计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(9):82−86 [doi: 10.15888/j.cnki.csa.005924]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于知识图谱的香山文化信息组织与检索系统①

陈威宇, 姜　赢, 罗盛亨, 黄嘉文, 吴昊驰

(北京师范大学珠海分校 管理学院, 珠海 519087)

摘　要: 香山文化的内涵包含多种元素, 内容极其丰富, 但是相关研究孤立分散主要原因在于缺乏系统性的信息组织. 在概述知识图谱信息组织应用基础之上, 提出基于知识图谱的香山文化信息组织思路. 利用骨架法构建了基于本体的香山文化知识图谱, 建立了香山文化分散研究要点之间的关联. 基于知识图谱的香山文化检索系统通过可视化方式展现香山文化中复杂的知识点和知识关系, 勾勒出香山文化总体轮廓. 基于知识图谱的信息组织在处理领域复杂关系的分析与挖掘上有一定优势, 知识图谱检索系统可视化模式丰富多彩、特色鲜明.

关键词: 香山文化; 知识图谱; 信息组织; 检索系统

引用格式: 陈威宇,姜赢,罗盛亨,黄嘉文,吴昊驰.基于知识图谱的香山文化信息组织与检索系统.计算机系统应用,2017,26(9):82–86. http://www.c-s-a.org.cn/1003-3254/5924.html

Xiangshan Culture Information Organization and Retrieval System Based on Knowledge Graph

CHEN Wei-Yu, JIANG Ying, LUO Sheng-Heng, HUANG Jia-Wen, WU Hao-Chi

(School of Management, Beijing Normal University, Zhuhai 519087, China)

Abstract: Xiangshan Culture has multiple elements and rich contents, but the isolation and the scattering of each research is due to the lack of systematic information organization. After summarizing the applications of knowledge graph information organization, this paper provides a method of Xiangshan Culture information organization based on knowledge graph. By using the Bone-method, the knowledge graph of Xiangshan Culture has been built based on ontology, with the connections among the scattered research topics of Xiangshan Culture. The knowledge graph-based Xiangshan Culture retrieval system visualizes complex knowledge instances and knowledge relations, in order to sketch the overall contours of Xiangshan Culture. The information organization based on knowledge graph has the advantage of analyzing and digging complex domain relations, while the knowledge graph retrieval system has vivid visualization models and distinctive features.

Key words: Xiangshan Culture; knowledge graph; information organization; retrieval system

香山文化在地缘上是指包括今天的中山、珠海、澳门在内的地域文化. 它在本质上集中体现了岭南文化中粤、闽、客三大民系的文化特征, 是中原文化、土著文化、西洋文化、南洋文化相互碰撞和不断融合的产物, 是相对岭南文化而言的子文化, 是岭南文化的重要组成部分[1]. 2006 年首发的《香山文化——历史投影与现实镜像》一书, 率先提出了香山文化这一概念[2], 社会各界尤其是学术界、新闻界对此给予广泛关注. 同年由广东省社会科学界联合会主办“香山文化学术研讨会”[3], 86 篇会议论文涉及香山文化的基本概念、本质特征、演变轨迹、历史名人、香山文化的传承与创新、香山文化研究的理论与方法等方面.

2006 年香山文化概念提出半年即“蹿红”, 但随后迅速降温, 近几年相关研究越来越少, 犹如昙花一现.

① 基金项目: 文化部科技创新项目(201505); 广东省自然科学基金(2016A030313386); 广东大学生科技创新培育专项资金(pdjh2017a0898)
收稿时间: 2016-12-19; 采用时间: 2017-01-09

可见, 香山文化“立得住”是做到了, 但“推得开”有一定困难, 更是没有达到“影响大”的层次. 究其原因, 主要在于香山文化研究内容分散, 缺乏系统性. 香山文化的内涵包含多种元素, 内容极其丰富, 大多数学者从各自研究领域出发, 针对香山文化的某一个方面进行研究[4](例如: 香山民俗、香山方言、香山买办、香山华侨、香山商业、香山文化), 研究内容较为分散, 难以反映香山文化全貌. 因此, 需要通过系统性的研究, 利用信息组织技术建立分散研究要点之间的关联, 勾勒香山文化总体轮廓, 归纳总结香山文化更加全面而宏观的文化精神.

另外, 研究香山文化的学者大多数是历史、社会等人文学科领域的专家和教授, 研究方法仅限于传统文献调查、实地调研, 案例实证分析等等[5,6]. 本文提出发挥交叉学科的优势, 利用数学、信息科学成熟的技术(例如: 数据挖掘、信息组织、知识组织、知识推理等)对香山文化进行更为量化的分析和研究, 揭示香山文化现象与文化本质之间的更深层次的因果逻辑, 为当今大香山经济圈的文化发展提供借鉴.

1 国内外研究现状

2010 年初, 以 Google 公司为代表的研究机构提出知识图谱(Knowledge Graph)的概念与实现框架[7]. 知识图谱以本体(Ontology)技术为核心[8], 通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法结合, 并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论. 知识谱图特别适合于解决内容关系复杂领域的知识管理问题, 在国内外医疗卫生、电子商务、生物化学、国防军事、人文历史等各个领域将有广泛的应用. 其中, 知识图谱在国内外历史文化遗产保护的典型应用案例较多. 例如, CultureSampo[9](芬兰历史知识图谱)是芬兰政府 2010 年建设的文化公共发布门户网站, 它利用本体映射技术和本体推理技术, 将来自芬兰 20 个博物馆、图书馆、档案馆中的素材整合, 建立成芬兰历史知识图谱. 目前总共容纳了 128, 714 件芬兰文化遗产物件, 包括博物馆藏品、历史照片、地图、油画、诗歌、古籍、民歌等, 还包括 276, 681 个历史事件、人物、地点、时间等抽象文化概念知识. 它提供基于知识图谱的查询服务: 在文化遗产物件及抽象文化概念知识之间推荐和跳转, 查询历史人物之间的知识关联, 查询用户地理位置周边的文化遗产物件, 以时间轴为线索浏览芬兰重要历史时间及相关文化遗产物件. 2012 年武汉大学信息资源研究中心与中华书局合作项目“中华史籍分析系统”[10], 对二十四史中的人物、时间、地点实体进行了全面标注. 该系统自建知识图谱记录总共 268491 条. 知识库构建知识类 122 个、对象属性 32 个、数据属性 28 个, 推理属性 15 个和实例 179503 个, 时空分析人物 308 个, 地图地点标注 12736 个. Google、Facebook 等国外知名互联网公司是知识图谱的倡导者. Google 公司已建立了 5 亿个对象, 35 亿个事实和关系, 足以证明知识图谱技术的可行性. 随后, 国内百度、搜狗以及复旦大学 GDM 实验室相继推出了其中文知识图谱, 可见知识图谱在中文领域应用的可行性[11].

在此背景下, 本文提出以挖掘、研究、弘扬香山文化为主旨, 利用现代信息组织技术手段构建香山文化知识图谱及检索系统, 不仅仅是对历史进行系统全面的梳理和对历史文化遗产保护, 更重要的是力求通过弘扬和传承, 对大香山经济圈的经济和社会发展起到促进作用.

2 基于知识图谱的香山文化信息组织

2.1 香山文化知识图谱的主要内容

知识图谱的理论模型主要包含知识分类、知识点, 知识属性、知识属性值、知识点之间关联. 领域中经常出现的词汇, 这些词汇就是知识点. 由于知识点很多, 需要分门别类组织一下, 知识分类可以有多层, 最终形成一个树形结构. 知识点表达具体个体的概念; 知识分类表达抽象分类的概念. 为了更深入细致的描述知识点, 可以为知识点添加知识属性. 知识属性是描述知识点的某个方面. 什么样的知识点具有什么样的知识属性, 是根据知识点所归属的知识分类来确定的. 可以用知识属性来描述某个知识点, 并将某个知识属性值赋予这个知识属性, 形成一个完整的对知识点的描述. 知识属性和知识属性值都是用来描述知识点: 知识属性与某个知识分类关联, 可以被归属于该知识分类的知识点所共享重用; 知识属性值与某个知识点关联, 只能和某个知识属性一起组合起来描述一个知识点. 知识属性和知识属性值是对知识点的内部特征的描述, 而知识关系也是用来描述知识点的, 只不过它所描述的

是知识点对外关系/关联(与谁关联以及如何关联). 知识属性和知识关系有类似的特征: 什么样的知识点之间具有什么样的知识关系, 是根据知识点所归属的知识分类来确定的. 知识关系所关联的对象就是知识点. 所以知识关系值, 也就是“宾语”(Object)本身就是知识点. 即知识关系关联了知识点与知识点. 而知识属性可以理解为知识点, 与字符串、数值等之间的关联.

将所收集的香山文化知识点进行分类, 得到 14 个知识分类: 历史事件、地点(行政区域)、学校、文化遗存、香山人物、组织机构、文学、艺术、时间、称号、职务、饮食文化、香山方言、香山民俗. 经过这样的划分, 知识图谱的架构体系以及脉络十分清晰, 从多维度出发, 而且分类细腻, 基本涵盖了香山文化的所有内容. 这有助于我们能更清晰地研究香山文化. 具体来说, 山文化知识图谱的框架设计包括以下知识分类:

历史事件: 军事事件、政治事件、教育事件、文化事件、社会事件、科技事件、经济事件、自然事件.

组织机构: 军事组织、国际组织、工商机构、政府机构、文化教育机构(这个又可分为宗教组织和教育单位)、社会组织、社会群体、经济组织(个体商店和公司企业).

香山人物: 世纪伟人、乡贤俊彦、买办家族(唐、徐、莫、郑四大家族)、从商人士、军政要人、华侨华人、思想先驱、文化名家、留学人士、航空翘楚、英烈志士、香山居民.

时间按照具体的时间点和时间段添加子类, 以具体的年份作为知识点. 地点以行政区域进行子类划分. 艺术、文学以作品类型添加子类. 学校以在读和毕业分开. 香山方言以语系的不同进行划分. 香山民俗按照习惯活动、礼节、节日以及艺术进行子类划分. 饮食文化以烹饪方式、饮食方式以及具体的美食相关进行子类划分.

如图 1 所示, 香山文化的买办文化板块中, 近代中国著名的买办、实业家徐氏家族在中国早期工业化的过程中起到的积极作用: 1872 年李鸿章委派唐廷枢为总办, 徐润、盛宣怀为会办, 改组轮船招商总局, 徐润统管财务账目、人事大权; 1877 年招商总局吞并了当时轮船运输业的老大——美国旗昌轮船, 增加了码头和船只, 扩大了经营, 成为能与太古轮船公司抗衡的唯一对手, 后来还不断投资大型企业, 包括投资张之洞在汉阳创办的湖北铁厂等企业.

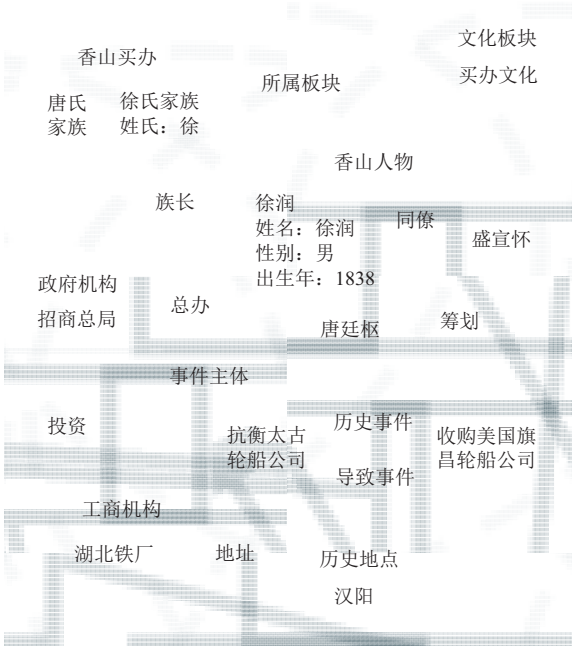

图 1 香山文化知识图谱信息组织示例图

2.2 香山文化知识图谱构建的基本思路

首先对香山文化资料收集与梳理: 香山文化文献资料特别是晚清到民国间香山文献数量之宏大, 内容之丰富, 在广东省名列前茅, 这是大香山经济圈最大的文化资源, 是香山文化的根基, 也是了解和诠释香山文化的重要依据, 只有摸清楚了香山文献的家底, 对其有了更深入更真切的了解, 才能够对香山文化的含蕴有更准确、清晰的把握和界定, 为香山文化知识图谱的构建与挖掘打下基础. 作为原始文献资料收集的补充, 利用田野考古的形式, 深入基层、深入民间调研, 包括中山、珠海诸多村庄, 深入澳门、东莞、顺德、江门等地, 寻访名人故居、名人坟墓, 访问知情人士, 记录口述史料和真情实感, 收集大量第一手资料.

接下来, 从收集到的香山文化资料挖掘出香山文化知识分类、知识点、知识属性和知识属性值, 最终建立香山文化知识图谱. 香山文化的常见的词汇需要按照这种知识模型组织起来, 建立起香山文化知识图谱.

最后, 本项目香山文化知识图谱服务平台采取 MVC 模式构建了 B/S 应用系统. 具体来说, 后台使用 Jena TDB 数据库作为知识图谱持久化的工具, 利用 Jena 的 RDF API 对知识图谱进行读写操作. 笔者研发的中间件嵌入到 Struts 框架作为业务逻辑层, 提供知识图谱索引、检索、提取和推理等核心功能 API. 特

别在检索部分还使用 Lucene 对知识分类、知识点、知识属性、知识属性值和知识关系的 LocalName 进行索引, 能够实现模糊检索. 系统前台使用 D3.js 工具将三元组转换成结点和边, 最终使用 Javascript 构建出知识图谱检索结果的 Graph 图.

香山文化知识图谱的构建, 是知识图谱构建人员和文化领域专家共同努力的成果. 笔者邀请了中山大学的一位历史系教授以及北京师范大学珠海分校的一位研究历史文化的教师参与我们的香山文化知识图谱构建. 他们对整合好的香山文化知识图谱原始资料库进行人工筛选和补充, 最终建立知识图谱中所有知识点. 在这个过程中, 他们细致而专业的历史文化理论知识使我们能够顺利地完成知识图谱本体库的构建.

3 香山文化知识图谱应用系统

3.1 香山文化知识图谱构建系统(后台系统)

Protégé[12]软件是斯坦福大学基于 Java 语言开发的本体编辑和知识获取软件, 或者说是本体开发工具, 也是基于知识的编辑器, 属于开放源代码软件[13]. 它提供了大量的知识模型与动作, 可以创建并操作各种表现形式的本体. Protégé已成为目前使用最广泛的本体论编辑器之一, 是一套用于对本体知识进行描述、表达和推理的软件. 它拥有一个灵活的架构, 支持插件开发, 并且提供了一套 Java API 供编程人员使用[14]. 笔者可以利用这个开源软件, 实现香山文化知识图谱的构建. 基于 protégé的香山文化本体库的构建实现如图 2 所示. 最终, 香山文化知识图谱构建有 110 个知识分类(class), 其中包括 4 个父类(superclass)和 106 个子类(subclass), 以及 2482 个知识点(individual)和 67 条关系属性(property), 经过统计香山文化本体已经包含了 3740 条本体数据记录.

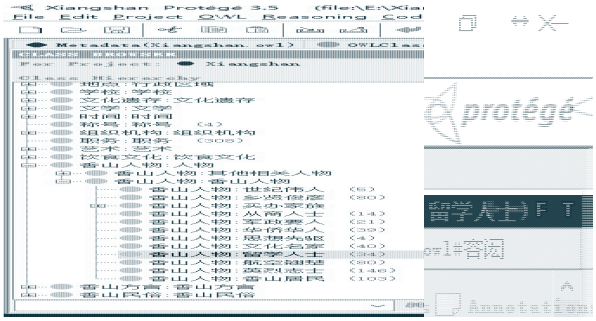

图 2 基于 Protégé的香山文化知识图谱构建

3.2 基于知识图谱的香山文化检索系统(前台系统)

在设计并构建好香山文化知识图谱之后, 笔者研发了基于知识图谱的香山文化检索系统, 并使用了 D3.js 工具来实现知识图谱的可视化效果. 在研发过程中, 笔者利用具名图对香山文化知识图谱中由 RDF 三元组描述的资源进行四元组拓展, 为它们加入时间维度描述, 即变成“<subject>-< predicate >-<object>-<time graph>”. 笔者定义两个新的 Graph: 一个 Graph 是用来存放在古代乃至近现代香山文化发生或者存在的人、物、事等, 另一个 Graph 是用来存放现代香山依然存在并且正在发展的人、物、事等. 而正在 RDF 中, 默认存在的 Default Graph, 笔者就用来存放一些在时间维度上不属于古也不属于今的客观事实. 通过这样的资源再搭建, 我们在香山文化知识图谱检索系统上建起了一条时间线, 将香山文化的时间逻辑环扣在了一起.

例如, 通过在图谱中检索“马应彪”, 会出现与马应彪相关的知识点、知识属性和知识关系, 如图 3 所示.

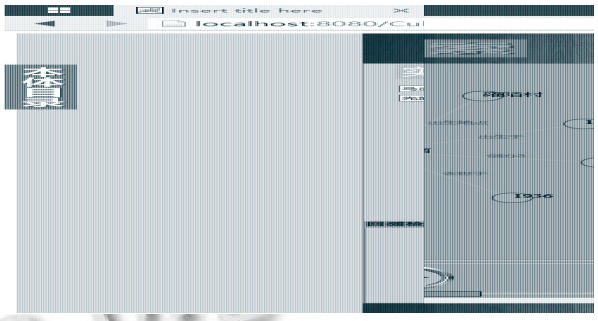

图 3 基于知识图谱的香山文化检索系统

如图 4 所示, 以“马应彪于 1900 年创办了先施公司”、“马景华在 2014 年经营先施公司”、“马景煊就任现任先施公司董事长”、“马景煊与马景华是堂兄弟关系”这四个为例, 来展示知识图谱的古今演化效果.

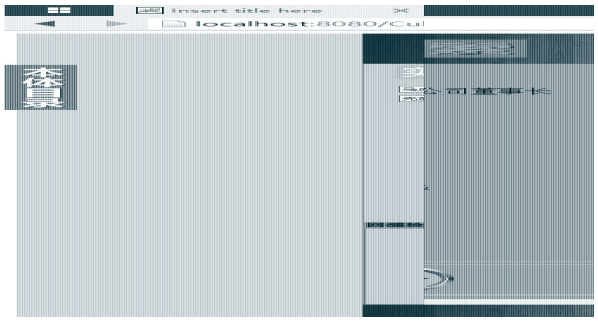

图 4 香山文化知识图谱时空演化“今”效果图

4 总结

本文提出利用信息组织的成熟技术对香山文化进行定性和定量相结合的分析和研究, 构建了一个涵盖香山文化各个方面内容的香山文化知识图谱, 力求解决香山文化相关的历史人物、历史事件、文化遗产等复杂关系的描述、组织、检索和等知识管理技术问题, 丰富了文化遗产保护与开发的技术手段.

参考文献

1 胡波. 香山文化的现代诠释. 学术研究, 2006, (6): 115–121.

2 赵立彬. 文化自觉: 从文化学视角看香山文化研究的意义——兼评《香山文化——历史投影与现实镜像》. 广东社会科学, 2007, (2): 122–128.

3 王杰, 胡波. 香山文化显芳华——香山文化学术研讨会综述. 学术研究, 2007, (10): 151–155. [doi: 10.3969/j.issn.1000-7326.2007.10.026]

4 张文平. 浅析香山的商业文化与买办文化. 文史博览(理论), 2011, (7): 25–26.

5 尹绪忠. 论香山文化对近代中国政治发展的启蒙作用. 学术研究, 2007, (5): 144–149.

6 李芳清. 香山文化: 中国近代文化的奠基石. 广东社会科学, 2007, (6): 124–129.

7 Singhal A. Introducing the knowledge graph: Things, not strings. https://mondaybynoon.com/introducing-the-knowledge-graph-things-not-strings/. [2016-03-20].

8 Grau B, Horrocks I, Motik B, *et al*. OWL 2: The next step for OWL. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(4): 309–322. [doi: 10.1016/j.websem.2008.05.001]

9 Mäkelä E, Hyvönen E, Ruotsalo T. How to deal with massively heterogeneous cultural heritage data: Lessons learned in CultureSampo. Semantic Web, 2012, 3(1): 85–109.

10 董慧, 徐雷, 王菲, 等. 基于语义系统的中华史籍分析研究. 图书馆理论与实践, 2015, (4): 1–5, 46.

11 百度下一代搜索引擎雏形曝光 应用知识图谱技术. 电脑编程技巧与维护, 2013, (19): 4.

12 Stanford University School of Medicine. What is protégé. http://protege.stanford.edu/overview. [2012-12-01].

13 邓仲华, 黄鑫, 陆颖隽, 等. 论中文古籍版本本体库的构建. 图书情报知识, 2014, (4): 80–87, 93.

14 何来坤, 缪健美, 刘礼芳, 等. 基于 Ontology 与 Jena 的研究综述. 杭州师范大学学报(自然科学版), 2013, 12(5): 467–473.