识。这一系列改进将为数字文物平台的建设和发展提供新的思路和方法，促进文物管理与展览工作的深入开展。

1　相关技术介绍

1.1　知识图谱

知识图谱是一种基于图形结构的语义知识表示方式，旨在捕捉现实世界中实体之间的关系和属性[7]。由一系列实体（节点）及其之间的关系（边）组成，形成一个具有语义关联的实体网络，具体结构模型如图 1 所示。在知识图谱中，实体通常代表现实世界中的事物，如人、地点、组织、事件等，而关系则表示这些实体之间的联系，如位置、时间等[8]。在本系统中，实体节点表示的是具体的文物实体，关系是文物与标签分类、文物附属数据之间的关系。

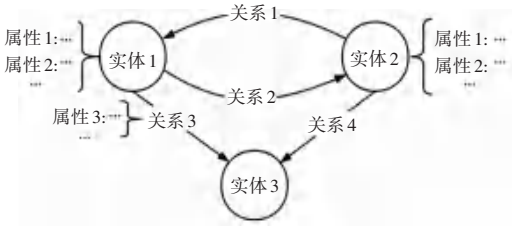

图 1　知识图谱结构模型图

Fig. 1　Knowledge graph structure model diagram

知识图谱的构建过程通常包括 3 个主要步骤：知识抽取、知识表示和知识推理[9]。其中，知识抽取是指从各种数据源中提取结构化的知识，可以通过自然语言处理、信息抽取等技术来实现。知识表示是指将抽取得到的知识转换成计算机可理解的形式，通常采用三元组（$N-R-N$）的形式表示出实体和关系[10]。知识推理是指基于已有的知识对新的知识进行推断和推理，从而丰富知识图谱的内容和结构。

知识图谱在各个领域都有着广泛的应用，如搜索引擎、推荐系统、智能问答等[11]。在数字文物领域，知识图谱可以帮助整合和展示文物信息，使得用户能够更加全面地了解文物之间的关联和历史背景。同时，知识图谱还可以为文物保护和研究提供支持，帮助研究人员发现隐藏在文物背后的规律和价值。因此，在知识图谱基础上建立起数字文物平台，能够更加有效地管理、展示人类的文化遗产，并推动文物研究工作的进一步发展。

1.2　可视化展示

ECharts 是一个由百度开发并维护的开源数据可视化库，专门用于在 Web 上创建丰富、交互式的图表和数据可视化界面。Echarts 基于 HTML5 Canvas 技术实现，具有强大的功能和灵活的扩展性，可以轻松地处理大规模数据集并实现各种复杂的可视化效果[12]。

ECharts 提供了丰富的图表类型，包括折线图、柱状图、饼图、散点图、雷达图、K 线图等等，同时支持堆叠、缩放、拖拽、动画等交互功能。同时，还提供了丰富的配置选项和 API 接口，使开发人员能够根据具体需求定制和调整图表的样式、布局和行为。

在基于知识图谱的数字文物平台中，利用 Echarts 等技术将文物知识图谱以生动、直观的图形化形式展现给用户。具体流程是将由图数据库中获取到的 node 节点和 links 关系的数据传入到前端配置项中，由此即可渲染出动态的文物知识图谱，图形化展示不仅使用户更容易理解文物之间的关联和分类情况，还为用户提供了一种全新的探索文物世界的方式。用户可以通过交互式界面深入了解不同文物之间的关系，探索各文物间的历史渊源、文化内涵以及艺术特点。同时，图谱的可视化也为文物研究者提供了新的研究思路和方法。研究人员可以通过图谱界面进行数据分析，探究文物背后的潜在规律和趋势。

此外，数字文物平台也为用户提供个性化的浏览体验，用户可以根据自己的兴趣和需求，自由地浏览和检索各类文物信息。技术方面除 Echarts 之外，也可借助其他前端技术和框架实现图谱可视化，如 D3. js、React 等，进一步提升用户的交互体验和页面性能。

1.3　Aho-Corasick 匹配算法

Aho-Corasick（AC）算法是一种高效的多模式匹配算法，用于在一个长文本中同时查找多个模式

串的出现位置[13]。该算法由 Aho 和 Corasick 提出，是基于确定性有限自动机(DFA)的改进算法。算法的核心思想是构建一个特殊的有限状态自动机(Finite State Automation，FSA)，该自动机在输入文本上进行模式串匹配。使用过程主要分为以下几个步骤：

(1)构建关键字树(Trie)：将所有待匹配的模式串构建成一个特殊的树结构，通常是一个前缀树[14](Trie)。该树的结构就使得算法能够高效地查找匹配，Trie 树结构图示例如图 2 所示。

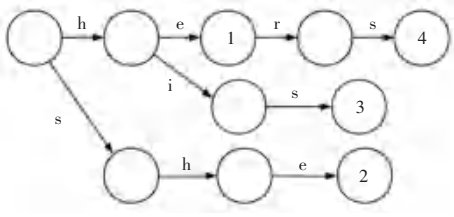

图 2 Trie 树结构图

Fig. 2 Trie tree structure diagram

(2)构建失败指针(Failure Pointer)[15]：对于每个节点，在构建 Trie 树的过程中，算法会设置一个指向其他节点的失败指针，该指针指向当前节点的“最长可匹配前缀”的后缀节点。目的是在匹配失败时，能够快速地跳转到下一个可能匹配的位置，从而提高匹配效率。

(3)匹配过程：将待匹配的文本从头到尾按字符依次输入到状态自动机中，根据当前字符和状态自动机的状态进行状态转移[16]。在每一步状态转移中，算法会根据当前字符和当前状态，找到下一个状态并进行跳转，直到输入文本被完全扫描完毕。当某个状态达到一个终止状态时，表示找到了一个匹配结果。

AC 算法的时间复杂度[17]主要取决于构建状态转移图的过程，即构建失败指针的过程。一旦状态转移图构建完成，匹配的时间复杂度是输入文本的长度加上所有模式串的长度之和。这种算法在实际应用得到了被广泛使用，尤其在字符串匹配和关键字检索等领域有着重要的应用价值。

2 数字文物平台设计

数字文物平台融入知识图谱的全流程，包括数据预处理、知识图谱构建、可视化展示和智能问答功能的设计。

首先需从 MySQL 数据库中提取文物数据，并转换为 CSV 格式，以便进一步规范数据格式。使用 Python 语言进行数据预处理，包括清洗和格式化等，以确保数据的准确性和一致性。系统采用 Neo4j 图数据库存储文物数据[18]，利用其图形结构的特性构建出知识图谱，该过程涉及图谱结构的设计、节点和关系的定义，以及索引的建立，用于支持后续的高效数据检索[19]。

通过前端技术和 Echarts 图形库，实现知识图谱的图形化展示。用户可以通过交互式界面，直观了解文物之间的联系，或基于搜索、过滤、解析等功能，获取所需的文物数据信息。

基于知识图谱的关联关系设计智能问答功能，利用 Aho-Corasick 算法和规则匹配方法对用户提出的问题进行解析与意图识别[20]，最终提供相对准确且即时的答案。智能问答功能将提高用户获取文物信息的效率和便捷性，同时提升平台的可交互性。此外，为更好满足使用的多语言需求，特引入多语言翻译服务，该服务能够将文物数据、展示结果等内容实时翻译为用户选择的语言类型，从而使得平台更加用户友好，并提供更为全面和便捷的使用体验。

平台整体框架图如图 3 所示，通过简化和优化流程，数字文物平台将为用户提供一个全新的、交互式的文物管理和展示平台。

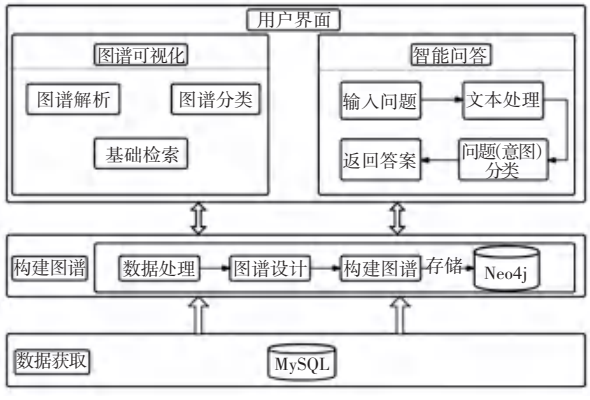

图 3 数字文物平台框架图

Fig. 3 Digital cultural relics platform framework diagram

3 功能实现与展示

3.1 知识图谱构建

构建知识图谱是一项复杂而系统化的工作，需经历多个严格的步骤，并需预先考虑后续问答模块和可视化界面的使用做出预先考虑，以保证数据的结构化和一致性。本文中所采用的数据经前期准备与处理皆已存入 MySQL 数据库中，该过程不做过多介绍，这里将主要讨论图谱的构建。

首先，需从 MySQL 数据库中获取文物的详细数

据，包括名称、简介、图片信息、标签等。这些数据经过提取后被转换成 CSV 格式文件，旨在作为下一步处理做备用。利用 Python 中的 pandas 库对这些 CSV 文件进行文本预处理，经过数据清洗、格式化以及异常值的处理等过程，从而确保数据的质量和可用性。

在知识图谱的构建过程中，重点考虑了文物的分类及其关联关系。设计文物类别总节点，如历史时代、材质构造、所属馆藏等，并建立了相应的子节点以及彼此之间的关系。此外，根据文物的信息，创建出其他数据节点，如历史来源、相关人物等，用来丰富和完善知识图谱的内容。

构建知识图谱的目的是将文物的信息以结构化的形式存储，便于用户根据图的关联详细探索文物的各个方面。因此，可以将文物的基本数据信息以属性的形式配置在文物实体节点中，以利于用户的后续查询和展示。数据库存储文物实体示例如下：


```
{
  "identity": 27,
  "labels": [
    "其他博物馆",
    "石刻艺术",
    "魏晋南北朝"
  ],
  "properties": {
    "name": "武士画像砖",
    "comid": "71106101012311",
    "description": "画像砖是用拍印和模印方法制成的图像砖。作为古代民间美术艺术的一枝奇花。",
    "labelname2": "其他博物馆",
    "proname": "武士画像砖",
    "pclassname": "71213141533580",
    "visitnum": "42",
    "fmtp": ……
  },
  "elementId": "27"
}
```


通过以上步骤，即得到一个全面、丰富的知识图谱，局部结构如图 4 所示。该图谱将为数字文物平台的功能实现奠定了坚实数据基础，可为用户提供更智能、全面的文物知识服务。

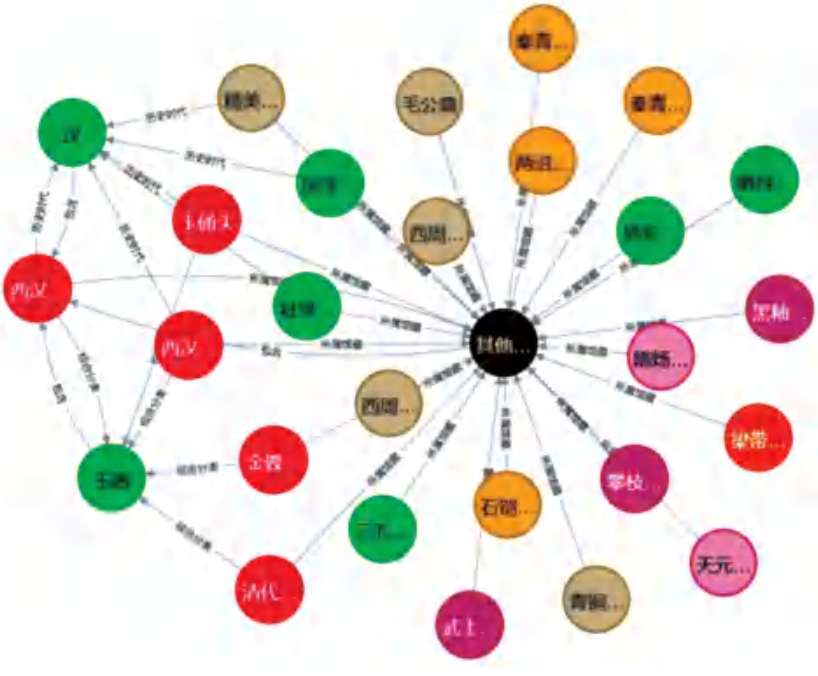

图 4　知识图谱局部结构图

Fig. 4　Knowledge graph local structure diagram

3.2　图谱可视化

图谱可视化主要包括展示图谱、实现分类以及解析实体之间的关系。图谱展示使用户能够直接地了解文物之间的关联和相关背景，从而深入探索文物世界。图谱分类功能能够对图谱中的文物实体根据标签进行分类，而解析实体之间的关系则帮助用户更清晰地认识到文物之间的联系和影响。

可视化界面根据用户需求提取图数据库中的节

点和边数据,代表图谱中的实体和关系。将数据传送至前端,利用 Echarts 图形库进行图形化渲染,并动态生成图谱界面,示例如图 5 所示。

图 5 图谱解析

Fig. 5 Graph analysis

3.3 算法与特征词汇表

AC 算法基于字典树的数据结构,以及有限状态机的概念。在数字文物领域,文物数据量通常较大,涵盖了丰富的文物信息和关键词。考虑到 AC 算法能够高效地同时匹配多个模式串,而无需遍历整个数据集,因此在处理大规模的文物数据时,选用 AC 算法是一种较为合适的选择。

应用 AC 算法首先需从文物数据中提取关键词,并将这些关键词构建成一个单词树。在单词树中,每个节点代表一个关键词的字符,从根节点到叶子节点的路径构成一个完整的检索项。由此单词树能够高效地存储和管理大量的文物关键词。

利用 AC 算法,对构建好的单词树进行处理,构建一个用于多模式匹配的自动机结构,后期使用中当用户提出问题时,将问题作为输入文本,利用构建好的自动机进行搜索匹配,同时匹配多个文物实体关键词,并记录匹配到的关键词及其位置。最终为实现智能问答功能提供关键信息。

同时,需构建出疑问特征词汇表为智能问答模块提供支持,该表旨在收集和整理常见的疑问词汇,以及与问题相关的其他关键词,用来帮助系统准确识别用户提问意图。后期也需不断优化和更新词汇表,提升系统的问题理解和解答能力,为用户提供更准确的答案。

3.4 智能问答

在智能问答功能模块中,当获取到用户输入的问句文本后,需对其进行分析处理,通过 AC 算法迅速捕获关键词,基于疑问特征词汇表得到问句中的疑问词,即可全面理解用户的查询目的。由此系统就能够在用户输入文本的基础上较为准确地识别出相关的主题和问题要素,为后续处理做好有效准备。

随后,根据疑问词的类型对问句进行分类,由此达到意图识别的效果,运用预定义的分类规则和语义模板,选择出相应的 Cypher 语句模板进行数据库查询,使系统能够更加智能地根据问句类型进行针对性搜索,从而提高问答系统的全面性和准确性。其中,解析文物图谱关系的 Cypher 语句示例如下:

```
" MATCH (startNode), (endNode) WHERE startNode.name =~'.*{}.*' AND endNode.name =~'.*{}.*'\
        WITH startNode, endNode\
        MATCH paths = allShortestPaths((startNode)-[*]->(endNode))\
        RETURN relationships(paths) AS r LIMIT 3"
```

最后,将从图数据库中获取到的数据信息经过数据格式转换后,进行答案封装,将结果以自然语言的形式返回给用户,确保用户得到的信息具有清晰的结构和良好的可读性。

通过以上流程的设计与实现,智能问答功能不仅能够为用户提供高效、准确的信息查询服务,同时也能极大提升平台的互动性与智能化水平。具体实现功能 Web 界面如图 6 所示。

图 6 智能问答功能界面图

Fig. 6 Intelligent Q&A function interface diagram

4 结束语

将知识图谱技术引入到数字文物平台后,文物数据得以高效管理和展示,用户能够更便捷地获取详细的文物信息。同时,使得文物之间的关联和分

类更直观，为用户提供了探索文物世界的全新方式。此外，智能问答功能的引入进一步提升了平台的智能化水平，用户可以实时获取问题解答，能够加深对文物知识的感悟和理解。未来，随着技术的加速发展和用户需求的不断变化，数字文物平台在知识图谱的基础上必会有更广阔的发展前景。

参考文献

[1] 夏杰长，李銮淏. 数字技术赋能文化强国建设的作用机制和优化路径[J]. 中国流通经济，2023，37(11)：3-11.

[2] 张建国. 数字博物馆对文物保护与全球化传播的保障策略研究[J]. 情报科学，2022，40(2)：59-64.

[3] 刘学锋. 基于数字人文背景下档案知识图谱的构建路径[J]. 黑龙江档案，2023(5)：157-159.

[4] 王春法. 关于智慧博物馆建设的若干思考[J]. 博物馆管理，2020(3)：4-15.

[5] 李凌霄. 面向频谱管理的知识图谱构建及应用[D]. 北京：北京邮电大学，2023.

[6] 刘怡彤，张静，姜润发. 基于 NLP 的图书馆智能问答系统研究[J]. 信息与电脑(理论版)，2024，36(1)：117-120.

[7] 秦川，祝恒书，庄福振，等. 基于知识图谱的推荐系统研究综述[J]. 中国科学：信息科学，2020，50(7)：937-956.

[8] 徐增林，盛泳潘，贺丽荣，等. 知识图谱技术综述[J]. 电子科技大学学报，2016，45(4)：589-606.

[9] 杨玉基，许斌，胡家威，等. 一种准确而高效的领域知识图谱构建方法[J]. 软件学报，2018，29(10)：2931-2947.

[10] 田玲，张谨川，张晋豪，等. 知识图谱综述—表示、构建、推理与知识超图理论[J]. 计算机应用，2021，41(8)：2161-2186.

[11] 武毓琦. 基于知识图谱的学习者个性化推荐方法研究[D]. 广州：广东技术师范大学，2023.

[12] 崔蓬. ECharts 在数据可视化中的应用[J]. 软件工程，2019，22(6)：42-46.

[13] 姜海洋，李雪菲，杨晔. 基于距离比较的 AC 自动机并行匹配算法[J]. 电子与信息学报，2022，44(2)：581-590.

[14] 陈永杰，吾守尔·斯拉木，于清. 一种基于 Aho-Corasick 算法改进的多模式匹配算法[J]. 现代电子技术，2019，42(4)：89-93.

[15] 魏先燕，卢加奇，冯燕茹，等. 基于云服务的恶意内容检测方法研究[J]. 现代信息科技，2023，7(12)：155-157，161.

[16] 刘碧纯. 基于 MPI 和串匹配算法的关键词查重并行算法的研究[D]. 哈尔滨：哈尔滨理工大学，2022.

[17] 杨武，方滨兴，云晓春，等. 入侵检测系统中高效模式匹配算法的研究[J]. 计算机工程，2004，30(13)：92-94.

[18] 田梦晖，陈明，席晓桃. 融合 Albert 模型的珍稀濒危植物知识图谱的构建[J]. 湖南农业大学学报(自然科学版)，2023，49(5)：616-623.

[19] 崔鸣石，邬雪阳，朱宏伟，等. 基于知识图谱的电力通信设备故障智能诊断方法[J]. 科技和产业，2024，24(5)：212-221.

[20] 魏晓玲，蔡敏. 基于知识图谱的创新创业智能问答系统[J]. 电脑编程技巧与维护，2023(1)：101-103，153.