文章编号：2095-2163(2022)05-0066-04 中图分类号：TP311.5 文献标志码：A

旅游景点知识图谱的构建与应用

原 越，范家豪，寇 哲，安超凡
（北京信息科技大学 计算机学院，北京 100101）

摘 要：本文针对现有搜索引擎提供的旅游景点信息缺乏关联度和完整性，不利于用户挖掘有效信息的特点，通过设计旅游景点本体、针对爬取语料的处理构建旅游景点知识图谱。在此基础上设计并实现了一个 B/S 架构的旅游景点应用系统。

关键词：知识图谱构建；本体构建；旅游景点

Construction and application of knowledge graph of tourist attractions

YUAN Yue, FAN Jiahao, KOU Zhe, AN Chaofan
(Computer School, Beijing Information Science and Technology University, Beijing 100101, China)

【Abstract】 Aiming at the lack of relevance and completeness of the tourist attractions information provided by the existing search engines, which is not conducive to the user's digging of effective information, this paper constructs the tourist attractions knowledge map by designing the tourist attractions ontology and processing the crawling corpus. On this basis, a B/S architecture tourist attractions application system is designed and implemented. This research could help users search for information more efficiently and quickly, and understand scenic spots intuitively and concisely, thereby promoting the development of the tourism industry. The fruits also are helpful for further research on the construction of domain knowledge maps.

【Key words】 knowledge graph construction; ontology construction; tourist attractions

0 引 言

得益于信息技术的发展，线上景点数据日趋丰富。目前在各个领域都存在着知识图谱技术的广泛应用。针对旅游领域来说，现有的基于网页进行搜索的方式能够提供的旅游景点信息缺乏关联度和完整性，不利于用户获取有效信息。而知识图谱作为一种关联知识组织技术，能够描述现实世界中存在的各种实体和属性，以及实体之间的关系[1]。同时，知识图谱能够帮助计算机更好地理解人类语言模式，从而更加智能化地对用户所需要的各类信息进行反馈[2]。

在知识图谱构建技术方面，刘峤等人[3]介绍了自底向上的知识图谱构建技术。在知识获取方面，姚萍等人[4]介绍了知识获取的过程。在知识图谱的应用方面，邵嘉进等人[5]将知识图谱用于旅游领域，构建了每个用户的画像，并为每个用户推荐旅游景点。

与传统的富文本形式相比，知识图谱更注重强调对实体的覆盖，其包含的语义丰富，语义关系的建模具有多样化的特点。例如针对景点、酒店、餐饮这 3 个实体，可以引申出相同地理位置的景点、相同等级的景点以及景点附近相邻的酒店和餐饮三个关系，对比分散性查询信息可以提供极大的便利。因此，本文通过对景点数据的高效整理来构建旅游景点知识图谱，且在此基础上设计并实现了基于知识图谱的旅游景点应用系统。

1 知识图谱的构建

1.1 知识图谱的架构

知识图谱在逻辑上可分为模式层与数据层两个层次：模式层在数据层之上，是知识图谱的核心，通常采用本体库来管理知识图谱的模式层[6]。两者之间的联系以及区别通过如下例子进行说明。

(1)模式层：实体-属性-属性值

(2)数据层：故宫-面积-72 万平方公里

(3)模式层：实体-关系-实体

(4)数据层：故宫-位于-北京

这里，对构建流程可做探讨分述如下。

(1)原始材料的收集。根据互联网的信息进行爬取。

(2)本体构建。是对获取的知识进行加工，模式层的建立。

(3)信息抽取。需要将内容抽取成三元组的形

基金项目：北京信息科技大学 2021 年大学生创新创业训练计划(51002110805)。

作者简介：原 越(2001-)，女，本科生，主要研究方向：知识图谱。

收稿日期：2021-11-22

式。

(4)知识存储。将最终的内容存储到图数据库中,利用自身的语句进行增删改查。

1.2　知识图谱的构建流程

目前,多用 2 种构建知识图谱的方法。一是自顶向下:从模式层开始构建,再往里面添加实例,直至构建的完成。二是自底向上:从数据层开始构建,再向上抽象出来概念和关系,直至构建完成。

本文选择 2 种方法相融合进行旅游景点知识图谱的构建。首先定义并构建本体,建立模式层;根据模式层抽取出景点、旅店、餐馆的概念、属性、属性值以及各自的概念、属性、属性值之间的关系;最后将所有知识存入 Neo4j 图数据库。构建流程如图 1 所示。

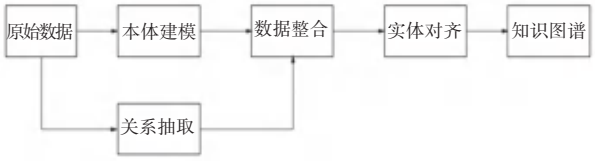

图 1　知识图谱构建流程图

Fig. 1　Flow chart of knowledge graph construction

2　旅游知识的本体设计

景点的知识的本体构建选用经典的七步法,并且根据实际情况简化不必要的步骤。主要分为 5 个步骤:确定本领域的范畴 、获取爬取的信息、旅游领域标签确定、抽取本体概念、本体建模。本体构建流程如图 2 所示。

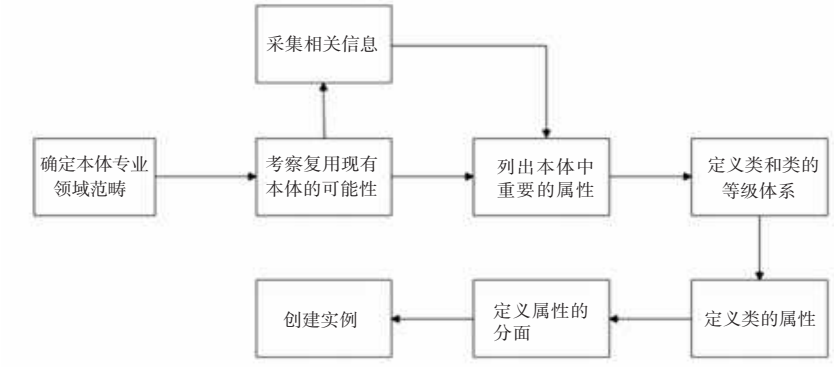

图 2　本体构建流程图

Fig. 2　Flow chart of ontology construction

本体建模的过程:使用 Protégé 软件利用自上而下法进行本体构建、旅游景点知识图谱的构建。对此拟展开研究论述如下。

(1)确定本领域的范畴。不同的领域本体的构建内容和方式是不同的,研究旨在确定本体领域研究范围重要的因素。本文着重研究的是旅游领域知识图谱的构建,所以重点在于以旅游景点为中心的本体构建[7]。

(2)获取有关信息。通过爬取得到的数据作为景点信息的来源,在其中获取旅游景点、旅店、餐馆的相关的信息。

(3)旅游领域标签的确定。根据爬取的信息,确定初步的旅游领域的标签,如:旅游攻略、酒店、餐馆、景区风景、等级等等。

(4)抽取本体概念。通过筛选抽取出来本体的概念有地区、等级两个类。

研究得到的本体类层级如图 3 所示。根据旅游景点的等级和采集到的数据,可以将景点分为 AAAA 和 AAAAA 级景区。地区可以根据中国的省份进行划分,分为:黑龙江省、四川省、吉林省等等。

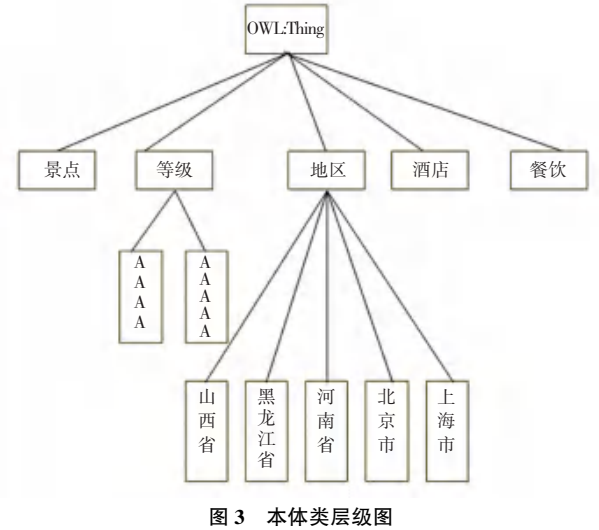

图 3　本体类层级图

Fig. 3　Hierarchical diagram of ontology classes

(5)本体建模。本体建模的具体实现如图 4 所示。

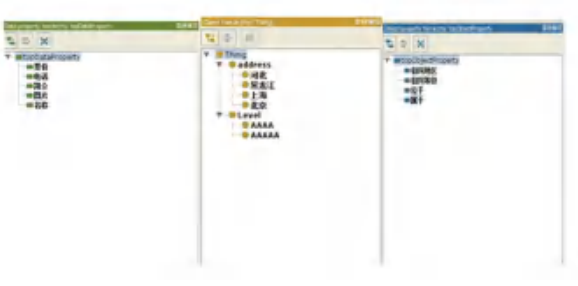

图 4 本体建模图

Fig. 4 Ontology modeling diagram

(6)部分结果展示。见图 5。

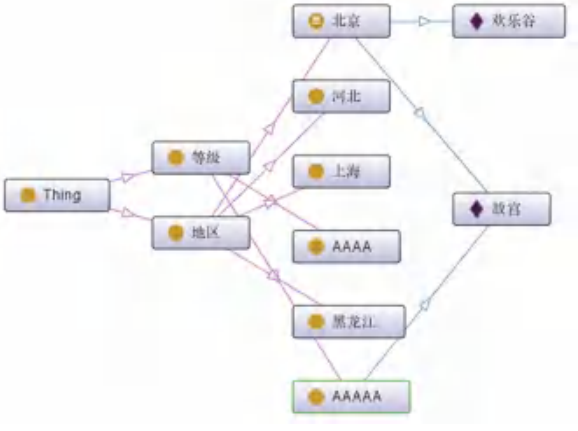

图 5 部分结果展示

Fig. 5 Part of the results show

3 基于旅游知识图谱的系统设计与实现

3.1 系统功能设计

相对于文本形式的表达,将各景点以及景点周边设施的信息以图的形式展现可以帮助用户在尽可能短的时间内多角度、全方位地搜索到自己感兴趣的目标,功能需求如下:

(1)多样化输入法格式。因为每一个用户的搜索形式不同,有的用户可能只输入关键词进行搜索,而有的用户会输入整句进行搜索,所以系统要尽可能满足不同用户的输入法格式,无论是输入关键词、还是整句,都能搜索到相应内容。

(2)多景点图谱可视化展示。以图的形式将各景点的信息展示出来,以导航形式将信息串联起来,如此用户使用起来将更加方便。

(3)当用户点击某个景点节点时,与其处于相同等级或者相同地理位置的景点就会被列举出来,这时用户可以从中选择自己感兴趣的景点。若要获得某一景点的周边服务,例如酒店、餐饮时,只需要点击该景点,就会实现跳转,获得更加详细的信息。

3.2 系统关键技术

系统采用 B/S 架构,对于设计实现中用到的关键技术,这里将给出剖析表述如下。

(1)数据爬取与预处理。通过爬取 https://www.ctrip.com 和 https://www.qunar.com 作为信息的来源。在爬取方法上,采用分布式爬虫的方法,能够提高爬取数据的效率;在搜索策略上,采用广度优先的网页搜索策略;在分析方法上,采用基于文本的网页分析算法,很大程度上借用了文本检索的技术,利用网页的标签,快速有效地对网页信息进行分类和聚类;在爬虫技术上,分为 URL 读取和存储、页面元素分析、数据存储三个模块,最终将其存储到 csv 文件中。

(2)知识图谱存储。利用将 Phantom JS 和 Selenium 结合搭建爬虫框架,通过正则化表达式导入存入 csv 文件中,利用图数据库 Neo4j 自身的 import 语句导入 Neo4j 中。Neo4j 数据库中的知识图谱如图 6 所示。知识图谱中,实体包括:景点、餐馆、酒店。景点的属性有图片、名称、电话、简介、票价、等级、地址;酒店属性包括酒店的名称、类型、酒店链接、评分、地址、电话、简介、好评率;餐馆包括餐馆的名称、餐馆菜系、地址、特色美食、人均费用、评分这几个属性。关系则包括:同地区、相同等级、近邻。

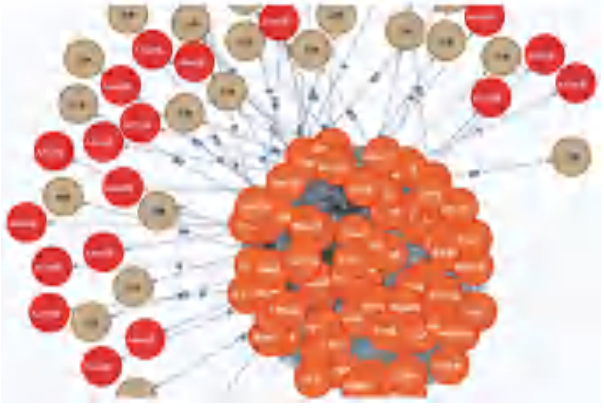

图 6 Neo4j 数据库中知识图谱

Fig. 6 Knowledge graph in Neo4j database

(3)信息检索。将前端网页输入的关键词请求传送到后台,调用相关的语句,匹配关键词,返回信息,展示给用户。

(4)搜索结果可视化。本文采用 roc-echarts 实现知识图谱的网页可视化,后台通过 Cypher 语句完成信息检索。

3.3 系统效果展示

景点搜索页面是主要应用知识图谱的页面,页面主要分为四大块,分别为:左侧、上面、中间、下面。旅游景点网站主要展示这个景点的信息,并将信息分成 4 部分,分别为:基本信息、简介、小贴士、知识

图谱的可视化。从知识图谱中可以清晰地看到，景点的信息、和本景点相同地点的景点、和本景点相同等级的景点。本次研发得到的景点信息页面如图 7 所示。

图 7　景点信息页面

Fig. 7　Scenic spots informations page

4　结束语

目前的旅游攻略仅有富文本信息流单一形式，用户难以获取足够的、有关联的、真正感兴趣的旅游信息，所以专门针对该需求，本次研究开了发旅游知识图谱。研发中，利用 Phantom JS 和 Selenium 结合搭建爬虫框架，获取精细粒度旅游数据，以景点、旅店、餐厅为对象实体构造知识节点，持久化在图数据库中；在前端使用 roc-charts 技术实现图谱的可视化。最终解决信息孤岛，实现旅游信息的知识图谱，并基于图谱实现搜索系统，最终成果将封装成 Web 形式提供服务。

参考文献

[1] 邹莹莹. 基于用户生成内容的旅游知识图谱构建和信息服务研究[D]. 广州：华南理工大学，2020.

[2] 董杭. 知识图谱技术在银行网络安全中的应用研究[J]. 电子世界，2021(18)：63-64.

[3] 刘峤，李杨，段宏，等. 知识图谱构建技术综述[J]. 计算机研究与发展，2016，53(03)：582-600.

[4] 姚萍，李坤伟，张一帆. 知识图谱构建技术综述[J]. 信息系统工程，2020(05)：121，123.

[5] 邵嘉进，陈成栋，陶俊樾，等. 基于画像的旅游推荐服务实现[J]. 电脑编程技巧与维护，2021(07)：147-149.

[6] 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(04)：589-606.

[7] 张宇飞. 河北省旅游景点知识图谱的构建与应用[D]. 邯郸：河北工程大学，2020.

(上接第 65 页)

4　结束语

本文研究本地差分隐私机制 *RR* 和 *OUE* 在权重聚合的联邦学习上实现隐私保护并优化服务端聚合速度近似从 $O(n^2)$ 减少为 $O(1)$，同时评估 *cells* 的个数、ε、客户端数量、扰动机制和数据分配方式对模型精度影响，精度损失会随着 *cells* 数量、ε、客户端数量的增加而减小。同等 ε 下，*NonIID* 会比 *IID* 的精度损失高，*RR* 和 *OUE* 的精度损失类似；同时实现自适应隐私预算策略提升模型训练速度，近似提升一个 ε 等级的收敛速度。

在此基础上，本文仍存在需要改进之处。本文并未评估梯度聚合框架下的效果。同时还未评估联邦学习下本地差分隐私对隐私泄露的影响。另外，本文现有的数据集不够全面，只有 2 类典型数据集，有待进一步研究解决。

参考文献

[1] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models [C]//2017 IEEE Symposium on Security and Privacy (SP). San Jose: IEEE, 2017: 3-18.

[2] GENTRY C. Fully homomorphic encryption using ideal lattices [C]//Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing. New York, USA: ACM, 2009: 169-178.

[3] LE PHONG T, AONO Y, HAYASHI T, et al. Privacy-preserving deep learning via additively homomorphic encryption [J]. IEEE Transactions on Information Forensics and Security, 2018, 13(5): 1333-1345.

[4] DWORK C. Differential privacy[C]//International Colloquium on Automata, Languages, and Programming. Venice, Italy: Springer, 2006: 1-12.

[5] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria: ACM, 2016: 308-318.

[6] DWORK C, ROTHBLUM G N. Concentrated differential privacy [J]. arXiv preprint arXiv: 1603.01887, 2016.

[7] BUN M, STEINKE T. Concentrated differential privacy: Simplifications, extensions, and lower bounds [M]//HIRT M, SMITH A. Theory of Cryptography. TCC 2016. Lecture Notes in Computer Science. Berlin/Heidelberg: Springer, 2016, 9985: 635-658.

[8] WARNER S L. Randomized response: A survey technique for eliminating evasive answer bias [J]. Journal of the American Statistical Association, 1965, 60 (309): 63-69.

[9] WANG Tianhao, BLOCKI J, LI Ninghui, et al. Locally differentially private protocols for frequency estimation[C]//SEC' 17: Proceedings of the 26th Usenix Conference on Security Symposium. Berkeley, CA: Usenix Association, 2017: 729-745.

[10] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]// 20th International Conference on Artificial Intelligence and Statistics. Seattle, Washington: PMLR, 2017: 1273-1282.

[11] ZHAO Yue, LI Meng, LAI Liangzhen, et al. Federated learning with non-iid data[J]. arXiv preprint arXiv: 1806.00582, 2018.