基于图数据库的贵州省大数据政策知识建模研究*

张维冲[1,2] 王芳[1,3] 黄毅[1,2]
(1. 南开大学商学院，天津 300071；2. 中电科大数据研究院有限公司，贵阳 550081；
3. 南开大学网络社会治理研究中心，天津 300071)

摘要：新时期政策制定的科学化、智慧化、精准化，需要对大规模、碎片化的政策性公文进行知识建模与关联聚合，实现知识层面的数据融合与集成。本文以716篇贵州省大数据政策为样本数据，对大数据关键表述进行知识表示与知识抽取，并基于图数据库Neo4j对Cypher语言的知识查询、知识推理等关键技术进行研究。通过对单一关系分析、复杂关系分析、公文引文分析3类实例的验证，结果表明，该方法较好地实现了基于政策/政令多粒度知识发现的公文间关系分析与推理，为提升政策制定的系统性和科学性提供方法参考。

关键词：大数据政策；政府公文；Neo4j；知识图谱
中图分类号：G356
DOI：10.3772/j.issn.1673-2286.2020.04.005

国家治理现代化离不开政府治理现代化，建设数字政府是推进政府治理和国家治理现代化的重要途径。当前数字政府建设中，普遍存在政务数据“拥而难用、汇而不慧”的现象。虽然政务数据共享开放工作不断深入，跨部门数据流通渠道逐渐建立，但对海量政务数据仍然缺乏有效的整合分析，数据挖掘分析多停留在简单的相关性分析层面，碎片化政务数据难以转化为可供决策使用的知识和智慧[1]。政府公文，作为政务数据的一种，既是政府部门日常办公处理的重要内容，也是重要的知识资源[2]。自然语言处理、知识图谱与深度学习等技术的迅速发展，为政府公文的知识发现、管理与利用创造了基本条件。为实现新时期政策制定的科学化、智慧化、精准化，持续推进政府治理能力现代化，亟需将大规模、碎片化的政策性公文中的知识进行关联聚合，以实体为基本单位对政务数据进行挖掘分析，揭示各实体间的复杂关系，实现知识层面的数据融合与集成，更大程度地释放政策数据价值，进而为政府、企业、组织、公众提供知识服务。

然而在政策公文时空关联研究上，现有成果主要集中于处理小规模数据的信息计量方法[3-4]，以及基于词语、句子的浅层统计分析方法[5-7]等。赵洪等[8]基于卷积神经网络构建基于大规模政府公文智能处理的算法，实验结果表明有较好的性能，但未对公文间的关联关系进行分析。鉴于政府公文智能处理研究对政策文本挖掘所具有的基础指导及其应用价值，本研究将在其基础上进一步研究公共政策间的关联及聚合技术。

政策文本关联聚合表现为多维关系和多粒度信息对象的组织、关联、排序与呈现，是基于信息组织的多源信息单元融合与重组技术。在多维关系的文本关联聚合研究上，已有研究开展了基于语义关系的聚合[9]、基于引用关系的聚合[10]和基于社会关系网络的聚合[11]等。同时，面向不同的文本信息粒度，依靠内容的相似度计算，进行基于多粒度信息单元的聚合[12]及多源文本片段的信息融合[13]等。这些研究为关联文本间关系的有效揭示提供了很有价值的研究参考，但在对不同类型信息资源的解构与重组技术上也表现出较大差异，表明

*本研究得到提升政府治理能力大数据应用技术国家工程实验室2017—2018年度开放基金重点支持项目“基于NLP和深度学习的大规模政府公文智能处理技术研究”（编号：HX20180069）资助。

深度关联聚合技术研究存在较大的领域特性和策略差异，在特定的文本对象上还需进行更深入的研究。

大数据作为信息化发展的新阶段，对人类社会生产生活都产生巨大影响。把握大数据发展方向，推动大数据开发应用，发展大数据产业，对于地方经济社会发展具有十分重要的战略意义和现实意义。2013年以来，贵州省深入实施大数据战略行动，持续推动大数据探索实践，取得了显著成效，其政策制定的成功经验值得借鉴。

本文即以贵州省大数据政策为样本数据，对其涉及大数据的关键表述进行建模分析，抽取细粒度知识元组，并基于图数据库Neo4j对Cypher语言的知识查询、知识管理、知识推理等关键技术进行研究，从而实现基于政策/政令多粒度知识发现的公文间关系分析与推理，为提升政策制定的系统性和科学性提供方法参考。

1 图数据库Neo4j简介

在面临大规模知识管理需求时，需要考虑用数据库管理系统（Database Management System，DBMS）对知识进行存储。常用的数据库管理系统可分为关系型数据库管理系统（Relational DBMS）、图数据库管理系统（Graph DBMS）、RDF存储系统（RDF Stores）[14]。其中，图数据库（Graph Database）是基于图论（Graph Theory）思想和算法而实现的高效处理复杂关系网络的新型数据库系统，善于高效处理大规模、复杂、互连、多变的数据，其计算效率远远高于传统的关系型数据库[15]，在社交网络、实时推荐、征信系统、人工智能等领域被广泛应用。常见的图数据库有Neo4j、Microsoft Azure Cosmos DB、ArangoDB、OrientDB等，根据DB-Engines网站发布的图数据库使用热度[16]，近年来Neo4j一直排名居首。

Neo4j是基于Java的高性能、高可靠性、可扩展性强的开源图数据库，完全兼容ACID，即原子性（atomicity）、一致性（consistency）、隔离性（isolation）、持久性（durability）。Neo4j目前应用广泛，社区活跃，生态成熟，企业版支持高可用集群。国外的ebay、Walmart和PitneyBowes等公司均选用Neo4j图数据库，实现对企业级大数据中关系的有效处理。Neo4j社区版具有英文版[17]和简体中文版[18]两种。

基于Neo4j的图数据库模型如图1所示，Neo4j的信息建模包括节点、边和属性3种构造单元。数据库图形中的节点可以与其他任何节点建立关系，每个节点可以设置多个属性。图形中的每一个关系必须拥有一个开始节点和一个终止节点，每个关系也可以设置多个属性。

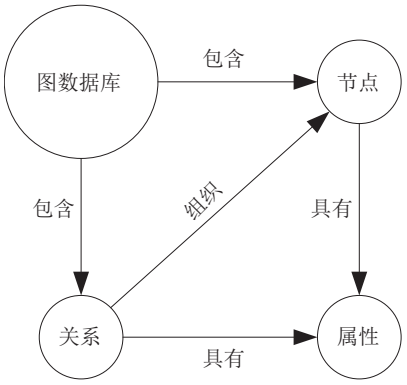


图1 基于Neo4j的图数据库模型

Neo4j的数据导入有5种方式：①Cypher语句中的CREATE命令；②Cypher语句中的LOAD CSV命令，加载CSV数据；③官方提供的Java API—Batch Inserter；④官方提供的Neo4j-import工具；⑤第三方开发的工具。Neo4j数据维护简单，每个节点对应于关系数据库中的一个记录，节点和边的属性相当于记录中的字段，属性内容和个数可以动态变化，节点之间的边也可以自由删减且不会影响已有数据结构的逻辑。Neo4j数据查询使用Cypher语言[19]，Cypher是一种声明式、表达能力强的描述性图形查询语言，主要使用的关键字有create（主要用于创建图形节点、关系及属性）、match（在已有图形数据库中匹配目标信息）、where（是match功能的条件）、return（完成匹配后，返回指定值）。

围绕Neo4j的配套应用开发也日趋成熟。例如，Neo4j Bloom可视化工具能够实现Neo4j中数据集的导航和编辑，性能上可以轻松、平滑地在普通个人电脑上显示数万节点和它们之间的关系；Neo4j函数存储包APOC（Awesome Procedure On Cypher），包含丰富的函数和存储过程，涵盖各种图论算法，是Cypher的有力补充。

2 基于Neo4j的科技政策知识建模

2.1 政策公文的知识建构

简单而言，知识是人类通过观察、学习和思考有关客观世界的各种现象而获得和总结出的所有事实、概

念、规则或原则的集合[20]。知识世界是意义世界的镜像。通观知识的复杂性，从微观的知识元、知识结构，到宏观的知识体系，知识世界的建构模型必然是一个多建模实施主体、多层次、多类型、多视角构成的“知识联结网”。政府公文领域的知识建构也是如此，由文献单元向知识单元的深入必然是一个复杂的过程。

关于知识基本构成单元的形式和概念还没有形成统一认识，知识元[21]、知识单元[22]、知识因子[23]和知识基因[24]等是主要的代表性观点。图书情报学和计算机科学是知识构成单元的主要研究领域。尽管观点不一致，但多数学者认同知识元是知识控制与处理的基本单位，是知识结构的基元。具有如下特点[25]：①知识元具有语义相对完整性，即有实际意义和相对独立性；②知识元用于表达特定的知识，如一个科学概念或一条基本原理；③知识元相对于它所表达的特定知识而言，应该是最小的、不可再拆分的；④知识元表现为具体的知识内容，在文献中表现为概念、原理、方法、定理、定律、结论等形式。宋艳辉等[26]通过重点分析知识单元研究中有争议或者理解不一致的问题后认为，知识元可以按照一定的知识关联进行自由组合，组合而成新的知识体，称为知识单元；知识单元通过知识关联组合而成新的知识单元，或者更高层级、宏观意义上的知识体系。关于知识体系如何分类，陈洪澜[27]列举了知识分类的10种方式，分别是按照知识效用分类、按照研究对象分类、按照知识属性分类、按照知识形态分类、按事物运动形式分类、按照思维特征分类、按照自然现象和社会现象分类、按照知识研究方法分类、按照知识的内在联系分类、按照学科发展趋势分类。这些方法各有长短，需要根据客观条件和主观需求灵活应用。

由以上看来，政府公文作为一种知识资源，通过知识体系建构的方式对内容进行解构是可行的，并且公文中知识建构所达到的广度、深度和精度决定了对政策知识建模的效果。本文尝试基于知识元理论构建知识表示体系，进而运用知识抽取（实体识别、关系抽取）、知识融合（知识异构、实体匹配）、知识存储（图数据库管理系统）等知识挖掘技术，对政策/政令中主体、对象、事项等知识进行关系推理。

2.2 构建过程

为实现对科技政策文本的知识建模、知识获取、知识存储与可视化图谱展示，本文基于已有数据和算法研究，构建贵州大数据政策知识图谱，具体过程如下。

（1）数据收集。在自建大规模公文数据库中以全文包含“大数据”为检索条件，以“发布机关代码=202 or 204”为限定条件（贵州省级政府及下辖部门），总计获得贵州省各级单位发布的全文中包含“大数据”的政策性公文716篇，保存为数据集data.json。该数据也可通过网络公开数据获得。

（2）“大数据”关键表述抽取。即抽取每篇公文中包含“大数据”的句子，结果保存为数据集data_keysentence.txt，核心程序段如表1所示。政策性公文是一类主题复合型文本，几乎每篇公文包含的主题都不止一项。如《省人民政府办公厅关于支持贵安新区发展若干政策措施的意见》中仅有部分语段涉及“大数据”，这时候就需要将全部这些语段抽取出来，剔除非相关的冗余语句。该步骤也是对知识粒度的第一次细化。

（3）实体抽取。Python环境下调用hanlp模块对形成的关键表述语料进行实体抽取，并人工校对，结果保存为数据集data_entity.txt。该步骤是对知识粒度的第二次细化。经过实验对比，研究发现，基于pyhanlp的enableOrganizationRecognize模块较基于pyltp的NamedEntityRecognizer模块和基于StanfordCoreNLP的ner模块，在机构识别、命名实体抽取方面的效果更优，更适合于新词发现。将抽取结果同原始数据合并，得到的数据结果样例如图2所示。

（4）关系构建。基于对实体抽取结果的内容分析，制定实体类别标签体系，同时将实体与发布机关、公文标题进行关联，构建实体、关系标签体系，标签类别如表2所示。依据标签体系分别对实体和关系进行标注，并补全实体标注信息（实体名称、实体标签）和关系标注信息（实体1标签、实体1名称、关系、实体2标签、实体2名称）。至此，贵州省大数据政策的知识获取工作完成，得到细粒度知识元组与关系对。

（5）知识存储。将上述步骤形成的实体和关系数据转化为图数据库Neo4j需要的格式[28]，并进行数据批量导入。本文采用的Neo4j版本为微云数聚研制的Neo4j简体中文版，支持数据驱动节点、连线的颜色、节点的大小、连线的粗细等不同显示，支持批量执行Cypher语句。

（6）知识分析。基于Cypher语言进行图数据库分析。

如图3所示，Neo4j类似于一个基于网页的外壳环境，其应用界面分为3个部分。①左侧功能列表，可收起为边框扩展栏。边栏扩展显示不同的功能面板，用于常

表1 “大数据”关键表述抽取及实体抽取算法核心程序段

输入：数据集data.json，内容包含编号、机构代码、中文标题、全文、发布机关、发布日期、参考文献

输出：数据集data_keysentence.txt，内容新增大数据关键表述；数据集data_entity.txt，内容新增实体抽取结果

```
mport json
import re
from pyhanlp import *
import pandas as pd
text = open（‘data/input.json’，‘r’，encoding=‘utf-8’）. read（）
sent = open（‘data/data_keysentence.txt’，‘w’，encoding=‘utf-8’）
ent = open（‘data/data_entity.txt’，‘w’，encoding=‘utf-8’）
text_json = json.loads（text）
dd = {}
i = 1
bb = []
#抽取“大数据”关键表述
for j in text_json:
    try:
        id = j[‘编号’]
        body = j[‘全文’]
        body = body.replace（‘<BR>’，‘\n’）
        s1 = []
        w1 = []
        print（id)
        if ‘大数据’ in body:
            s = re.split（r ‘[。？！…]’，body）#切割为单句
            for x in s:
                if ‘大数据’ in x:
                    s1.append（x.strip（））
        print（id，s1，file=sent)
    print（i）
i = i + 1
```

```
#调用hanlp模块
CustomDictionary=JClass（“com.hankcs.hanlp.dictionary.CustomDictionary”）
CustomDictionary.add（“大数据”）
segment = HanLP.newSegment（）.enableOrganizationRecognize（True）
#实体抽取
for j in text_json:
    try:
        id = j[‘编号’]
        body = j[‘全文’]
        body = body.replace（‘<BR>’，‘\n’）#用<BR>替换换行符\n
        s1 = []
        w1 = []
        kk = []
        print（id）
        if ‘大数据’ in body:
            s = re.split（r ‘[。？！…]’，body）#分句
            for x in s:#对每句话而言
                if ‘大数据’ in x:
                    ww = segment.seg（x.strip（））#分词
                    for term in ww:#对每个词而言
                        zz = ‘{}\t{}’.format（term.word，term.nature）
                        if ‘nt’ in zz:#如果该词的词性为nt
                            if term.word not in bb:
                                bb.append（term.word）
                            if term.word not in kk:
                                kk.append（term.word）
        if kk:#如果kk为真
            w1.append（kk）#合并kk中的词
        print（id，w1，file=ent）#保存到文件
    print（i）
    i = i + 1
```

编号	机构代码	中文标题	全文	发布机关	发布日期	参考文献	大数据关键描述	实体抽取结果	
1a201309170105	202	关于转发《国家林业局关于进一	各市（州）林业（绿化）局，贵	贵州省林业厅	2013-9-2	#N/A	[推动大数据、移动互联等新一代信息技术的集成	林业	
1a201406090089	202	贵州省商务厅关于印发《贵州省	各市（州）、贵安新区、仁怀市、	贵州省商务厅	2014-6-5	《贵州省关于加快发展电子商	[贵州省将依托贵安新区电子信息产业园和中关村	贵安新区电子信息产业	中关村贵阳科
1a201408250314	202	贵州省大数据产业发展领导小组	各市、自治州人民政府，贵安新区	贵州省大数据产业发展	2014-7-10	《关于加快大数据产业发展应	[大数据是大产业、大机遇、大红利，发展大数据	云上贵州·大数据国际	大数据产业创
1a201505112264	202	省人民政府印发《关于加快大数	各市、自治州人民政府，贵安新区	贵州省人民政府	2014-2-25	《贵州省信息基础设施条例》	[现将《关于加快大数据产业发展应用若干政策的	大数据管理平台	大数据交换与
1a201505112268	202	省人民政府关于加快商贸流通业	各市、自治州人民政府，贵安新区	贵州省人民政府	2014-3-20	《贵州省现代流通业发展规划	[积极打造电子商务大数据云计算产业，结合我省	电子商务大数据云计算	商贸流通业
1a201505112274	202	省人民政府关于支持工业企业加	各市、自治州人民政府，贵安新区	贵州省人民政府	2014-5-22		[并加挂大数据产业办公室牌子]	大数据产业办公室	
1a201505112276	202	省人民政府关于加快现代服务业	各市、自治州人民政府，贵安新区	贵州省人民政府	2014-6-4	《贵州省现代服务业集聚区发	[把我省建设成为西部地区重要的金融商务会展中	大数据产业基地	大健康产业基
1a201505112280	202	省人民政府关于贯彻落实国务科	各市、自治州人民政府，贵安新区	贵州省人民政府	2014-7-16	1a201303080077;1a200804150	[在能源、资源深加工、特色轻工业、以大数据为	大数据信息产业	
1a201505112281	202	省人民政府印发《贵州省关于加	各市、自治州人民政府，贵安新区	贵州省人民政府	2014-8-1	《中共贵州省委关于进一步实	[推进医药服务与大数据、医药流通与物联网结合	医药服务大数据	新医药产业
1a201505112282	202	省人民政府关于印发贵州省社会	各市、自治州人民政府，贵安新区	贵州省人民政府	2014-8-21	1a201301180128;1a200711270	[要以大数据为引领，充分利用大数据云服务，利	大数据云服务	信用体系建设
1a201505112288	202	省人民政府关于印发贵州省信息	各市、自治州人民政府，贵安新区	贵州省人民政府	2014-11-21	1a201405090197	[推动以大数据为引领的信息产业发展水平和信息	大数据信息产业	大数据开发利
1a201505112293	202	省人民政府印发《关于支持健康	各市、自治州人民政府，贵安新区	贵州省人民政府	2015-2-26	1a201108052096;《中共贵州省	[融合医疗服务、大数据信息服务、健康管理和促	大数据健康信息服务	健康养生产业
1a201505112430	202	省人民政府办公厅关于支持贵安	各市、自治州人民政府，贵安新区	贵州省人民政府办公厅	2014-9-29	《贵安新区规划技术导则》；《	[大力发展大数据产业、先进制造业、高端服务业	贵安新区布局大数据产	全省大数据
1a201505112441	202	省人民政府办公厅关于加强政府	各市、自治州人民政府，贵安新区	贵州省人民政府办公厅	2014-12-30	1a201412010046	[省大数据产业发展领导小组办公室牵头制定“云	“云上贵州”平台	网站建设
1a201505112446	202	省人民政府办公厅关于加快发展	各市、自治州人民政府，贵安新区	贵州省人民政府办公厅	2015-1-15	《省人民政府关于加快现代服	[打造物流网络体系；着力推动物流业与大数据融	物流业	
1a201505112453	202	省人民政府办公厅关于进一步加	各市（自治州）人民政府，贵安新	贵州省人民政府办公厅	2015-3-16	1a201406091481	[努力将物联网、大数据、地理信息、移动通信、	林业	
1a201505111875	202	省人民政府办公厅关于印发贵州	各市、自治州人民政府，贵安新区	贵州省人民政府办公厅	2015-3-31	1a201408060020;1a201505112	[面向综合型生产性服务业聚集带、现代物流集散	服务业	物流业
1a201505111877	202	省人民政府办公厅关于印发贵州	各市、自治州人民政府，贵安新区	贵州省人民政府办公厅	2015-3-27	《《贵州省新医药产业发展规	[结合大数据产业，推进医疗、预防、保健及管理	贵阳医学院大健康学院	贵州省医疗工
1a201505111884	202	省人民政府办公厅关于印发《贵	各市、自治州人民政府，贵安新区	贵州省人民政府办公厅	2015-4-29	1a201402110554	[加强大数据体系建设]	大数据体系建设	
1a201505111821	202	省人民政府办公厅转发省发展改	各市、自治州人民政府，贵安新区	贵州省人民政府办公厅	2014-5-16	1a201402180191;1a201404090	[重点开展大数据、医药、文化旅游、节能环保等	大数据产业	
1a201505111830	202	省人民政府办公厅关于转发省科	各市、自治州人民政府，贵安新区	贵州省人民政府办公厅	2014-7-10	《科技部办公厅、中央组织部	[依托贵州大数据基地和云工程]	贵州大数据基地	
1a201506020057	204	市人民政府印发关于加快发展大	各区、市、县人民政府，高新开发	贵阳市人民政府	2014-5-19	《贵州省大数据产业发展应用	[《关于加快发展大数据产业的实施意见》已经市	大数据产业	贵安新区大数
1a201505290182	204	市人民政府关于印发全国林业信	各区、市、县人民政府，高新开发	贵阳市人民政府	2015-4-30	《国家林业局关于启动第二批	[依托贵阳大数据和云平台建设，通过平台的大数	林业大数据服务	
1a201505290183	204	市人民政府办公厅印发《贵阳市	各区、市、县人民政府，高新开发	贵阳市人民政府办公厅	2015-3-24	《贵州省现代服务业集聚区发	[加快推进贵阳多彩贵州城、花溪青岩特色优势服	现代服务业	贵阳多彩贵州
1a201506020065	204	市人民政府印发关于贯彻落实省	各区、市、县人民政府，市政府有	贵阳市人民政府	2014-10-17	《贵阳游客满意度服务质量评	[完成全市A级景区门禁系统的安装、改造；初步完	旅游大数据运营中心	
1a201506020066	204	市人民政府关于印发贵阳市《社	各区、市、县人民政府，高新开发	贵阳市人民政府	2014-10-13	1a201407011772;《省人民政府	[要以大数据为引领，充分利用大数据云服务，利	大数据云服务	信用体系建设
1a201506020115	204	市人民政府办公厅关于印发贵阳	各区、市、县人民政府，高新开发	贵阳市人民政府办公厅	2014-9-10	《中共贵州省委办公厅、贵州	[承担制造业信息化及大数据产业发展相关工作；	制造业	
1a201506020237	204	市人民政府办公厅关于切实用好	各区、市、县人民政府，高新开发	贵阳市人民政府办公厅	2014-8-14	1a201505112415;1a201204280	[各地可围绕“5个100工程”、扶贫生态移民工程	大数据建设	
1a201506020098	204	市人民政府办公厅关于印发贵阳	各区、市、县人民政府，高新开发	贵阳市人民政府办公厅	2014-5-14	1a201308190020;1a201308140	[《贵阳大数据产业行动计划》已经市人民政府研	中关村贵阳科技园和贵	工信部中国软
1a201506020111	204	市人民政府办公厅关于印发大数	各区、市、县人民政府，高新开发	贵阳市人民政府办公厅	2014-9-2	《关于抓好下半年重点工作确	[（四）狠抓大数据产业。1.市工业和信息化委要	大数据产业	

图2 数据处理结果样例

表2 知识抽取标签体系

内 容	标签类别
实体	产业、产业园、基地、应用平台、重要开放窗口、组织机构、资金支持、研发组织机构、活动、重要发展平台、机制、工程、大数据应用、项目、领域、集聚区、示范区、战略定位、第三方权威组织机构、大数据产业空间布局、专项规划、试验区、创新创业载体、基础设施、专项行动、政策标题、政策发布机关
关系	政策发布机关→政策支持→产业；政策发布机关→鼓励发展→产业园；政策发布机关→鼓励发展→基地；政策发布机关→政策支持→应用平台；政策发布机关→鼓励→重要开放窗口；政策发布机关→政策涉及→组织机构；政策发布机关→政策支持→资金；政策发布机关→政策涉及→研发组织机构；政策发布机关→政策支持→活动；政策发布机关→鼓励发展→重要发展平台；政策发布机关→搭建→机制；政策发布机关→政策支持→工程；政策发布机关→政策鼓励→大数据应用；政策发布机关→政策支持→项目；政策发布机关→政策支持→领域；政策发布机关→政策支持→集聚区；政策发布机关→政策支持→示范区；政策发布机关→战略定位→规划；政策发布机关→政策支持→第三方权威组织机构；政策发布机关→战略布局→大数据产业空间布局；政策发布机关→发布→专项规划；政策发布机关→政策支持→试验区；政策发布机关→政策支持→创新创业载体；政策发布机关→政策支持→基础设施；政策发布机关→政策支持→中国·贵阳大数据旅游研究实训基地；政策发布机关→举行→专项行动；政策发布机关→发布政策→政策标题

见的查询和信息。研究通过构建“贵州大数据”专项选栏和下拉菜单，可实现快捷查看“战略定位”“政策支持产业”“政策支持项目”“政策支持应用平台”“政策支持示范区”等知识图谱。②右上方区域为编辑器。编辑器是输入和运行命令的主要接口，通过输入Cypher查询来处理图数据。③右下方区域为分析结果区，按操作顺序滚动呈现，点击图谱中任意节点即可查看节点信息，图查询后可导出表格结果和可视化结果。

3 建模实例

前述对贵州大数据政策/政令中的知识进行了表示、抽取与链接，本节对该建模方法进行实例验证，基于Neo4j图数据查询实现知识图谱生成和知识发现，包括单一关系分析、复杂关系分析和公文引文分析。

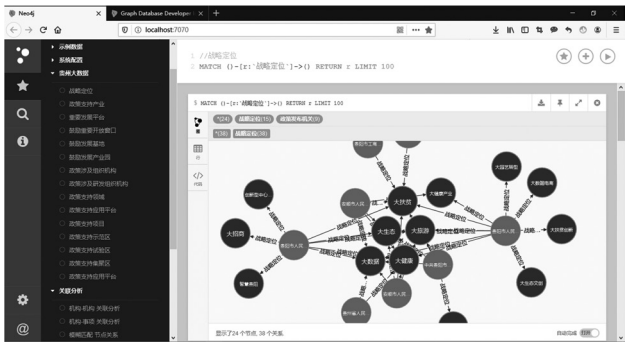


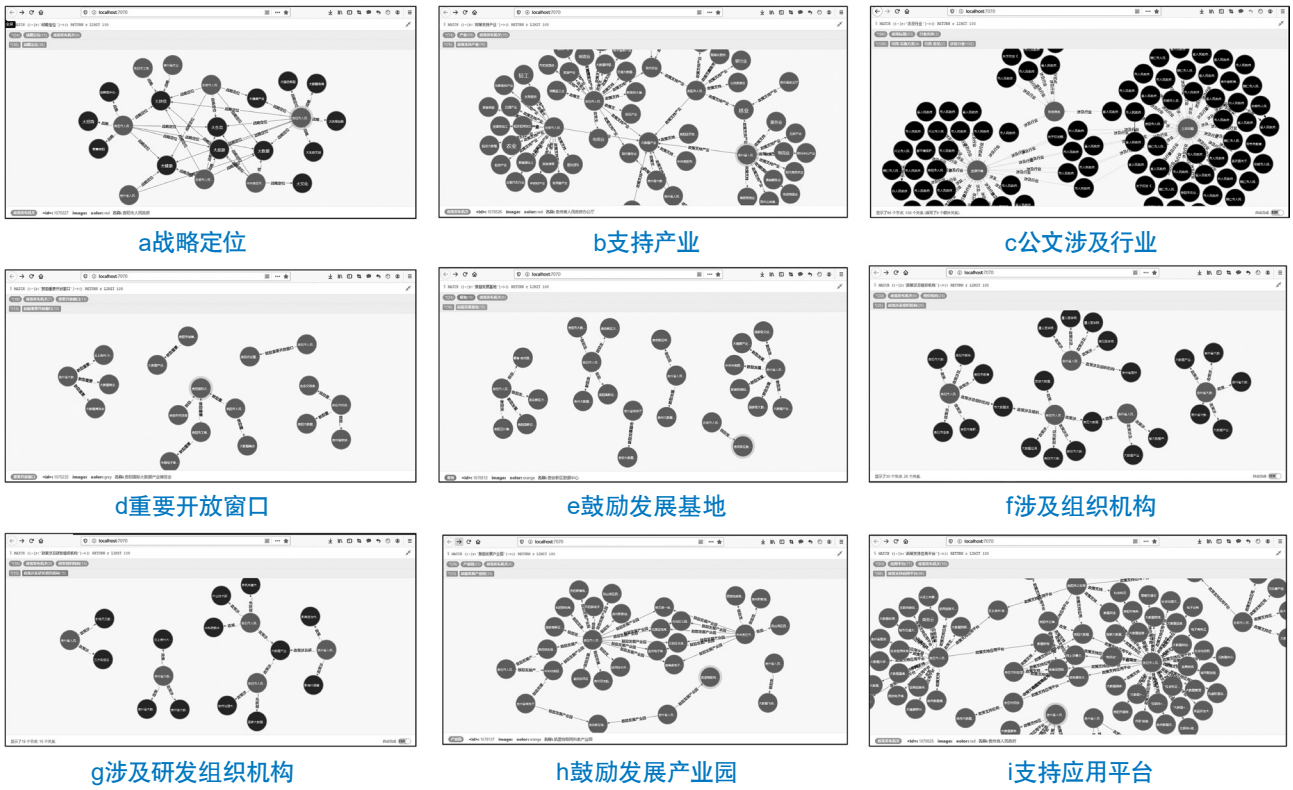
图3 Neo4j应用界面

3.1 单一关系分析

图4为9种单一关系分析实现的效果样例。实现方法为，编辑器输入Cypher查询语言“MATCH ( ) -[r:`战略定位|政策支持产业|涉及行业|鼓励重要开放窗口|鼓励发展基地|政策涉及组织机构|政策涉及研发组织机构|鼓励发展产业园|政策支持应用平台`]-> ( ) RETURN r LIMIT 100”。其中，“( )”为节点信息，“[]”内为关系信息，通过LIMIT后面的数字可限制节点的数量。研究还对大数据政策涉及的基地、资金支持、活动、重要发展平台、机制、工程、大数据应用、项目、领域、集聚区、示范区、第三方权威组织机构、大数据产业空间布局、专项规划、试验区、基础设施、专项行动等单一映射关系进行了分析。

从政策关联结果来看，贵州省各级政府高度重视大数据发展，发布的各类政策叠加效应明显，为加速资源集聚、推动大数据产业发展提供了充分保障。各政府部门结合当地实情与发展需要，深入推进大数据、大扶贫、大生态、大健康、大旅游、大文化、大招商、大生态文创等战略行动（关联关系如图4a），优化资源配置、强化技术支撑、创新发展模式，大力发展电子信息、电子商务大数据云计算、互联网金融、移动电子商务、大数据网络信息安全等大数据重点产业，以及中药材、食用菌、吊瓜、消费品工业、林业、煤炭、石材、现代高效农业、呼叫中心、白酒等地方性产业，通过大数据的发展支持航空、生态农业、医药制造业、旅游业、民族药材、磷化工、新型建筑材料、电子元器件、有机农产品、高端装备、新能源化工等相关产业发展，形成完整的大数据相关产业体系（见图4b和图4c）。

贵州省充分利用生态文明贵阳国际论坛、云上贵州·大数据国际年会、大数据博览会、大数据商业模式



a战略定位 b支持产业 c公文涉及行业

d重要开放窗口 e鼓励发展基地 f涉及组织机构

g涉及研发组织机构 h鼓励发展产业园 i支持应用平台

图4 单一关系分析实现效果样例

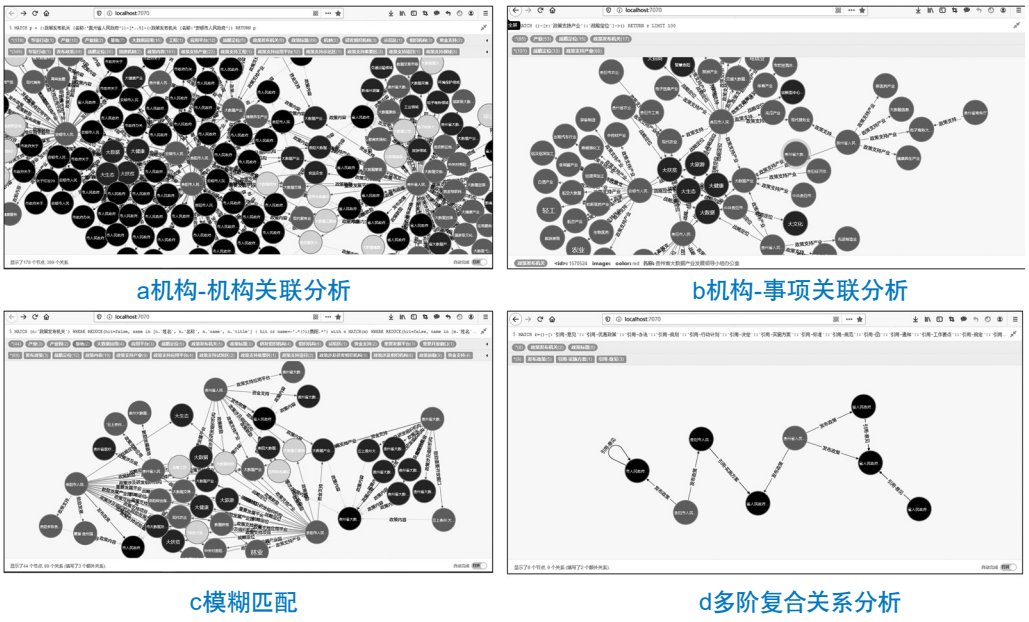
大赛、中国电子商务创新发展大会、贵州省旅游发展大会等重要开放窗口的作用（关联关系见图4d），举办形式多样的交流、展览、招商等活动，聚集优质发展要素，推进国际化进程。支持鼓励发展大数据发展基地，包括大数据产业基地、国家级大数据基地、国家级文化和科技融合示范基地、中关村贵阳科技园大数据基地、教育实践和培训基地、贵安新区大数据存储基地、贵阳高新云平台应用基地、贵州大数据综合试验区实验基地、惠普·贵州国际金贸云基地、贵阳高新云计算基地、贵阳云计算大数据创新孵化基地、贵安大数据中心、贵安新区数据中心等（关联关系见图4e）；支持发展大数据飞地产业园区、凯里物联网科教产业园、贵安新区电子信息产业园、富士康产业园、三大运营商数据中心、贵安新区节能环保产业园、贵阳跨境电子商务产业园等大数据产业园（关联关系见图4h）。作为贵州省会，贵阳市重点规划建立了“一轴两基地多园”的大数据产业空间布局。贵州省不断完善大数据系统应用，深入推进应用平台建设，包括数据铁笼、社会和云、数据民生、城市交通大数据、农业大数据信息管理、支撑数据采集系统、数据共享开放平台、数据增值应用平台、互联网+医疗健康数据、“云上贵州”系统、旅游块数据、互联网数据交换、互联网数据中心、电子政务网络、综合治税应用平台等（关联关系见图4i）。

3.2 复杂关系分析

常规关系分析方法有聚类分析、多维尺度分析、中心度分析等，但这些方法不足以展现大规模实体间的复杂关系，且不具备推理功能、因果关系分析功能。而图数据库有望弥补这些不足。Neo4j不仅能实现对众多客观实体的管理，还可以进行实体间复杂关系的查询与推理，支持逻辑语言查询与面向约束的推理。图5为4种复杂关系分析的实现效果样例。

图5a为机构-机构关联分析，查询贵州省人民政府和安顺市人民政府有哪些政策相互关联。实现语句为“MATCH p =（：政策发布机关 {名称：‘贵州省人民政府’}）-[*..3]-（：政策发布机关 {名称：‘安顺市人民政府’}）RETURN p”。其中，“[*..3]”限定关系不超过3阶。

图5b为机构-事项关联分析，查询贵州省人民政府的战略定位和支持的产业。实现语句为“MATCH（：政策发布机关 {名称：‘贵州省人民政府’}）-[r：‘政策支持产业’|：‘战略定位’]->（）RETURN r LIMIT 100”。



a机构-机构关联分析　　b机构-事项关联分析

c模糊匹配　　d多阶复合关系分析

图5 复杂关系分析实现效果样例

图5c为模糊匹配，检索名称中包含“贵阳”或“贵州”的政策发布机关。实现语句为“MATCH（n：‘政策发布机关’）WHERE REDUCE（hit=false，name in [n.‘姓名’，n.‘名称’，n.‘name’，n.‘title’] | hit or name=~ ‘.*（?i）贵阳.*’ ）with n MATCH（m）WHERE REDUCE（hit=false，name in [m.‘姓名’，m.‘名称’，m.‘name’，m.‘title’] | hit or name=~ ‘.*（?i）贵州.*’ ）with n，m MATCH r=allShortestPaths（（n）-[*]-（m））RETURN r LIMIT 100”。

图5d为多阶复合关系分析，即检索连续3阶以上的关联关系。实现语句为“MATCH r=（）-[：‘引用’]->（）<-[：‘发布政策’]-（）RETURN r LIMIT 50”。

通过对以上复杂关系的实验研究表明，基于Neo4j的政策多粒度知识关联聚合具有以下优势：①高效的信息检索能力；②能够通过节点、关系、属性值的设定实现很好的聚类性能和分类精度；③比传统的关系型数据库更适合处理大规模关联数据，特别适合多跳查询、计算最短路径、推理分析等特殊网络分析需求；④图数据库基于图论基础，支持路径搜索、中心性、社群发现等图论算法，对社会网络分析（SNA）的相关研究视角和方法依然适用，包括中心性分析，二元组分析、三元组分析，块模型分析、社群分析，凝聚子群分析，结构均衡分析，网络同质性分析，动态演化分析，统计网络分析等；⑤有望通过关联网络探究因果集群，进而探究多要素的因果关系；⑥性能上，实时速度每秒达数以百万跳，能够满足大规模图数据的运算[29]。

3.3 公文引文分析

政策性公文同学术论文一样，也存在广泛的援引、引用情况。公文引文分析通过公文的引用和被引等流动过程，揭示政策的引用现象、过程和规律，具有较强的实践价值和方法论意义。目前相关研究多为定性描述，缺少量化依据。李江等[30]深入探讨了适用于量化评价政策主题变迁、政策主体合作模式、政策体系的结构与演进的“政策文献计量”方法。刘晓光等[31]以“中央一号文件”为案例进行了引文量化分析，进一步指出公共政策研究在解决综合、复杂的现实问题时，仍面临不小的挑战。随着政策资源库的建设和语义解析技术的提升，大规模政府公文的引文关系自动解析成为可能，为通过公文引文关系探究政策演化变迁、政策扩散规律、政策影响力、府际关系等提供了新路径。

本文基于贵州大数据政策数据，按照被引用公文的类别，对引用关系进行了细分。生成的政府公文引用关系图谱如图6所示。实现语句为“MATCH（）-[r：‘引用-意见’|：‘引用-优惠政策’|：‘引用-办法’|：‘引用-规划’|：‘引用-行动计划’|：‘引用-决定’|：‘引用-实施方案’|：‘引用-标准’|：‘引用-规范’|：‘引用-函’|：‘引用-通知’|：‘引用-工作要点’|：‘引用-规定’|：‘引用-标准规范’|：‘引用-工作制度’|：

‘引用-制度’|:‘引用-计划’|:‘引用-预案’|:‘引用-指南’|:‘引用-通告’|:‘引用-批复’|:‘引用-公告’|:‘引用-法律’|:‘引用-工作规则’|:‘引用-意见办法’|:‘引用-决定规范’|:‘引用-行动计划规范’]->（）RETURN r LIMIT 100”。其中，各种引用类别可以任意组合。

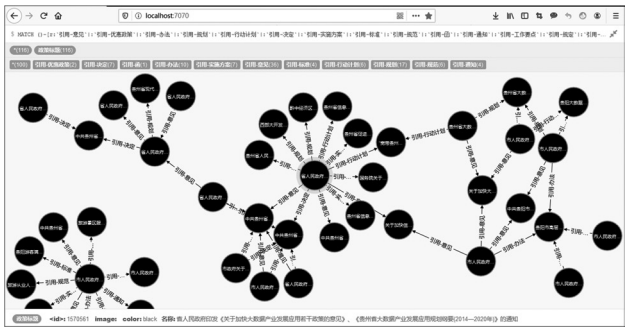


图6 公文引文分析实现效果样例

从方法的实现效果来看，基于知识建模的公文引文分析能够较好实现引文网络图谱的构建，通过Neo4j可视化图谱可以按照实际需求动态、清晰地呈现公文间的引用关系、关联网络，适合辅助探究综合、复杂问题。如根据图6样例分析可知，贵州省人民政府印发《关于加快大数据产业发展应用若干政策的意见》《贵州省大数据产业发展应用规划纲要（2014—2020年）》《省人民政府办公厅关于印发贵州省医药产业 健康养生产业发展任务清单的通知》等公文被引用率明显较高，与其他公文的关联关系较多，并且公文间的引用关系错综复杂，并非单纯的上下级公文引用。

文献的引用是文献价值、重要性及影响力的指标，而政策性公文作为一类由党政机关在行使职权或实施管理过程中形成的具有法定效用的文件材料，其相互引用必然具有更多的现实意义。文献计量学中的影响力评价（影响因子）、文献老化规律（半衰期）、共被引分析、引文耦合分析等理论与方法对于公文引文分析是否适合有待进一步验证，能否指导公文引文分析需要继续探索。未来公文引文分析有望不断丰富政策文献计量方法体系。

4 结语

本文通过对716篇贵州省大数据政策的知识建模，抽取细粒度知识元组，并基于图数据库Neo4j实现对政策/政令的多粒度知识发现和关联聚合，经过单一关系分析、复杂关系分析、公文引文分析3类实例验证，结果表明，Neo4j可较好地实现基于政策/政令多粒度知识发现的公文间关系分析与推理，本文所提方法为提升政策制定的系统性和科学性提供参考。基于图数据库Neo4j的政策多粒度知识关联聚合具有高效的检索能力、大规模关联数据处理能力、较好的聚类性能和分类精度、满足网络分析需求等优势，在探究公共政策的知识发现、演化变迁、扩散规律、府际关系等方面具有广阔的应用前景。

参考文献

[1] 李军，乔立民，王加强，等. 智慧政务框架下大数据共享的实现与应用研究 [J]. 电子政务，2019（2）：34-44.

[2] 赵国俊. 电子政务教程 [M]. 北京：中国人民大学出版社，2004.

[3] 汪波，李坤. 国家养老政策计量分析：主题、态势与发展 [J]. 中国行政管理，2018（4）：105-110.

[4] 刘亚亚，曲婉，冯海红. 中国大数据政策体系演化研究 [J]. 科研管理，2019，40（5）：13-23.

[5] 裴雷，李向举，谢添轩，等. 中国信息政策研究主题的历时演进特征（1986—2015年）[J]. 数字图书馆论坛，2016（7）：19-27.

[6] 刘刚，傅玮萍，马莺歌. 基于语义的政策血缘网络演化机理研究 [J]. 中文信息学报，2018，32（5）：114-127.

[7] 叶江峰，任浩，甄杰. 中国国家级产业园区30年发展政策的主题与演变 [J]. 科学学研究，2015，33（11）：1634-1640.

[8] 赵洪，王芳，王晓宇，等. 基于大规模政府公文智能处理的知识发现及应用研究 [J]. 情报学报，2018（8）：805-812.

[9] CHEN Z，GANGOPADHYAY A，HOLDEN S H，et al. Semantic integration of government data for water quality management [J]. Government Information Quarterly，2007，24（4）：716-735.

[10] DING Y，ZHANG G，CHAMBERS T，et al. Content-based citation analysis：the next generation of citation analysis [J]. Journal of the Association for Information Science and Technology，2014，65（9）：1820-1833.

[11] DIESNER J. From texts to networks：Detecting and managing the impact of methodological choices for extracting network data from text data [J]. KI-Kunstliche Intelligenz，2013，27（1）：75-78.

[12] ZHANG L. Grasping the structure of journal articles：Utilizing the functions of information units [J]. Journal of the American Society for Information Science and Technology，2012，63

（3）：469-480.

[13] KAPTEIN R，MARX M. Focused retrieval and result aggregation with political data [J]. Information Retrieval，2010，13（5）：412-433.

[14] 王鑫，邹磊，王朝坤，等. 知识图谱数据管理研究综述 [J]. 软件学报，2019，30（7）：2139-2174.

[15] 张帜. Neo4j权威指南 [M]. 北京：清华大学出版社，2017.

[16] DB-Engines. DB-Engines Ranking of Graph DBMS [EB/OL]. [2020-02-16]. https://db-engines.com/en/ranking/graph+dbms.

[17] Neo J Inc. Neo4j下载 [EB/OL]. [2020-01-15]. https://neo4j.com/download-center/#releases.

[18] 微云数聚. Neo4j简体中文版下载 [EB/OL]. [2020-01-15]. http：//we-yun.com/.

[19] 李雪. 一种基于Neo4j图数据库的模糊查询研究与实现 [J]. 计算机技术与发展，2018，28（11）：16-21.

[20] 中国电子技术标准化研究院. 知识图谱标准化白皮书 [R/OL]. [2020-01-15]. http://www.cesi.ac.cn/images/editor/20190911/20190911095208624.pdf.

[21] CHANG X，ZHENG Q. Knowledge element extraction for knowledge-based learning resources organization [J]. Advances in Web Based Learning Icwl，2007，4823：102-113.

[22] 王子舟，王碧滢. 知识的基本组分——文献单元和知识单元 [J]. 中国图书馆学报，2003（1）：4-10.

[23] 蒋永福，李景正. 论知识组织方法 [J]. 中国图书馆学报，2001（1）：3-7.

[24] 刘植惠. 知识基因探索（六）知识基因原理在情报分析研究中的应用 [J]. 情报理论与实践，1998（6）：63-65.

[25] 索传军，盖双双. 知识元的内涵、结构与描述模型研究 [J]. 中国图书馆学报，2018，44（4）：54-72.

[26] 宋艳辉，王小平. 从文献单元、信息单元向知识单元的嬗变 [J]. 科研管理，2015（S1）：459-464.

[27] 陈洪澜. 论知识分类的十大方式 [J]. 科学学研究，2007（1）：26-31.

[28] Beijing We-Yun Data Co. L. To Neo4j-The Smart Importer for Neo4j [EB/OL]. [2020-01-17]. http://we-yun.com:8000/ToNeo4j/.

[29] 约翰·斯科特. 社会网络分析法：第2版 [M]. 重庆：重庆大学出版社，2007.

[30] 李江，刘源浩，黄萃，等. 用文献计量研究重塑政策文本数据分析——政策文献计量的起源、迁移与方法创新 [J]. 公共管理学报，2015，12（2）：138-144.

[31] 刘晓光，王贤文. 公共政策的学术影响计量——基于2004—2017年“中央一号文件”的引文分析 [J]. 情报杂志，2019，38（8）：165-171.

作者简介

张维冲，男，1991年生，博士研究生，研究方向：知识发现、科学计量学。
王芳，女，1970年生，博士，教授，通信作者，研究方向：电子政务、网络社会、数据治理，E-mail：wangfangnk@nankai.edu.cn。
黄毅，男，1986年生，博士研究生，研究方向：机器学习、知识发现。

Knowledge Modeling of Big Data Policy in Guizhou Province Based on Graph Database

ZHANG WeiChong[1,2] WANG Fang[1,3] HUANG Yi[1,2]
(1. Business School, Nankai University, Tianjin 300071, China; 2. CEC Data Research Institute Co., Ltd. Guiyang 550081, China; 3. The Center for Network Society Governance, Nankai University, Tianjin 300071, China)

Abstract: In view of the scientific, intelligent and precise policy making, it is necessary to achieve knowledge modeling and association aggregation of large-scale and fragmented policy documents. This article took 716 big data policies issued by Guizhou Province as sample data, carried out knowledge representation and knowledge extraction for key sentences, and studied key technologies such as knowledge query and knowledge reasoning based on the graph database Neo4j. Through the verification of single-relation analysis, complex-relation analysis, and governmental document citation analysis, the results showed that the method realized the analysis and reasoning of the relations between documents based on policy multi-granularity knowledge discovery, and provides a reference for improving the systematicness and scientificity of policy making.

Keywords: Big Data Policy; Governmental Document; Neo4j; Knowledge Graph

（收稿日期：2020-01-12）