科技政策知识图谱构建研究*

张雨 吴俊
（北京邮电大学经济管理学院，北京 100876）

摘要：为助力广大中小企业快速查新，亟需使用人工智能手段对科技政策文本知识建模，构建基于知识图谱的结构化查询。本研究以采集到的全国26 660条科技政策文本为数据源，首先构建科技政策知识本体，之后通过Bi-LSTM深度学习模型完成三元组抽取，最后应用Neo4j图数据库完成知识存储与图谱化检索。所构建的科技政策知识图谱共有4万余个实体节点、15万余条关系，能够实现不同细粒度政策实体和关系的关联查询与可视化。这种基于科技政策本体构建政策知识图谱的方法既拓展了垂直领域知识图谱的新思路，也为开拓基于互联网的科技政策智能问答奠定基础。

关键词：知识图谱；科技政策；本体；Bi-LSTM
中图分类号：G353　DOI：10.3772/j.issn.1673-2286.2021.08.005
引文格式：张雨，吴俊. 科技政策知识图谱构建研究[J]. 数字图书馆论坛，2021 (8)：31-38.

近年来国家高度重视科学技术发展，为鼓励“政用产学研”协同创新，从中央到地方出台了一系列科技政策。一方面，这些政策散布于政府及媒体网站，不方便用户集中查阅和整合分析；另一方面，政策内容包罗万象，涉及诸多行业，公文化的表达方式不利于企业或个人快速检索，把握不同政策间的内在联系[1]。如何准确而快速地挖掘科技政策文本的关键语义信息，以结构化形态展现科技政策主要条目的层级关系，以可视化方式展现多个政策主体以及政策属性特征间的不同关系，成为学术界和企业界亟待解决的问题。这一问题的解决不仅有利于垂直领域的知识发现与利用，也有助于广大中小企业及时跟踪了解各级政府的科技创新政策动向。

科技政策文本的语义挖掘与知识结构化解析可以借助知识图谱技术解决。知识图谱旨在通过提取知识实体及实体间关系，将原始的文本数据解析为表征知识本体的语义属性及脉络关系，进而以图网络形态帮助用户快速理解知识结构，揭示领域知识的特征和规律[2]。将知识图谱技术应用于科技政策的知识序化，展现政策内容与政策主体间的关系，揭示不同政策之间的内在联系，提升政策文本的使用价值，解决用户长期以来反映的“拥而难用、汇而不慧”难题。此外，随着深度学习技术的快速发展，通过深度神经网络模型智能识别与提取实体间关系也为领域知识图谱的专业化和精细化提供了新手段。

具体而言，本文以采集到的各级政府公开发布的科技政策文本为数据源，构建政策知识本体，应用Bi-LSTM模型抽取政策文本实体及属性特征，使用Neo4j图数据库构建科技政策知识图谱并实现可视化查询，以提升科技政策的利用效率，更大程度发挥科技政策的效用。

1 相关研究

1.1 知识图谱

知识图谱较早由Google基于语义网研究提出，旨在实现语义搜索的智能化，提升用户对知识的搜索质量与体验[3]。知识图谱主要由<实体，关系，实体>和<实体，属性，属性值>三元组构成，优点在于构建的语义

* 本研究得到国家重点研发计划项目“基于模式创新的科技咨询服务平台研发与应用示范”（编号：2018YFB1403600）资助。

知识库以图形化形式展示现实世界中的实体及其相互关系。

随着知识图谱相关技术的快速发展，在通用知识图谱之外，众多行业领域知识图谱逐渐兴起。一般而言，通用知识图谱以常识性知识为对象，以大规模开源知识为支撑，构建广域语义知识库，主要应用在智能搜索领域，知名的通用知识图谱开源库有FreeBase、DBPedia、Wikidata等。行业知识图谱则面向垂直行业，以从特定领域采集的文本信息为支撑，聚焦定域的语义知识库，具有鲜明的行业应用特征，对专业性与准确度要求更高[4]。此外，行业知识图谱更加强调领域知识的有序化、结构化和可视化，以提高管理决策效率为主要目标。以金融股权知识图谱[5]为例，它从股权角度出发，通过股权穿透式查询，可从全局实现风险识别，通过持股比例判断机构风险水平，为企业风险识别与预测提供新方法。在图书情报领域，白如江等[6]提出科学事件元数据模型，以文献摘要为挖掘对象，构建科学事件知识图谱。在医疗领域，曹明宇等[7]构建肝癌知识图谱，并进一步设计了肝癌知识问答系统，能够有效回答肝癌相关的疾病症状、治疗药物及治疗手段等问题。

作为智能互联时代知识优化和推荐的重要手段，知识图谱技术已成为学术界和工业界研究的焦点，被广泛应用于个性化推荐、语义搜索、智能问答、风险识别及预警等领域。

1.2 科技政策

科技政策是政府为促进科学技术发展以及利用科学技术为国家目标服务而采取的集中性和协调性措施，是科学技术与国家发展的有机整合[8]。随着各级政府科技投入的加大，科技政策引领科技发展的作用日益凸显。相应的，科技政策的相关研究也呈现内容多元、方法多样的特点。

已有的科技政策研究方法大致可分为3类。第一类是利用学者Rothwell等[9]提出的政策工具法进行定量研究。徐硼等[10]基于政策工具视角深入剖析了我国科技创新政策，指出3种政策工具在应用层面存在结构失衡的问题，并对未来科技政策制定提出了改进策略。针对我国科技政策间协调性差，政策体系不完善的问题，仲伟俊等[11]基于政策工具分析框架对科技政策进行具体分析，总结现有科技政策的合理性和不足，探讨完善政策的路径。第二类是利用文本挖掘方法对科技政策文本内容进行词频和语义分析。例如：宋伟等[12]以地方政府发布的人工智能科技政策文本为对象，通过文本语义分析，指出人工智能政策主题存在群聚化特点；祝鑫梅等[13]对1979—2017年国家层面的245篇政策文本进行分析，对政策文本的高频主题词进行可视化，揭示了政策主题循环往复、螺旋上升的演化过程。第三类方法是将文献计量法应用于科技政策文献的量化研究，识别并发现政策文献的知识分布与演化等规律。如黄萃等[14]以1949—2010年中国科技政策数据为研究对象，绘制我国科技政策主题词的聚类图，展示中国科技政策的主题热点与演化路径，进一步总结中国政府执政理念的变化。

1.3 科技政策的知识图谱研究

由于科技政策文本数量日趋庞大，语义关系日渐繁杂，将知识图谱技术用于科技政策领域，以实现政策主体、政策属性与关系的结构化和显性化越来越重要。既有的研究呈现两种特点。一是聚焦科技政策领域的文献研究，主要是应用Citespace等可视化工具构建领域知识图谱，展示领域研究的发展脉络和热点动向，用以预测前沿趋势，推动科技政策制定过程的科学性和规范性。例如：李梅芳等[15]以Research Policy期刊1974—2016年发表的2 855篇文献为研究对象，利用Citespace软件绘制文献共被引知识图谱以揭示科技政策领域国际研究的演化情况；赵绘存等[16]在李梅芳研究的基础上，对2007—2017年*Research Policy*发表的文章进行分析，通过VOSviewer软件构建作者网络共现关系图谱和国家合作网络图谱，发现科技政策研究个人合作强度不足，但国家合作网络联系紧密的特点。二是利用自然语言处理开源工具从政策文本中抽取知识实体与关系。张维冲等[17]利用HanLP等工具对716篇贵州省大数据政策文本进行实体抽取，构建大数据政策图谱；Wang等[18]利用正则表达式提取政策实体及属性，采用规则匹配与神经网络相结合的方法抽取关系，构建政策知识图谱分析平台。

既有研究至少存在两大不足：一是基于Citespace的科技政策研究知识图谱聚焦学术文献而非政策文本；二是已构建的科技政策知识图谱多采用自下而上的构建思路，关注政策实体与实体间关系，忽略了科技政策的扶持类和禁止类重要属性信息。本文采用自上

向下的构建路径，首先考虑科技政策的新发展，构建政策知识本体，定义科技政策实体、属性和关系；然后应用BiLSTM模型，识别并提取政策实体、属性及关系，尤其是提取政策扶持类和禁止类实体信息；最后导入Neo4j图数据库完成政策实体、属性及其关系的可视化查询与检索。

2 研究框架

知识图谱在逻辑结构上由模式层和数据层两部分构成[19]，模式层通过本体库规范目标领域内的实体、属性以及不同对象之间的关系，数据层则以<实体，关系，实体>或<实体，属性，属性值>三元组的形式表征知识结构，通过知识抽取实现模式层的实例化。本文构建的科技政策知识图谱，首先定义模式层，明确科技政策主体、客体、政策元数据特征，借此梳理得到科技政策的实体、属性与关系，形成科技政策本体模型；然后在数据层针对采集的科技政策文本应用知识抽取技术提取实体和属性信息，并将三元组信息存入图数据库，完成知识图谱的构建。具体可划分为数据获取、本体构建、知识抽取、知识存储4个部分，具体的构建流程如下。

（1）数据获取。科技政策文本来源主要从各级政府的网站通过爬虫程序采集获取，将采集的政策文本存放在数据库中，方便后续处理。

（2）本体构建。分析科技政策知识结构，确定政策本体中的概念体系，确立类、属性及关系，构建科技政策本体。本体中的概念主要包括政府机构、政策类别以及区域等；属性是对政策文本知识粒度的进一步细化，包括发布时间、政策扶持条文和政策禁止条文等；关系包括政策与机构之间的发布关系，以及政策间的引用关系等。

（3）知识抽取。基于构建的政策本体模型，应用深度学习算法从政策文本中抽取实体、属性及关系信息。实体抽取包括抽取政策标题、发布单位等；属性抽取主要利用深度学习模型从政策全文中抽取出包含情感态度的扶持内容和禁止内容；关系抽取主要包括科技政策与政府机构之间的发布关系和政策文件之间的相互引用关系等。

（4）知识存储。将上述处理流程中获取的实体和关系数据转换成数据格式并批量导入Neo4j图数据库中，采用图结构存储知识，并通过Neo4j实现可视化，直观展示科技政策实体之间以及实体与属性之间的关系。

3 科技政策知识图谱的构建过程

3.1 政策文本收集及预处理

为构建一个较为全面的科技政策知识图谱，本文从多个渠道检索科技政策文本，既包括各级政府网站等官方平台，也包含各类政策咨询服务平台。通过Python爬虫共获取28 741条科技政策文本，经过合并、去重、删除无效数据的整合分析后，数据量缩减到26 660条（缩减7.24%），之后将政策文本存储在MySQL数据库中，为后续知识三元组抽取做准备。在存储过程中，也对政策文本数据集进行预处理，包括去除文本空格、网页标识符等。

3.2 模式层构建

模式层是知识图谱的概念模型和逻辑基础，能够对数据层进行规范约束，在研究中多采用本体作为知识图谱的模式层。本体定义知识图谱的数据模式，是对知识图谱的抽象化表示，通过本体库而形成的知识图谱不仅层次结构较强，而且冗余程度较小[4]。通过研读科技政策文本内容，对科技政策实体、属性和关系进行定义，构建科技政策本体模型（见图1）。

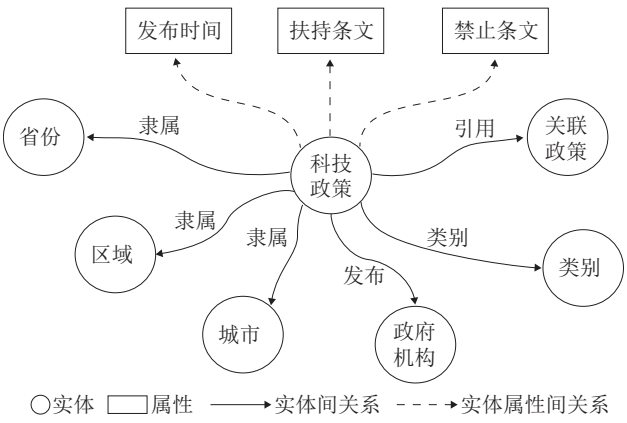

图1 科技政策本体模型

3.2.1 实体定义

实体是知识图谱中的重要节点，在政策本体中实体可以是政策文件、政策类别、政策发布机构等元数据，

也可以是政策文件中提及的关联政策，还可以是抽象的政策概念。本文基于科技政策数据的外部特征和内在知识元素来构建科技政策知识图谱中所需实体，主要包括科技政策、类别、政府机构、关联政策、地区、省份、城市七大实体类型，如表1所示。

表1 科技政策本体中的实体类型及描述

序 号	实体名称	含 义
1	科技政策	科技政策的标题
2	类别	科技政策所属类别，共有30种政策类别，每个政策可能会对应一个或多个类别
3	政府机构	科技政策的发布单位，同时也是政策的来源，一个政策只对应一个发布单位
4	关联政策	科技政策内容中所提及的政策，可能会对应多个政策
5	地区	科技政策的发布单位所在地区，分为中央、东部、中部、西部、东北部
6	省份	科技政策的发布单位所在省份
7	城市	科技政策的发布单位所在城市，若细分到县或区统一划到市行政级别

3.2.2 属性定义

科技政策本体中的实体属性及描述如表2所示。现有实体大多为科技政策的外部特征信息（包括发布时间），而缺乏用户亟需的、蕴含在政策内容中的关键信息。本文将扶持类政策条文和禁止类政策条文归为科技政策的实体属性，目的是帮助广大中小企业用好政策红利，规避风险和政策禁区。

表2 科技政策本体中的实体属性及描述

序 号	属性名称	描 述
1	发布时间	政策的发布时间
2	扶持条文	政策内容中所涉及的对资金补贴和政策扶持等内容
3	禁止条文	政策内容中所涉及的惩罚性和禁止性内容

3.2.3 关系定义

鉴于科技政策的制订与出台存在时序性和关联性，例如，A政策参考了B政策的规定、解释、标准等，或者A政策以B政策为指导思想制定，在这些情形下A政策与B政策具有时间和语义的关联，因此构建政策实体之间的引用关系，能够清晰梳理政府机构政策制定思路以及政策发展脉络，明晰政策导向[20]。此外，在科技政策的本体概念模型中，实体之间的关系还包括类别、发布、隶属三大主要关系。据此，本文确定科技政策本体中的关系类型及描述如表3所示。

表3 科技政策本体中的关系类型及描述

序 号	关系名称	描 述
1	类别关系	存在于科技政策与类别之间，表明科技政策拥有的特定政策类别
2	发布关系	存在于政府机构与科技政策之间，表明政府机构发布科技政策
3	引用关系	存在于科技政策之间，表明科技政策引用的相关政策文件
4	隶属关系	存在于科技政策与地区、省、市之间，表明科技政策是在一定地区内执行

3.3 数据层构建

数据层构建是以模式层构建的科技政策本体为基础，从已获取的语料中抽取结构化信息，主要包括实体抽取、属性抽取、关系抽取三部分[21]。实体是知识图谱的最基本元素，因此实体抽取是知识抽取中最基础和关键的部分，其任务是从语料中识别出命名实体。文本语料经过实体抽取，得到的是离散的命名实体，还需要提取实体之间的关联关系，将实体联系起来形成网状的知识结构。属性抽取是从语料中抽取特定实体的属性信息，刻画完整的实体。关系抽取是构建知识图谱的关键一步，其主要任务是从文本内容中挖掘出实体与实体之间的语义关系，构建<实体，关系，实体 >的三元组，用于后续知识图谱的构建。

3.3.1 实体抽取

（1）提取科技政策文本中的省份、城市、区域实体。chinese_province_city_area_mapper（cpca）是一个用于识别中文字符串中省、市和区的Python开源库。利用cpca库从政策来源字段中提取出政策所属省市，其提取效果如表4所示。经过数据统计发现，政策数据集中的省市信息提取率超过90%，提取效果较好。为避免空值信息对后续统计产生影响，在现有数据的基础上，人工校对空值，使之能够获取每条数据政策的省市信息。然后将省、市字段按照其地理位置划分为东部、

表4　省市实体提取效果

政策来源	省	市	区	机　构
抚松县人民政府	吉林省	白山市	抚松县	人民政府
广州番禺区发展和改革局	广东省	广州市	番禺区	发展和改革局
惠州市龙门县人民政府	广东省	惠州市	龙门县	人民政府
湛江市吴川市经济信息化和科技局	广东省	湛江市	吴川市	经济信息化和科技局
昆明市五华区人民政府	云南省	昆明市	五华区	人民政府

西部、中部、东北部、中央5个类别，归纳为地区实体。

（2）提取科技政策实体。通过使用Python中内置的“re”模块来使用正则表达式检查政策文本是否引用某个政策，之后将同一政策文本中引用的多个政策进行合并。具体操作：首先制定过滤规则，匹配政策文本包含“《 》”“贯彻落实”“依据”“参照”等标志词的字符串，然后从匹配成功的文本字符串中提取科技政策实体并存储。以某政策内容为例：“各有关企业：现将《关于组织申报2018年度市级知识产权优势企业的通知》（渝知发〔2018〕45号）印发与你们，请符合相关条件的企业自行申报，并将申报材料传一份至县科委”，在本例中提取结果为《关于组织申报2018年度市级知识产权优势企业的通知》。本研究所收集的全部科技政策文本，经过处理后共提取出13 389条引用的科技政策，占全体的50.22%。

3.3.2　提取政策扶持条文和政策禁止条文属性

情感分析是自然语言处理的任务之一。从自然语言处理技术角度来看，情感分析的任务是从文本中提取该文本表达的情感倾向。本研究需从政策内容中提取出政策扶持条文和政策禁止条文，因此适用于用情感分析方法解决此问题。

情感分析根据处理文本颗粒度的不同，可分为篇章级和句子级。篇章级情感分析的目标是判断整篇文档表达的是褒义还是贬义的情感；句子级情感分析的任务是判断一个句子表达的是褒义还是贬义的情感。不过，篇章级情感分析只能得到每一条政策数据的情感等级，而无法获得每一条政策数据中包含的带有情感色彩的政策内容，因此舍弃该方法，采用句子级情感分析来完成该任务。

情感分析在某种程度上是文本分类的一种，所以本文利用深度学习方法，根据政策内容训练Bi-LSTM模型对科技政策文本情感分类，具体实现流程如图2所示。

步骤一，确定政策文本按照句子级别划分的类型，具体分为扶持型、禁止型、普通型。如果句子中出现负面词汇“严禁”“整治”“控制”等，或具有明显的惩罚性表达，则划分为禁止型；反之，如果内容中多出现正面词汇，如“补贴”“扶持”“鼓励”等，则划分为扶持型；两者都不是，则划分为普通型。

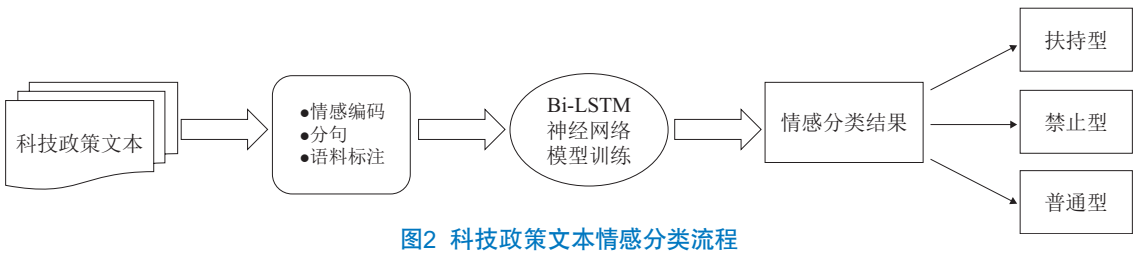

图2　科技政策文本情感分类流程

步骤二，将每条政策内容划分为多个句子，并按照上述规则对句子类型进行数据标注和编码。本研究将26 660条政策文本划分为30多万条句子，由人工对句子进行标注。为获得较好的训练效果，防止过拟合，尽量保证训练集中每一类的句子数目一致。

步骤三，将训练集与测试集输入Bi-LSTM模型中训练神经网络模型，并把epoch设置为60，模型训练结果F1值达到69%。然后，基于已训练好的模型，对其余句子进行预测，并将结果存入CSV文件中。

步骤四，得到带有标签的句子后，将所有的句子整合，提取出每条政策对应的扶持条文和禁止条文。

经过上述处理步骤，获得政策知识图谱所需实体、属性及关系三元组，其中，提取46 392个实体、23 400

个属性以及158 432条实体间关系。之后将三元组存储于Neo4j图形数据库中，实现科技政策知识图谱的可视化查询。

3.4 知识存储

Neo4j数据库是一个高性能的图形数据库，具备高可用性、易扩展性、完整的数据库事务支持和快速检索4个特征，具有强大的可视化能力，也是目前使用最多的图数据库[22]。Cypher是Neo4j的官方查询语言，是一个类SQL语言，可以方便地对图形数据库进行查询和更新。Neo4j支持多种数据导入方式，既可以使用Cypher语言中的LOAD CSV语句直接导入，也可以采用Neo4j-import命令将CSV文件批量导入。其中：第一种方法导入速度较慢；第二种方法速度较快，但需在初始化时进行数据导入。

本文使用第二种方法，首先建立知识网络的关系映射表，然后将科技政策文本中抽取的三元组处理成Neo4j要求的格式，使用Neo4j-import命令批量导入数据库中，构建科技政策知识图谱。该知识图谱包含46 392个节点、158 432条边。数据导入后可利用Cypher语言对构建的科技政策知识图谱进行可视化查询。由于实体节点较多且可视化空间有限，科技政策图谱部分展示如图3所示。每个节点代表一个实体，节点之间的连线代表两两实体间关系，单击实体或关系可查看对应属性信息。

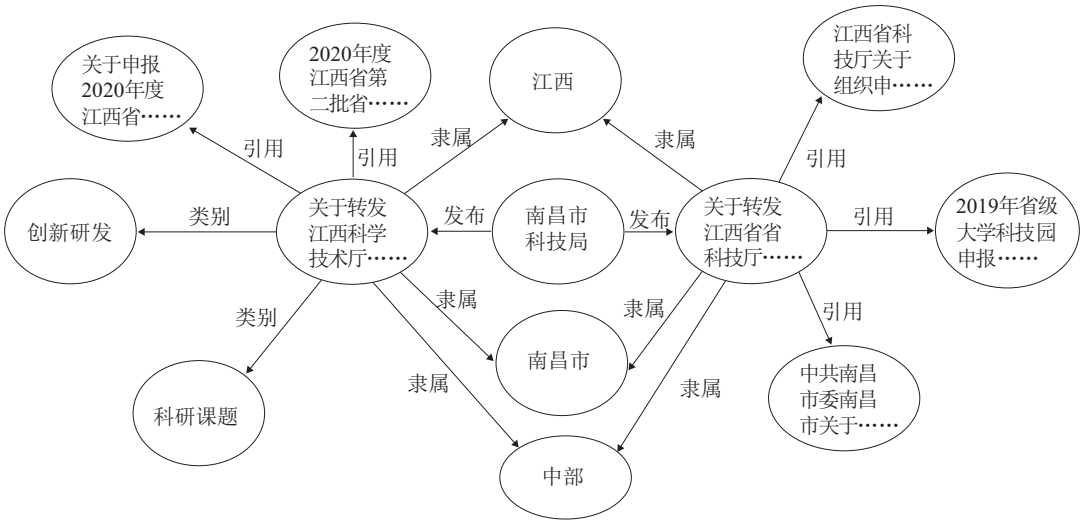

图3 科技政策知识图谱部分展示

4 科技政策知识图谱的可视化查询

构建科技政策知识图谱的最终目的是从海量的政策数据中提取关键语义信息，实现科技政策领域知识的可视化查询和知识发现服务，为政府、企业和个人提供工作抓手。基于前文构建的科技政策知识图谱，利用Neo4j的Cypher查询语言可以实现对科技政策实体和实体关系的查询，并将查询结果可视化呈现，便于用户高效地掌握关键信息，发现事物之间的潜在联系。

4.1 科技政策实体关系查询

在Neo4j数据库中使用Cypher语句中的MATCH子句可实现对科技政策、政府机构、关联政策等实体和相关关系的查询操作。以查询某一科技政策为例，输入“MATCH（m：科技政策）-[r]->（n）WHERE m.name = '内蒙古自治区科技成果转化专项资金管理办法' RETURN m,r,n;”其中“m”为实体信息，“[ ]”内填写需要查询的关系类型，WHERE对查询数据进行过滤，RETURN表示返回结果。该语句生成的知识图谱可查询出与《内蒙古自治区科技成果转化专项资金管理办法》政策相连的所有实体和关系，发布机构、政策类型、隶属的地区以及关联政策等数据都会被呈现。其中，不同实体类型的节点会通过颜色区分，单击实体或可查看其包含的扶持条文和禁止条文等属性信息。该政策中涉及的政策扶持信息包括科技成果转化引导资金、主要支持的重点领域和补贴标准，可帮助用户抓住机会，顺应政策导向，实现供需方的精准匹配，

而政策禁止内容则总结了政策内容的禁止性规定，能够帮助用户及时规避风险。

4.2 科技政策引文查询

引文分析是文献计量学的一种方法，并被广泛应用于知识发现中，主要是通过对文献对象的引用与被引用关系，反映文献之间的外在联系，揭示学科领域的结构和演化规律。政策数据与文献数据类似，也存在引用关系。科技政策引文图谱可为政策的制定提供决策支持，也能梳理不同政策间的关系，明晰政策制订依据。输入“MATCH (m)-[r:引用]->(n) RETURN m,r,n;”可以生成政策实体间“引用”关系图谱，能够实现政策溯源，展现政策体系演进过程，反映中央政府颁布的政策与其他政策的关联关系，以及中央政策与地方政策主题、政策目标的衔接性。

5 结论与启示

为全面深化改革，党的十八届三中全会提出推进国家治理体系和治理能力现代化的宏伟蓝图。实现国家治理现代化，首当其冲的任务是要实现政府治理现代化，而建设数字政府是实现政府治理现代化的重要途径。随着政府信息公开广度和深度的不断延伸，如何有效挖掘海量政策数据，发挥政策对科技创新的指引作用日益引起各界重视。

本文提出的科技政策知识图谱构建方法，为面向政策领域的知识图谱应用提供了鲜活的实例，所构建的科技政策本体库，可以为研究者开展其他政策图谱编绘提供参考，采用的知识抽取与存储技术在其他垂直行业以及金融、教育、医疗等领域也有广阔的应用前景。

未来研究可以从两方面延展。一是充实并完善政策本体。可以考虑借鉴政策评价相关理论，从政策目标、政策工具等出发，丰富政策实体的表征维度，在实现科技政策查询的基础上，满足各级政府开展政策评价的新需求。二是引入新兴技术，完善政策语义知识库的广度与深度。将BERT等考虑上下文语义信息的预训练语言模型与Bi-LSTM模型结合，从更细的粒度抽取政策实体和关系，提升科技政策知识图谱的适应性与易用性。此外，在本文构建的知识图谱基础上，还可以进一步延伸开发在线政策智能问答系统，满足政府与企业、科技提供商与技术应用者对前沿科技政策信息的准确定位与实时获取。

参考文献

[1] 李辉，曾文，吴晨生，等. 中文科技政策数据分析方法研究——以新能源汽车领域科技政策为例 [J]. 现代情报，2018，38（6）：68-72.

[2] PAULHEIM H. Knowledge graph refinement：a survey of approaches and evaluation methods [J]. Semantic Web，2017，8（3）：489-508.

[3] 漆桂林，高桓，吴天星.知识图谱研究进展 [J]. 情报工程，2017，3（1）：4-25.

[4] 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）：589-606.

[5] 吕华揆，洪亮，马费成. 金融股权知识图谱构建与应用 [J]. 数据分析与知识发现，2020，4（5）：27-37.

[6] 白如江，周彦廷，王效岳，等. 科学事件知识图谱构建研究 [J]. 情报理论与实践，2020，43（9）：107-114，124.

[7] 曹明宇，李青青，杨志豪，等. 基于知识图谱的原发性肝癌知识问答系统 [J]. 中文信息学报，2019，33（6）：88-93.

[8] 李建花. 科技政策与产业政策的协同整合 [J]. 科技进步与对策，2010，27（15）：25-27.

[9] ROTHWELL R，ZEGVELD W. Reindusdalization and Technology [M]. London：Logman Group Limited，1985：83-104.

[10] 徐硼，罗帆. 政策工具视角下的中国科技创新政策 [J]. 科学学研究，2020，38（5）：826-833.

[11] 仲伟俊，蔡琦. 科技政策分析框架研究 [J]. 科技管理研究，2014，34（22）：23-27.

[12] 宋伟，夏辉. 地方政府人工智能产业政策文本量化研究 [J]. 科技管理研究，2019，39（10）：192-199.

[13] 祝鑫梅，余晓，卢宏宇. 中国标准化政策演进研究：基于文本量化分析 [J]. 科研管理，2019，40（7）：12-21.

[14] 黄萃，赵培强，李江. 基于共词分析的中国科技创新政策变迁量化分析 [J]. 中国行政管理，2015（9）：115-122.

[15] 李梅芳，王梦婷，齐海花，等. 科技政策国际研究的演化 [J]. 科学学研究，2018，36（9）：1565-1574.

[16] 赵绘存，高峰，闫杰. 2007—2017年国际科技政策研究热点与前沿——基于科学知识图谱视角 [J]. 科技管理研究，2018，38（3）：42-49.

[17] 张维冲，王芳，黄毅. 基于图数据库的贵州省大数据政策知识建模研究 [J]. 数字图书馆论坛，2020（4）：30-38.

[18] WANG P, LI Z S, LI Z Y, et al. A government policy analysis platform based on knowledge graph [C] //2019 2nd International Conference on Artificial Intelligence and Big Data. Chengdu: IEEE, 2019: 208-214.

[19] 黄恒琪, 于娟, 廖晓, 等. 知识图谱研究综述 [J]. 计算机系统应用, 2019, 28 (6): 1-12.

[20] 张超, 官建成. 基于政策文本内容分析的政策体系演进研究——以中国创新创业政策体系为例 [J]. 管理评论, 2020, 32 (5): 138-150.

[21] JI S X, PAN S R, CAMBRIA E, et al. A survey on knowledge graphs: representation, acquisition and applications [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020 (2): 1-25.

[22] YAN J H, WANG C Y, CHENG W L, et al. A retrospective of knowledge graphs [J]. Frontiers of Computer Science, 2018, 12 (1): 55-74.

作者简介

张雨，女，1996年生，硕士研究生，研究方向：知识图谱。
吴俊，男，1971年生，博士，副教授，通信作者，研究方向：文本挖掘、知识图谱，E-mail：Wujun1127@126.com。

Research on the Construction of Science and Technology Policy Knowledge Graph

ZHANG Yu WU Jun
(School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: In recent years, governments at all levels have introduced numerous policies on scientific and technological innovation. Faced with massive policies, it is difficult for enterprises to make full use of the data resources. In order to encourage enterprises to accurately locate and quickly search for policies, it is necessary to use artificial intelligence methods to achieve knowledge modeling and build structured queries based on domain knowledge graph. This paper takes 26 660 science and technology policies as data sources, constructs the domain ontology of science and technology policies, and completes knowledge extraction through Bi-LSTM deep learning model. Finally, the data are stored in the graph database Neo4j and the knowledge graph of science and technology policy is constructed. The constructed science and technology policy knowledge graph has more than 40 000 entity nodes and more than 150 000 relationships, which can realize the associated query and visual presentation of different fine-grained policy entities and relationships. The method proposed by the research enhances the new ideas of the knowledge graph of the vertical domain and lays the foundation for the intelligent question answering system for science and technology policies.

Keywords: Knowledge Graph; Science and Technology Policy; Ontology; Bi-LSTM

（收稿日期：2021-07-06）