计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(6):121−125 [doi: 10.15888/j.cnki.csa.007424]

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

基于环境信息融合的知识图谱构建方法①

宋　伟[1], 张游杰[2]

[1](太原科技大学 计算机科学与技术学院, 太原 030024)
[2](中电科华北网络信息安全有限公司, 太原 030032)
通讯作者: 宋　伟, E-mail: sweety@stu.tyust.edu.cn

摘　要: 针对多源异构的环境数据难以利用的问题, 在通用知识图谱的基础上, 融合各类环境数据构建环境知识图谱. 首先利用网络爬虫等获取环境数据, 并进行数据预处理; 进而利用结构化数据转化、文本抽提以及数据融合等技术, 研究基于环境信息融合的知识图谱构建方法; 最后将生成的知识图谱存入图谱数据库，并搭建知识图谱应用平台, 提供递归查询功能, 实现环境知识图谱的可视化, 以期为相关人员提供有益参考.

关键词: 知识图谱; 数据融合; 可视化; 环境信息

引用格式: 宋伟,张游杰.基于环境信息融合的知识图谱构建方法.计算机系统应用,2020,29(6):121–125. http://www.c-s-a.org.cn/1003-3254/7424.html

Knowledge Graph Construction Method Based on Environmental Data Fusion

SONG Wei[1], ZHANG You-Jie[2]

[1](College of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)
[2](CETC North China Cyber Security Co. Ltd., Taiyuan 030032, China)

Abstract: Aiming at the problem that using multi-source heterogeneous environmental data is difficult, based on the general knowledge graph, this study constructed the environmental knowledge graph by fusing various environmental data. Firstly, the environmental data are obtained by using Web crawlers and preprocessed later; then, data conversion, text extraction, and data fusion technologies are used to build the knowledge graph; finally, the generated knowledge graph is stored in the graph database, and then we build a knowledge graph application platform to provide a way to searching information recursively and to realize the visualization environmental knowledge graph. Such platform is more comprehensive thus it may offer better references for the relative personnel.

Key words: knowledge graph; data fusion; visualization; environment data

随着大数据时代的来临, 数据的种类多种多样, 数据规模日益增大, 传统的数据管理模式和类 SQL 语句查询受到一定的限制. 知识图谱 (Knowledge Graph, KG) 作为一种新的知识表达方式及数据管理模式, 旨在描述客观世界的概念、实体、事件之间的关系, 其基本的组成单位是“头实体-关系-尾实体”三元组, 实体包含属性键值对, 实体之间通过关系进行描述, 形成网状的结构. 网络中的节点代表物理世界中的实体或概念, 实体之间的各种语义关系构成网络中的边.

知识图谱标准化白皮书中[1]中明确指出大数据时代, 应对碎片化的数据进行整合, 消除“信息孤岛”和“数据烟囱”, 将数据转为可供决策使用的知识和智慧. 在环境问题上, 各方数据孤立的现象尤为明显, 环境空气质量、水质质量、环保工作等数据存在与不同的机构之间, 没有一个统一的对环境情况的描述.

本文针对上述问题进行了研究, 介绍了环境知识

① 收稿时间: 2019-10-23; 修改时间: 2019-11-22; 采用时间: 2019-11-29; csa 在线出版时间: 2020-06-10

图谱的一般构建过程, 实现了环境信息知识图谱的构建, 将不同来源的异构环境信息进行了融合.

1 相关工作介绍

目前国际上较知名的知识图谱为 DBpedia, 是一个以维基百科为数据源的通用知识图谱, 用以增强维基百科的搜索功能[2]. 与 DBpedia 对应的为 CN-DBpedia, CN-DBpedia 是由复旦大学知识工厂肖仰华等从中文百科类网站 (如百度百科、互动百科、中文维基百科等) 的纯文本页面中提取而来的通用知识图谱, 其中包含 900 余万的三元组关系[3]. 上述知识图谱是从各类网页上采集而来的通用图谱, 针对特定的领域, 通用知识图谱没有很好的表现. 王雪芹等以 1997–2017 年的 CNKI 为数据源, 构建了针对矿区生态环境研究知识的专业图谱[4]. 孙强强等提出建立基于知识图谱环境科学知识挖掘, 是未来环境治理研究的发展方向[5]. 环境相关的知识图谱集中在环境治理方法图谱, 对环境情况的表现不足.

2 知识图谱的一般构建过程

为知识图谱的应用是以知识图谱的构建为基础的, 其中构建知识图谱的主要过程包括实体抽取和实体间关系的建立. 知识图谱在逻辑上分为模式层和数据层, 可视为一张图 G, 由模式图 G_s、数据图 G_d 以及 G_s 和 G_d 之间的关系 R 组成, 即 $G=<G_s, G_d, R>$. 模式层基于数据层之上, 是知识图谱的核心, 其表现形式为: 实体-关系-实体, 关系-属性-属性值. 数据层由一系列事实组成, 如: AQI-中文名-空气质量指数.

知识图谱的一般构建方法有自顶向下构建 (top-down) 和自底向上构建 (bottom-up) 两种[6]. 自顶向下是首先为图谱定义好全局本体, 即从数据源中先提取本体和模式信息, 再将实体加入图谱中. 而自底向上方法对实体进行归纳, 提取出置信度高的加入图谱中. 这两种方法不是孤立进行的, 可以两者交替结合. 本研究在构建知识图谱时采用两种方法的结合, 先通过一个通用知识图谱构建本体库, 再自底向上提取数据扩展知识图谱.

多数据源融合构建知识图谱, 如图 1 所示. 由不同来源、不同结构的数据, 如结构化、半结构化和非结构化数据, 通过关系抽取、属性抽取、实体消歧, 转化为符合图谱构造的三元组形式. 最后编写相应的展示平台, 对知识图谱提供一个外部展示及交互接口.

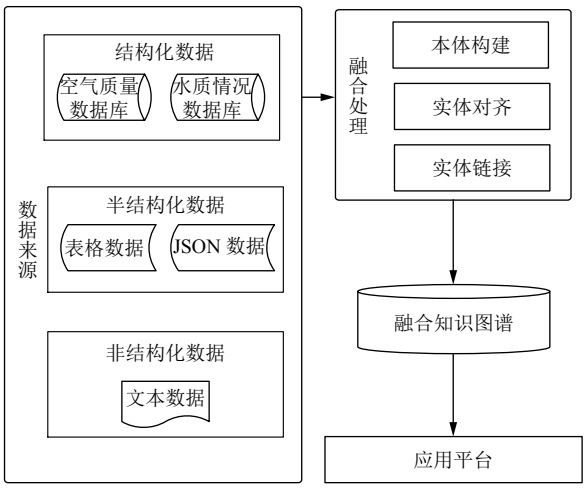


图 1 多数据源知识图谱构建过程

3 多数据源融合的环境知识图谱构建

3.1 数据源

用于建立知识图谱的数据源可以是结构化数据、半结构化数据和非结构化数据, 现有的一些通用知识图谱也可以作为数据的来源[7,8].

1) 结构化数据. 当前空气质量数据库数据.

2) 半结构化数据. 主要是获取历史天气情况的 JSON 格式数据、历史水质情况的表格格式数据.

3) 非结构化数据. 主要是文本数据, 如各地环保厅网站、环保局网站的工作动态文本, 和环境介绍的描述文件等.

4) 通用知识图谱. 使用思知 (OwnThink) 通用知识图谱, 包含了 2500 万实体和千万级别关系的中文图谱, 以文本三元组格式保存.

这些数据源共同为作为环境图谱的数据来源. 其中, 对于结构化数据中的空气质量数据, 和历史空气质量数据的数据频率不一致. 历史空气情况的频率是一天一个地区只有一条记录, 而实时爬取的空气数据每个小时都有一条记录, 需要将记录取均值, 从而与历史数据频率一致, 便于处理. 水质情况数据中存在大量缺失的缺失值, 这时可以采用均值法或剔除法. 非结构化的文本数据提取出的很多名词有些是中文名词, 有些是英文缩写, 但是指代的为同一实体. 通用图谱是一个大文本的三元组文件, 普通的文本编辑工具不能打开处理, 将三元组导入图谱时又要将其处理为特定的格式, 为各个实体和关系添加唯一的 Id 和生成对应 csv 文件, 需要使用大文本处理工具.

3.2 数据获取及数据预处理

使用 Scrapy 爬虫框架, Scrapy 是一个 Web 页面抓取框架, 可用于抓取 Web 站点并利用 Xpath 从页面中提取结构化数据. 从环境生态部网站、各环保厅网站、各环保局网站采集工作动态文本内容, 忽略其中的图片及附件等内容. 对采集的数据保存为文本格式, 并用 Python 进行预处理, 将其中的网页标签和乱码做删除处理.

爬虫工作流程如图 2 所示, 具体可描述如下:

(1) 设置待爬取网站的种子 URL, 这是一个列表形式, 用于定义初始请求. Scrapy 根据种子的初始请求开始进行抓取.

(2) 将种子 URL 的生成待爬取网页地址, 然后把网页下载下来, 存入已下载网页集合中, 标记为已爬取网页.

(3) 分析已爬取网页中的 URL, 将 URL 放入待抓取 URL 队列中, 重复 (1)~(3) 步.

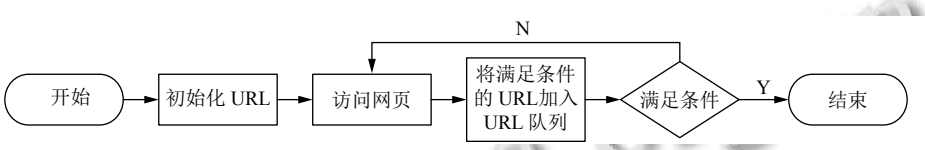


图 2 爬虫流程

从中国环境监测总站信息发布网站上采集每小时的空气数据存入 MySQL 数据库中, 将其中含有缺失值的项删除. 空气质量数据主要包括检测站点的名称、监测站代码、AQI 指数、可吸入颗粒物的值等.

从国家地表水水质自动检测实时数据发布系统采集水质情况数据, 删除缺失的项. 水质数据包括检测的站点名、水酸碱度、水中溶解氧的含量等一系列数据.

3.3 本体库构建

本体 (ontology) 是对概念进行建模的规范, 是描述客观世界的抽象模型, 以形式化的方式对概念及其之间的联系给出明确定义[9]. 本体可以借助本体编辑软件进行手工方式构建, 也可以以数据驱动的自动化方式构建本体, 通过分析关系数据库中表的信息和字段信息, 构建相应的概念模型[10].

图谱的数据源来自于空气质量检测数据和水质情况数据, 而关系数据库包含完整的表结构和完整性的约束条件, 可以从关系型数据库中抽取出关系模式, 根据关系型数据库中表信息和字段信息, 建立相应的概念模型, 利用规则将关系模式转为本体模型[11,12].

针对 JSON 格式和表格形式的半结构化数据, 通过将半结构化数据转换为结构化数据, 再通过规则将其转化为表名转为概念名: 将关系模式中字段名转为本体属性名等. 环境信息知识图谱结构如图 3 所示.

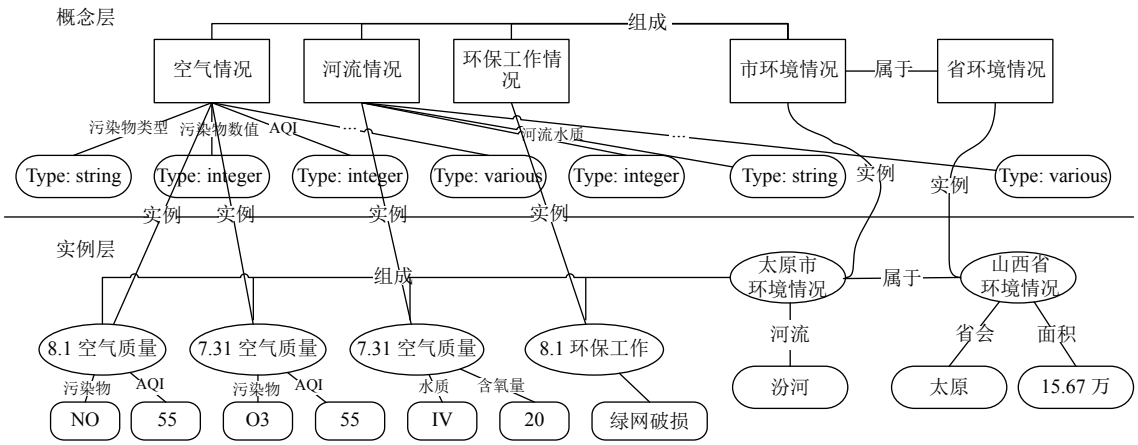


图 3 环境信息知识图谱结构

在与通用知识图谱融合过程中, 主要需要进行实体对齐操作 (entity alignment), 判断新提取出的实体和通用知识图谱中的实体是否指向同一对象, 将这些实体进行合并, 并用唯一表示对该实体标记, 最后将实体

抽取出的新关系添加到图谱中. 如: 空气质量和大气质量指向的语义相同, 通过实体对齐可以将其定义到一个实体下. 本文通过已训练好的词向量模型 (Word2Vec) 的词相似度进行判断, 词向量将词进行了向量化. 通过半监督学习, 词越相似, 其余弦相似度越高. 通过对候选实体中相似度得分高的实体进行合并, 进行实体对齐操作.

从国家地表水水质自动检测实时数据发布系统采集水质情况数据, 删除缺失项, 水质数据包括检测站点名、水的酸碱度、水中溶解氧含量等一系列数据.

3.4 句法依存的三元组抽取

而对于文本类的非结构化数据进行处理, 主要是将文本提取为多个三元组的集合. 提取的方法有 3 种: (1) 无监督提取. 这种提取方法需要由领域专家手工编写规则或模式, 然后进行抽取. (2) 半监督提取. 人工给出部分种子实例, 由机器学习挖掘符合该模式的实例, 再将这些实例加入种子实例中. (3) 无监督提取. 将句子中符合一定语法规则的关系组提取出来. 本文主要采用无监督的文本三元组提取, 基于哈工大 LTP 工具, 利用句法依存的关系提取三元组. 图 4 为句法依存示意图.

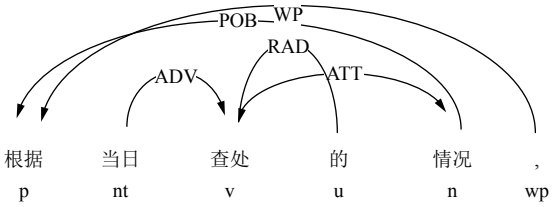


图 4 句法依存示意图

文本由多个句子组成, 一篇由 n 个句子组成的文档 D, 其中 S_n 表示第 n 个句子:

$$D = [S_1, S_2, \cdots, S_n] \tag{1}$$

对每个句子 S_n 进行分词处理, 将句子变为一系列的词汇和标点组成的向量, 其中 W_m 代表单个词汇或标点符号:

$$S_i = [w_1, w_2, \cdots, w_m] \tag{2}$$

再通过词性标注得到向量:

$$POS_i = [p_1, p_2, \cdots, p_m] \tag{3}$$

其中, p_m 代表每个词的词性, p_m 为 w_m 的对应词性.

通过对 S_n 进行句法分析, 得到:

$$PAR_i = [(r_1, t_1)_1, (r_2, t_2)_2, \cdots, (r_m, r_m)_k] \tag{4}$$

其中, r_m 表示对应的 w_k 和 w_m 词, t_m 表示句法依存的关系, 如: 主谓宾关系 (SBV)、定中关系 (VOB) 等.

抽取以谓词为中心的三元组过程如下:

遍历 PAR 向量, 寻找含有 VOB 和 SBV 关系的词. 即寻找一个句子中的谓语动词, 并将主语和宾语提取出来构成主谓宾三元组. 这样提取出的三元组不够完善, 由于句中进行了分词操作, 每个词都是独立存在的, 因此提取出的主语和宾语较短, 由于没有修饰词来说明实体, 抽取出的词语不能完整准确地表达出意思, 甚至会由于词汇太短从而出现语义不明的情况, 如表 1.

表 1 提取出的三元组

主语	谓语	宾语
靳队长	表示	存在
绿网	存在	破损

需要进一步将实体词进行完善, 补全主语和谓语的定语, 递归地把实体的修饰词补全, 形成完整的主语实体. 在递归补全实体的过程中, 对实体的修饰词长度进行限制, 过长的修饰词会淹没中心词造成中心语残缺.

先寻找句子中的动词作为三元组的中间词, 通过递归地把实体词的修饰语补充完整. 为了避免递归导致实体词过长, 设置修饰词的长度为 10, 超过修饰词长度上限就结束递归, 在完整表达实体语义的前提下减少过长修饰词出现的可能性. 具体过程如图 5 所示.

```
algorithm 1 CompleteEntity
function CompleteEntity(S,POS,PAR,Index)
    prefix ← NULL
    postfix ← NULL
    if POS[Index] == 'verb' then
        if 'VOB' in PAR then
            postfix+ = CompleteEntity(S, POS, PAR, Index + 1)
        end if
        if 'SVB' in PAR then
            prefix+ = CompleteEntity(S, POS, PAR, Index − 1)
        end if
    end if
    if len(prefix) >10 or len(postfix) >10 then
        return prefix + S[Index] + postfix
    end if
    return prefix + S[Index] + postfix
end function
```

图 5 完善实体伪代码

从表 2 结果看, 由于补全了实体的修饰语, 使得实体的描述更为准确, 而未补全的实体语义表述不明. 补全实体的过程是一个递归的过程, 把实体的前缀词和后缀词递归地加入实体中, 最终形成完整的实体.

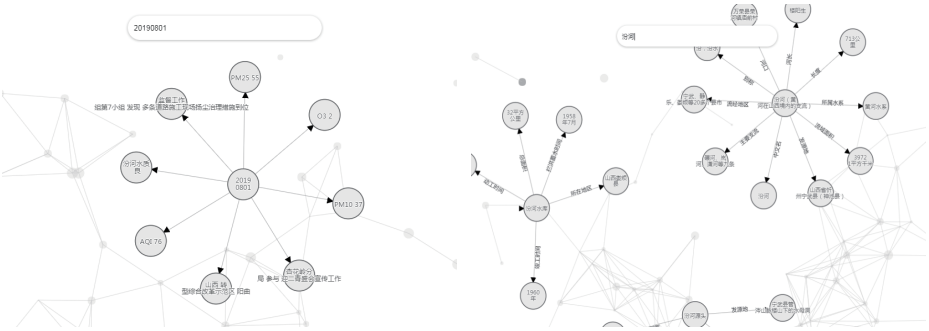
表 2 完善实体的三元组

主语	谓语	宾语
阳曲产业园区执法中队靳队长	表示	苫盖绿网存在破损情况
苫盖绿网	存在	破损情况

4 应用平台搭建

知识图谱的可视化主要是利用可视化技术构建的一种知识之间的关系网格图. 本文开发了一个知识图谱的可视化应用服务平台, 平台采用 Neo4j 作为图形数据库, 在前端使用 D3 构建可交互的数据图表, 使用 PHP 作为连接 Neo4j 数据库和返回查询数据的中间服务. 其主要功能有: (1) 为用户提供基础查询服务; (2) 递归查询各个实体; (3) 知识图谱实体关系网络的可视化, 实现概念、属性、实例等多个维度的知识图谱展示.

截取两幅图对展示平台进行简单说明. 图 6 表示某地某日的环境情况, 包括空气质量情况, 水质情况和当地环保工作的内容. 双击图中节点, 可进入下一层知识图谱, 如双击汾河, 可展示出其图谱内容, 图谱中每个节点又可以双击进入下一层, 实现递归查询图谱.



(a) 展示平台截取图1 (b) 展示平台截取图2

图 6 环境知识图谱展示

5 结论与展望

本文提出一种基于多数据源数据融合的知识图谱构建过程, 利用网络爬虫采集空气质量检测数据、河流数据和环境工作文本数据, 对数据进行融合处理, 旨在构建一个多源异构数据的融合知识图谱, 将构件流程工具化, 以便为环境工作相关人员提供更好的支撑.

相较与将各类信息分别使用不同的存储形式和不同的数据库类型, 把多源异构的数据以图谱形式存储可以进行数据的统一, 方便地使用类 SQL 语句进行查询, 作为智能推理和智能问答的基础. 文中描述的图谱构建过程也可以应用于其他领域, 将领域中不同类型和结构的数据统一导入图形数据库形成图谱.

目前知识图谱的构建方法还处于发展期, 部分技术及图谱构建算法还需要改进. 本文中所构建的多数据源知识图谱还有很多不足之处, 比如其数据源还不够完善, 应使用更多的相关数据源来扩展图谱, 尤其是相关的专业知识融入图谱中; 没有实现建立图谱的自动更新机制, 让知识图谱实现自增长.

参考文献

1 中国电子技术标准化研究院. 知识图谱标准化白皮书. 北京: 中国电子技术标准化研究院, 2019.

2 Bizer C, Lehmann J, Kobilarov G, *et al*. DBpedia - A crystallization point for the web of data. Journal of Web Semantics, 2009, 7(3): 154–165. [doi: 10.1016/j.websem.2009.07.002]

3 Xu B, Xu Y, Liang JQ, *et al*. CN-DBpedia: A never-ending Chinese knowledge extraction system. Proceedings of the 30th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Arras, France. 2017. 428–438.

4 王雪芹, 盛武. 我国矿区生态环境研究知识图谱分析. 河南理工大学学报 (社会科学版), 2019, 20(2): 47–53.

5 孙强强, 姜宛贝, 孙丹峰. 基于知识图谱和综合征的科学环境知识挖掘——以民勤荒漠化为例. 干旱区地理, 2018, 41(2): 426–434.

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

7 张瑶, 李蜀瑜, 汤玥. 大数据下的多源异构知识融合算法研究. 计算机技术与发展, 2017, 27(9): 12–16.

8 刘化冰. 知识图谱在知识产权大数据应用中的模式探索. 科技与出版, 2018, (12): 95–98.

9 黄恒琪, 于娟, 廖晓, 等. 知识图谱研究综述. 计算机系统应用, 2019, 28(6): 1–12. [doi: 10.15888/j.cnki.csa.006915]

10 Rong X. Word2Vec parameter learning explained. arXiv: 1411.2738, 2014.

11 阮彤, 孙程琳, 王昊奋, 等. 中医药知识图谱构建与应用. 医学信息学杂志, 2016, 37(4): 8–13.

12 李颖, 郝晓燕, 王勇. 中文开放式多元实体关系抽取. 计算机科学, 2017, 44(S1): 80–83.