[文章编号] 1003－4684(2019)05-0072-06

基于知识图谱的司法案件可视化研究与实现

陈建峡，黄煜俊，曹国金，杨 帆，李 超，马忠宝
（湖北工业大学计算机学院，湖北 武汉 430068）

[摘 要] 针对司法领域办案过程中所需知识分散、不完备、查询不便等问题，通过对收集到的司法案件进行分析归纳总结，首先采用哈工大的 LTP 语言技术平台对司法案件文本进行分词、词性标注及命名实体识别等处理，接着通过依存句法分析算法从处理过的文本中抽取出实体间的语义关系并存储为三元组的形式，然后将三元组形式的数据信息录入到 Neo4j 图数据库，利用 Neo4j 实现司法案件知识图谱的构建并对其进行可视化展示。最后以实例证明了该方法的可行性。

[关键词] Neo4j；图数据库；知识图谱；依存句法分析
[中图分类号] TP391 [文献标识码] A

司法领域的知识体系庞大，领域知识也比较复杂，随着数据量的不断增大，数据之间的关系也越来越复杂，只能处理简单数据关系的传统关系型数据库已无法胜任，知识图谱的兴起便是为了解决该难题。知识图谱是用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系[1]。近年来，知识图谱获得迅猛的发展，目前已成为分析学科领域热点和前沿的有力工具[2]。

2012 年 5 月份 Google 公司就首先提出了“知识图谱”的概念[3]，旨在提升其搜索引擎性能而建立的知识库。Zhang 等人认为知识图谱可以应用于展示领域知识整体结构、可视化分析检索结果[4]；CiteSpaceⅡ软件是一款针对所采集到的数据进行知识图谱分析，专门用于在科学文献中识别并显示科学发展新趋势和新动态的软件[5]；Wang 等人借助 CiteSpaceⅡ软件绘制了国际刑事司法研究领域的知识图谱，并作可视化分析，发现在国际刑事司法研究方面存在着注重理论与实证的两种趋势[6]。

目前，国内知识图谱的研究内容主要集中在知识图谱的构建和知识表示学习与推理的方法[7]。Wang 等人通过知识图谱梳理了我国近年来司法鉴定学科的研究热点与演进趋势，客观地展示其研究成果，为相关人员提供直观的参考依据[8]。

综上所述，面向司法领域知识图谱的研究还是比较匮乏。为此，本文提出了基于知识图谱的司法案件可视化分析方法，采用 LTP 语言技术平台对司法案件文本进行分词、词性标注、命名实体识别和依存句法分析等处理，获得诸如原告、被告、案件类型等关键信息，再利用 Neo4j 图数据库对其进行整理和编译，将其整合成为结构化语义网络构建司法案件知识图谱，最终实现特定查询的功能并优化了信息获取的速度。

1 相关技术概述

1.1 知识图谱简介

1.1.1 知识图谱基本原理 知识图谱由节点和连接节点之间的边组成，是知识的一种结构化图解表示，利用实体表示现实世界中存在的事物或抽象概念[9]，知识图谱中的每一个节点都与一个实体相对应，且每一个节点都具有一个全局的、唯一的 ID，而实体间的相互关系则被抽象为连接节点之间的边，边可以是无向的，也可以是有向的。

知识图谱主要采用三元组形式表示知识，其通用表达方式为：(head，label，tail)，简记为(h，l，t)，其中 head 与 tail 分别表示三元组中的头实体和尾实体，两个实体之间的关系用 label 作标记；h、t 都属于实体集合 E(Entities)，l 属于关系集合 R(Relationships)[10]。一个简单的知识图谱三元组实例如图 1 所示。

1.1.2 知识图谱构建流程 知识图谱的构建需要

[收稿日期] 2019－08－14
[基金项目] 湖北省科技厅自然科学基金青年面上项目(2017CFB326)
[第一作者] 陈建峡(1971－)，女，湖北省武汉人，湖北工业大学副教授，研究方向为人工智能，自然语言处理

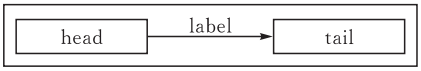


图 1 知识图谱三元组实例

借助来自不同研究领域的研究成果。通过知识抽取技术，从不同形式的数据源获取知识图谱构建的各类知识。采用知识融合剔除各类不合适的知识，提高知识图谱的质量与性能。知识图谱的构建流程见图 2。

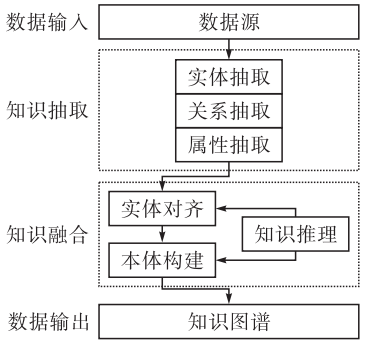


图 2　知识图谱构建流程

1)知识抽取

知识抽取是指使用自动化抽取技术，从开放链接数据、结构化数据、半结构化数据、自动化的 AVP(Attribute Value Pair，属性–值对)得到相应的结构化信息，包括实体、实体间关系以及属性[11]。通过知识抽取得到的常识性知识，以及各领域的专业知识是构建知识库的重要数据来源，决定了知识服务的质量。

2)知识融合

由于知识的来源广泛，可能包含大量的歧义、冗余甚至是错误的信息，质量难以保证，因此必须对原始知识进行融合。知识融合所涉及的技术主要有：实体消歧、共指消解，多数据源合并等[12]。知识融合的工作包括实体对齐和本体构建等环节，形成统一的知识表达表示形式，以便建立更加清晰完善的知识体系。

3)知识推理

知识推理是指根据知识图谱中已有的知识，使用一些推理模型与算法推断出新的未知的知识，用来提高知识的完备性，扩大知识的覆盖面。常用的知识推理方法包括基于规则的知识推理和基于统计的知识推理。

1.2　图数据库概述

图数据库是 NoSQL 的一种，适合表达图结构的数据，能很方便地建立和使用图形模型，并且图数据库扩展方便，适合分布式系统[13]。图数据库中最主要的组成要素是节点以及节点间的关系：每个对象是一个节点(Node)，节点可以具有多种属性(Property)，节点之间的关系(Relationship)表示为边，节点通过关系相连，形成关系网络的结构。图 3 展示了图数据库的数据结构。

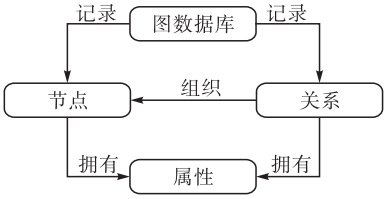


图 3　图数据库的数据结构

1.2.1　Neo4j 图数据库简介　Neo4j 是一款开源的、高效率的并提供很多编程语言的 REST API 接口的图数据库[14]，它具有 ACID 支持、强移植性、可扩展性、可更新性、高吞吐量等特征[15]提高了数据存储和检索的性能。

1.2.2　Cypher 语言简介　Neo4j 使用的 Cypher 语言是一种图数据库查询和更新语言，在关联查找能力上比 SQL 更强大，能轻易地找到实体间隔的轨迹。当入库的数据量较少时能使用 Cypher 查询语言逐一添加；当入库的数据量较多时可以从 csv 格式文件或关系型数据库中批量导入数据。

1.3　LTP 语言技术平台简介

LTP 是哈工大研发出的一种中文语言处理系统，具有语法分析、句法分析及语义分析等中文处理功能。LTP 底层以 XML 表示文本，以 DOM 处理文本。同时 LTP 能够通过 Web 以图形化界面直观地查看 XML 文件存储的处理结果。LTP 的语言技术平台架构如图 4 所示。

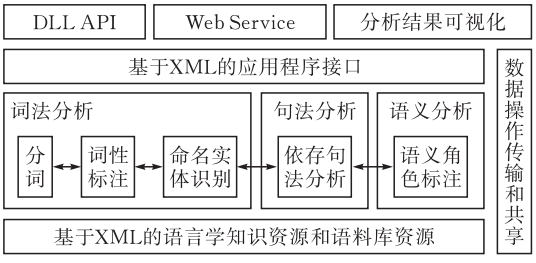


图 4　LTP 语言技术平台架构

2　司法案件数据获取

2.1　数据来源

目前关于知识图谱的研究领域中，其数据来源一共分为三大类：结构化数据、半结构化数据和非结构化数据。法院所保存的司法案件文本信息大都属于非结构化数据，因此，本文主要是基于这些非结构化数据进行知识图谱的构建。

2.2　数据获取

如图 5 所示，本文利用 ProcessOn 软件把司法案件整理为两大类。左边一类是在司法案件中出现

的涉案人身份，右边一类是司法案件中的涉案种类，形成司法案件知识导图。

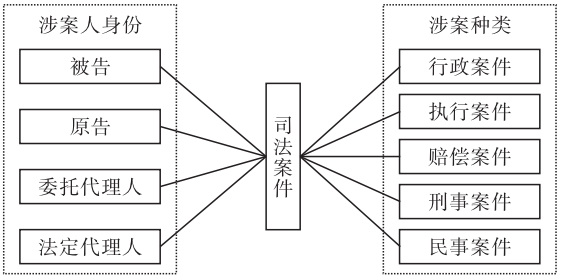


图5 司法案件知识导图

3 司法案件实体关系抽取

实体关系的抽取实际是将一个司法案件里的主要人物及其身份抽取出来。本文采用LTP首先对司法案件文本进行预处理，再依据依存句法分析算法，解释文本的句法关系；最后根据分析后的结果来构建实体关系三元组。司法案件实体关系抽取流程见图6，主要分为文本预处理、依存句法分析和实体关系三元组构建等三个模块。

3.1 文本预处理

从图6可得，文本预处理是司法案件实体关系抽取整个过程的第一步。首先从司法案件文本语料库获取需要进行关系抽取的文本。本文采用法院的案件判决书作为语料库，案件判决书原始文本部分截图见图7。

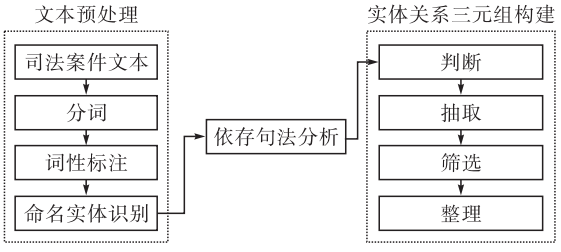


图6 司法案件实体关系抽取流程

湖北省武汉市中级人民法院

刑事判决书

武中刑一初字第1号

公诉机关湖北省武汉市人民检察院。

被告人张三，乳名小三，男，汉族，1997年8月10日出生于湖北省荆州市，初中文化，农民，住湖北省荆州市滩桥镇黄桥村。2013年10月5日因涉嫌犯故意杀人罪被刑事拘留，同年11月23日被逮捕。现羁押于滩桥镇看守所。

辩护人李四，湖北省智方律师事务所律师。

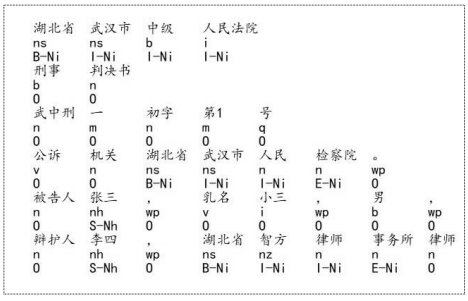
图7 案件判决书案例

文本选定后，进入分词阶段。利用LTP的分词模块，将连续的字序列按照一定的规范重新组合成词序列。即在句子的每个字标注一个词边界的标记，开始边界和结束边界中的内容就是一个词，这样便达到分词的效果。

分词后还需要对其词性进行标注。利用LTP的词性标注模块，依次标注出每个词的词性，标注的结果不仅应用于LTP的命名实体识别模块，还应用于后面的依存句法分析模块。

命名实体识别是文本预处理中最关键的一步。根据标注的词性，利用LTP的命名实体识别模块，识别出文本中具有特定意义的实体，主要包括人名、地名和机构名。

司法案件文本经过预处理之后的部分结果见图8。



湖北省	武汉市	中级	人民法院				
ns	ns	b	i				
B-Ni	I-Ni	I-Ni	I-Ni				
刑事	判决书						
b	n						
0	0						
武中刑	一	初字	第1	号			
n	m	n	m	q			
0	0	0	0	0			
公诉	机关	湖北省	武汉市	人民	检察院	。	
v	n	ns	ns	n	n	wp	
0	0	B-Ni	I-Ni	I-Ni	E-Ni	0	
被告人	张三	，	乳名	小三	，	男	，
n	nh	wp	v	i	wp	b	wp
0	S-Nh	0	0	0	0	0	0
辩护人	李四	，	湖北省	智方	律师	事务所	律师
n	nh	wp	ns	nz	n	n	n
0	S-Nh	0	B-Ni	I-Ni	I-Ni	E-Ni	0

图8 司法案件文本预处理部分结果

3.2 依存句法分析

依存句法是LTP中非常重要的句法结构分析子系统，也是实体关系抽取中的主要模块。其句法分析的基本任务是通过分析句子中各个成分之间的依赖关系，进而揭示句子的句法结构。该算法的结果是将句子中的依赖关系抽取出来。

LTP中的依存句法分析模块的算法核心依据神经网络依存句法分析算法，通过神经网络的方式，同时还加入丰富的全局特征和聚类特征设计了一个句法分析器来分析句子中的句法结构和句子中主要成分之间的依存关系。该神经网络依存句法分析算法模型的结构见图9。

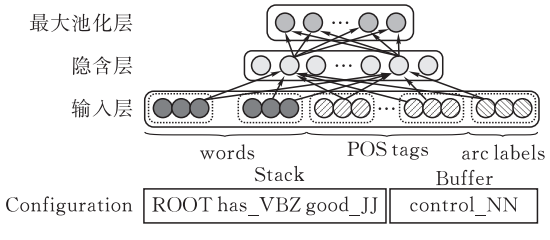


图9 依存句法分析神经网络

从图9可知，该模型包括输入层、隐含层和最大池化层。其中Configuration包括Stack、Buffer和依存弧集合的信息。依存弧集合模型根据Configuration信息来提取出一个特征向量，这个特征向量由(words，POS tags，arc labels)三个向量拼接而成。其中words是当前移位操作时栈和buffer中的部分词汇以及栈中部分单词的依赖单词；POS tags是上述单词的词性；arc labels是上述依赖单词

的依赖关系。这个向量因此包含了对应 configuration 的信息。

该模型的目标就是输入特征向量，然后预测出对应的转换类型，预测出转换类型就进行相应的转换操作，这样就更新了配置信息，然后得到新的向量，再输入模型中预测，如此循环。最后找出句子中依存关系。

预处理后的司法案件文本经过依存句法分析的部分结果如图 10 所示。

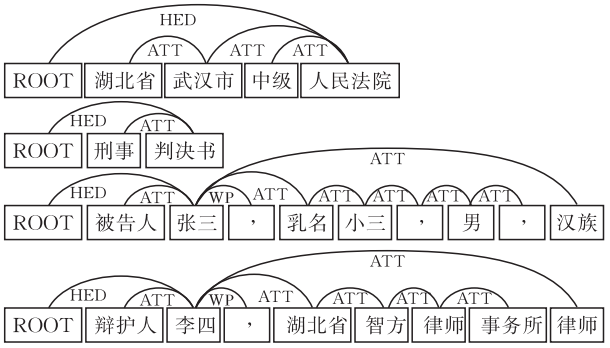


图 10　司法案件文本依存句法分析部分结果

3.3　实体关系三元组构建

实体之间的关系往往能用句子中的一个词语来描述。实体关系抽取实则是构建实体关系三元组的过程。如从“张三的父亲是李四”这句话中构建实体关系三元组，首先利用依存句法分析算法处理这句话，其结果见图 11。

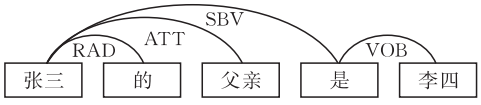


图 11　依存句法分析算法处理结果

从图 11 可以看出，“父亲”这个关系词依赖于“张三”，而“张三”以“主谓宾”的形式指向了“李四”，从此可判断出“父亲”就是描述“张三”和“李四”两个实体之间的关系。然后便能构建实体关系三元组：(张三，父亲，李四)。

本模块根据文本句子描述的特性，将句子中的关键信息以三元组的形式抽取出来，并对三元组进行有效性的筛选和整理，构成最适合需求的实体关系三元组，其抽取算法流程见图 12。

司法案件文本的实体关系三元组构建的最终结果截图见图 13，司法案件文本里的主要人物及其身份都被抽取出来。

4　司法案件知识图谱的构建与可视化

抽取出来的实体关系三元组形成完整严谨的知识语言逻辑体系，成为司法案件知识图谱的理论基础。通过定义实体与实体之间的关系，能够定义涉

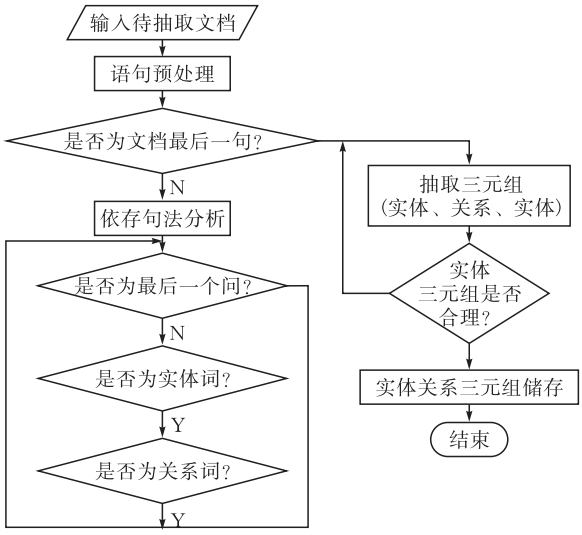


图 12　实体关系抽取算法流程

(张三，被告人，武中刑一初字第 1 号)
(李四，委托代理人，武中刑一初字第 1 号)
(杨杰，原告人，武中刑一初字第 1 号)
(杨彬，原告人，武中刑一初字第 1 号)
(王燕，法定代表人，武中刑一初字第 1 号)
(张明，法定代表人，武中刑一初字第 1 号)

图 13　实体关系三元组构建结果

案、种类、受理等多种关系，凭借着各种关系，多个不同实体间的关系就能够构成一套容纳司法案件中的实体与实体间关系的知识库。在此基础上，利用 Neo4j 图数据构建司法案件知识图谱。同时，司法案件知识图谱的可视化也利用 Neo4j 图数据库实现。

4.1　实体关系分类

司法案件知识图谱的实体节点数量很多，若要形成一个有明确知识架构的知识网络，就需要把相关的实体连接起来，即实体的关系。表 1 展示了各实体之间可能会有的关系，依据这些具体的关系能够将相关实体都联系起来。

4.2　司法案件知识图谱构建

本文利用 Neo4j 图数据库对实体及其关系分类后的数据信息进行存储。考虑到司法案件文本数量比较多，因此需要批量入库。首先将数据信息全部整合并转化为 csv 格式文件，采用“neo4j-admin import -mode=csv”的方式，将 csv 文件批量导入到 Neo4j 图数据库中，并在 Neo4j 图数据库中构建知识图谱体系。

实体节点 csv 文件，以被告实体为例，其 csv 文件除了保存被告人实体名外，还包含了其相关属性如性别、出生年份和出生地等信息。通过在 Neo4j 中输入实体节点批量导入代码，便能实现实体节点批量入库操作。其关键代码见图 14。

表 1 实体关系分类表

关系类别	说明
组成部门	法院由政工科、办公室、研究室、立案庭、刑事审判庭、民事审判庭、行政审判庭、审判监督庭、执行局、技术室、法警大队、财务科和监察室共 13 个部门组成
受理	本文研究的司法案件都是已被法院受理的
种类	司法案件分为行政案件、刑事案件、民事案件、执行案件和赔偿案件共 5 个种类的案件类型
涉案	司法案件中原告、被告、委托代理人和法定代表人共 4 种涉案身份
执行类型、刑事类型、行政类型、民事类型、赔偿类型	具体的司法案件的所属案件类型
原告方、被告方、委托代理方、法定代表方	司法案件中涉案人所属的身份。

```
--nodes importdata/court.csv
--nodes importdata/department.csv
--nodes importdata/anjian.csv
--nodes importdata/anjianleixing.csv
--nodes importdata/shenfen.csv
--nodes importdata/beigao.csv
--nodes importdata/weituodailiren.csv
--nodes importdata/yuangao.csv
--nodes importdata/fadingdaibiaoren.csv
--nodes importdata/xingshi.csv
--nodes importdata/minshi.csv
--nodes importdata/xingzheng.csv
--nodes importdata/peichang.csv
--nodes importdata/zhixing.csv
```

图 14 实体节点批量导入关键代码

以同样的方式对实体关系进行分类，生成对应的实体关系 csv 文件，通过在 Neo4j 中输入实体关系批量导入代码，便能实现实体关系批量入库操作。其关键代码见图 15。

```
--relationships importdata/zuchengbumen.csv
--relationships importdata/shouliguanxi.csv
--relationships importdata/leixingguanxi.csv
--relationships importdata/sheanguanxi.csv
--relationships importdata/beigaoguanxi.csv
--relationships importdata/yuangaoguanxi.csv
--relationships importdata/xingshiguanxi.csv
--relationships importdata/minshiguanxi.csv
--relationships importdata/xingzhengguanxi.csv
--relationships importdata/peichangguanxi.csv
--relationships importdata/zhixingguanxi.csv
--relationships importdata/beigaoguanxi.csv
--relationships importdata/yuangaoguanxi.csv
--relationships importdata/fadingguanxi.csv
--relationships importdata/weituoguanxi.csv
```

图 15 实体关系批量导入关键代码

4.3 司法案件知识图谱展示

完成数据的全部导入工作后，便能够使用 Cypher 语言对生成的知识图谱进行操作以及可视化。图 16 是绘制完成的司法案件知识图谱，鉴于司法案件文本数量比较多并且可视化空间有限，该图只展示了部分实体和实体关系。

4.3.1 知识查询可视化 通过 Cypher 语言查询数据库中的数据，可将查询结果以图形化的方式展现。点击实体节点还能展示其全部的属性，从而可以便捷高效地掌握信息，也同时为司法工作提供极大便利。实体节点“杜海龙”的部分属性如图 17 所示。

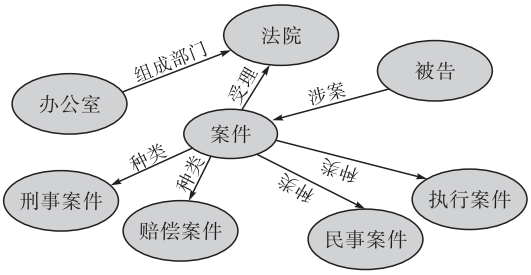


图 16 司法案件知识图谱部分展示

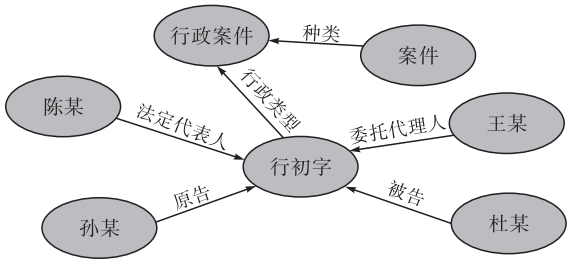


图 17 实体节点部分属性展示

4.3.2 拓展查询可视化 当使用拓展查询时，也能通过 Cypher 语句得到关于该实体的拓展信息，如查询刑事案件的拓展信息，拓展查询可视化部分结果展示见图 18。

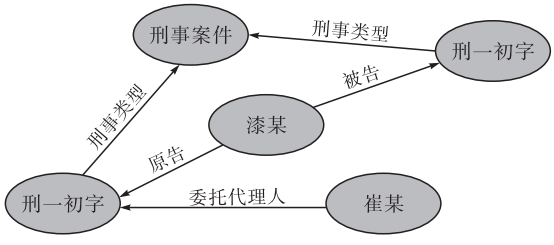


图 18 “刑事案件”拓展查询部分结果展示

5 结束语

本文首先通过对司法案件文本的知识信息数据进行归纳提取，利用 ProcessOn 软件完成司法案件的思维导图，理清了司法案件的分类关系及涉案人

在不同案件扮演的角色属性；再从实体关系的角度，通过依存句法分析算法深入不同实体之间的关系架构，利用 LTP 实现实体关系抽取，构建了实体关系三元组；最后把三元组形式的数据分类后整理成 csv 格式文件，使用 Cypher 语言批量导入到 Neo4j 图数据库，利用 Neo4j 实现司法案件知识图谱的构建与可视化，并实现了特定查询的功能。

对于司法领域而言，本文的司法案件知识图谱仅仅是一次初步的探索，其实体的属性还不够全面，对实体间关系的构建也还不算详尽，希望在以后的研究中将其不断完善更新并深入。

[参 考 文 献]

[1] 李莹，张曙光.知识图谱在学科发展分析中的应用[J].医学研究生学报，2013，26(8)：875-877.

[2] 秦长江，侯汉清.知识图谱——信息管理与知识管理的新领域[J].大学图书馆学报，2009(1)：30-37.

[3] 薛朋强.面向网络不良信息的知识图谱构建方法研究[D].乌鲁木齐：新疆大学，2017.

[4] Zhang Y，Dang Y，Hu P.Knowledge mapping for rapidly evolving domains：A design science approach[J]. Decision Support Systems，2011，50(2)：415-427.

[5] CHEN Chao-mei.Cite Space II：Detecting visualizing emerging And trends and transient patterns scientific literature[J].Journal of the American Society for Information Science and Technology，6，57(3)：359-377.

[6] 王云才，张民.基于知识图谱的国际刑事司法研究可视化析[J].上海公安高等专科学校学报，2014，24(3)：83-88.

[7] 郭琳.面向 Web 数据的知识图谱学习与推理关键技术研究[D].西安：西安邮电大学，2018.

[8] 王雅兰，朱尚明.面向科学计量分析的司法鉴定学科知识图谱构建与应用研究[J].中国司法鉴定，2017，90(1)：85-92.

[9] 吴运兵，杨帆，赖国华，等.知识图谱学习和推理研究进展[J].小型微型计算机系统，2016，37(9)：2007-2013.

[10] 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)：589-595.

[11] 盛晓昌.面向互动百科的知识抽取和知识库构建方法研究[D].杭州：浙江大学，2015.

[12] Liu Qiao，Li Yang，Duan Hong，et al. Knowledge graph construction techniques[J].Journal of Computer Research and Development，2016，53(3)：582-600.

[13] Hecht R，Jablonski S. NoSQL evaluation：A use case oriented survey[C]// International Conference on Cloud & Service Computing. 2012.

[14] Webber J. A Programmatic Introduction to Neo4j[C]. Conference on Systems，Programming，and Applications：Software for Humanity，2012：217-218.

[15] 王余蓝.图形数据库 Neo4j 的内嵌式应用研究[J].现代电子技术，2012，35(22)：36-38.

Research and Implementation of Visualization of Judicial Cases Based on Knowledge Mapping

CHEN Jianxia，HUANG Yujun，CAO Guojin，YANG Fan，LI Chao，MA Zhongbao

(*School of Computer Science，Hubei Univ. of Tech.，Wuhan* 430068，*China*)

Abstract：Aiming at the problems of scattered knowledge，incompleteness and inconvenient query in the process of handling the cases in the judicial field，a visualization analysis method of judicial cases based on knowledge mapping is proposed. Through analyzing and summarizing the collected judicial cases，the LTP language technology platform of Harbin Institute of Technology is used to process the texts of the judicial cases，word-of-speech tagging and named entity recognition，and then extract the processed texts through the dependency parsing algorithm. The extracted semantic relationship between entities is built up in the triple form，which is then entered into the Neo4j graph database to realize the construction of the judicial case knowledge map and visual displaying. Finally，the feasibility of the method is proved with an example.

Keywords：Neo4j；graph database；knowledge mapping；dependency syntax analysis

[责任编校：张岩芳]