文献与数据学报
第 4 卷　第 1 期　　2022 年 3 月

科技政策扩散路径生成模型关键技术研究 *

许乾坤　刘　耀
（中国科学技术信息研究所，北京　100038）

摘　要：[**目的 / 意义**] 通过分析科技政策文本组织结构和语义结构，发现科技政策文本中潜在的知识网络信息，构建科技政策扩散路径生成模型，将文本中包含的隐性知识显性化。[**方法 / 过程**] 通过对科技政策篇章的结构分析，深入挖掘科技政策内容中包含的特征词，将无结构的科技政策文本转化为结构化数据。显性扩散路径生成是通过分析政策内容与结构的特点，获取引用政策，构建政策扩散路径节点的语言模型。隐性路径生成是结合网络表示学习方法对科技政策结构化数据进行向量表示，构建科技政策篇章知识网络模型，发现科技政策文本中潜在的知识网络信息，为政策研究构建一个智能化研究的模型。[**结果 / 结论**] 通过实验证明，生成的科技政策扩散路径对验证了本文所构建科技政策扩散路径生成模型的有效性，为科技政策内容的深入研究提供了新的研究思路。

关键词：科技政策　扩散路径　知识网络结构　路径生成
分类号：G203
DOI：10.31193/SSAP.J.ISSN.2096-6695.2022.01.03

0　引　言

科技政策是对科技成果转化应用、促进科技创新，以及推动产业数字化智能化等科学技术的指引。只有科学技术得到发展，科技政策得到科学的指引，才能更好地推动我国社会经济的发展。在大数据背景下，政策文本数量越来越大，文本资源种类越来越复杂，内容越来越多样，研究需求越来越多变，传统的政策文本分析方法已经不能满足政策研究的需要。如何从海量的政策数据中凝练出重要信息，评估政策实行的结果，制定科学的科技政策是科技政策研究的重要挑战。目前，科技政策扩散研究方法多以对理论因素或模式特征的定性分析为主，侧重对政策扩散的理论分析，通过实证分析对前人提出的理论框架加以修正和完善。随着政策文本数量的激增和开放获取的便捷性，基于海量数据的量化分析方法逐渐成为政策扩散研究的主流

* 本文系国家社会科学基金一般项目“数字资源知识共享与知识再利用模式与方法研究”（项目编号：21BTQ011）研究成果之一。
[作者简介] 许乾坤（ORCID：0000-0003-0695-8802），男，硕士研究生，研究方向为自然语言处理与人工智能，Email：xuqk2019@istic.ac.cn；刘耀（ORCID：0000-0003-3729-3866），男，研究员，博士，研究方向为自然语言处理、知识工程，Email：liuy@istic.ac.cn。

方法。定量研究方法主要是就政策数量、发布机构、发布时间等进行研究，缺乏对政策内容的深入分析和自动挖掘，不能根据用户需求检索生成政策扩散路径，对政策内容的相关研究相对较少。

从改革开放以来，我国不断进行政策探索与创新，并将有效政策或者成功经验在全国范围内推广。政策如何在不同层级、不同区域的政府间进行传播，经过学者的不断努力已经形成较为成熟的研究体系，并由点及面形成了全国范围内的扩散，但仍然存在一些局限。随着其他学者更加深入的研究发现，政策内容包含的语义信息是影响政策扩散的重要因素，对政策内容和结构的深入挖掘是政策扩散研究中有待深入拓展的领域。

本文将无结构的科技政策文本转化为结构化数据，构建科技政策扩散路径生成模型。显性扩散路径生成是通过分析政策内容与结构的特点，引用政策是通过分析文本中的特点与结构，本文提出了构建显性生成政策扩散路径模型的方法。政策文本结构复杂，是显性的组织结构和隐性的语义结构的结合体，本文实现了对科技政策篇章文本的概念识别与关系标引的自动化，使用的 BiLSTM+CRF 深度学习方法和分类方法能够有效地标引科技政策文本中的概念以及预测概念对之间的关系，将抽取的概念节点经过网络表示学习算法得到向量表示，结合了 Node2vec 算法和篇章节点的知识网络对该模型进行了改进，计算得到的政策篇章知识网络结构，生成政策隐性扩散路径。实验表明，本文提出的显性扩散路径和隐性扩散路径生成方法可以有效的生成和展示政策扩散路径。

1 相关研究

1.1 政策文本挖掘研究现状

目前，随着政府信息公开化和互联网的蓬勃发展，对政策文本的深入研究日益受到重视，利用大批量的政策数据，分析和挖掘政策文本中包含的隐性知识变得越来越重要。文本挖掘和数据分析等领域不断更新方法，很大程度上拓展了政策文本的理论和方法研究。政策文本挖掘包括对政策文本内容分析、政策文本量化分析、政策文本数据挖掘方法等。政策文本的研究主要分为定性分析和定量分析，定性分析侧重政策的制定背景、目标、内容及效果的分析，对政策文本包含的隐含知识不能充分挖掘[1-3]；定量分析大多集中在对政策文本发布时间、发布地区和政策发布主题等基本信息的研究。随着政策文本研究数据的增长，定性分析的效率不高、研究成本较高等局限性显现，而定量分析通过将无结构化的政策数据转化为结构化数据，借助机器学习、深度学习的方法对政策文本中包含的隐性知识进行深入挖掘，大大降低了政策研究的成本，提高了过程的可复制性，其逐渐占据主流。

政策文本量化分析是对政策文本库中的文本内容和相关信息进行量化统计的方法，主要包括政策的发文主体、颁布日期、主题词、政策分类和政策之间的引用关系等。黄萃[4]采用文献量化的方法，对 4707 篇政策文献进行分析，通过共词分析和聚类分析，研究了四个领域的主题变迁，证明可以解释政策的变迁规律以及特点和趋势等。学者们早就认识到分析中大规模的文本数据的巨大成本阻碍了它们在政治学研究中的应用。Grimmer 等人[5]实现了自动化文本分析，大

大降低了分析大量文本集合的成本，并提供了构建模型的方法。李江[6]通过分析政策的分布特征、政策主体间的合作方式以及政策包含的体系结构等方面，验证了政策文献计量在四个方面的优势与劣势。丁洁兰[7]构建了二维词频分析框架，通过分析科技政策文本的关键词，研究科学计量方法在科技政策文本研究中的应用范围。

政策文本挖掘是将无结构化的科技政策文本转化为结构化的政策文本，从而对科技政策文本的内容语义和结构语义进行挖掘与分析。科技政策文本包含众多隐性知识，如何挖掘科技政策文本中的隐性知识并将政策联系起来，是政策研究的一个方向。通过使用自然语言处理、机器学习和深度学习等相关技术，对政策的文本进行分句、切分词、词性标注等。对于科技政策的内容语义挖掘，需要处理大批量的政策文本，通过命名实体识别、分类和知识网络构建等方法，发现政策文本间潜在的语义联系。Nowlin 等人[8]提出了一个问题定义模型，使用定量文本分析估计该模型，借助潜在狄利克雷模型，分析了七个不同层面。Leifeld 等人[9]借助社会网络分析，对政策中包含的概念和相关的结构特征进行分析，并且将网络分析作为政策研究的方法。科技政策文本中包含丰富的语义信息，通过运用自动标引概念与概念间关系的方法，结合先验知识在一定程度上优化了科技政策文本结构化分析的方法。研究科技政策扩散路径的过程，可以帮助各级政府优化、调整政策内容，提高科技政策创新水平。

1.2 网络表示学习相关技术研究

随着数据量的增长，各领域的数据都以网络的形式来表示，如信息网络、社交网络等，为了更方便高效地处理网络数据，利用网络表示学习算法对网络分析而言是近年来一个热点研究方向。网络表示学习的应用有着重大的意义，通过得到网络的特征表示，将网络中的节点和边得出向量表示，该结果可以通过深度学习或者机器学习方法对各种基于网络的表示进行更深层次的研究。在应用网络表示学习的早期，通常是将数据点的特征向量构建亲和度图，再将该亲和度图映射到低维空间中，面临时间复杂度很高问题，对于大规模的网络很难有效地处理。图分解技术的提出，目的是将节点进行低维空间表示，先使用矩阵分解得到图的低维表示，然后使用随机梯度下降法优化图的低维表示，仍存在时间复杂度高的问题。Bryan 等人[10]在 2014 年借助于语言模型中的深度学习技术 Word2vec 来学习图的邻接矩阵，用 DeepWalk 算法隐含表示图的结构。根据图中随机游走的路径节点出现的概率与自然语言处理中词频的分布十分相似，通过随机游走的路径节点生成的路径，作为语言模型 Word2vec 的输入并且得到对应的向量表示。Aditya Grover 等人[11]在随机游走 DeepWalk 的基础上对图网络路径节点生成的策略或方法进行了改进，提出了 Node2vec 算法，该算法通过改进随机游走的路径节点生成的策略，在随机游走的过程中增加了广度优先搜索和深度优先搜索，与 DeepWalk 算法相比提高了随机游走的路径节点生成的质量，在时间和空间复杂度上，有着更优的节点生成策略。Tang 等人[12]提出了 LINE 方法，借助于边缘采样算法，对任意的信息网络结构都可以将其嵌入到低维的向量空间中，也可以保留局部的网络拓扑结构。Shaosheng Cao 等人[13]在 2015 年提出能最大程度保留网络的高、低阶的结构信息算法 GraRep，通过改变图网络中转移概率表的不同阶的节点转移概率矩阵来捕获不同阶的结构信息，从而得到不同阶的节点特征表示，对矩阵进行分解，将不同阶的节点特征表示进行归一化处理后，将所有的节点特征合并起来作为最终的节点表示，并选择最高阶的

结构信息，改变了缓慢和复杂的抽样过程。Pan 等人[14]提出了基于三方深度网络表示模型，即 TriDNR 模型。该模型充分利用了节点的标签信息，以向量的形式对节点进行表示，生成网络结构，借助于机器学习方法进行分析。TriDNR 模型借助于随机游走模型最大限度对节点之间的关系进行表示，接着通过固定节点的共现性获取节点与词之间的相关性，最后对标签与词之间的关系进行建模。将 TriDNR 模型获取的节点、内容和结构信息输入到神经网络模型中，取得了不错的结果。

网络节点通常是文本中特征或者标签信息的集合，集合中包含的信息可以计算网络节点之间的相似性。例如，对于两篇科技政策来说，如果它们之间存在相同的概念或者术语，在构建每一篇科技政策的网络拓扑结构时，会存在相似的网络节点，在计算网络拓扑结构之间的关系时，从而构建政策之间的关系。本文从政策文本的结构和内容出发，对科技政策文本进行深层结构化分析，使用命名实体识别的方法对概念提取，借助网络表示学习技术对政策文本构建知识网络和多样化排序技术，实现科技政策扩散路径自动生成，在一定程度上丰富了科技政策的研究方法。

2 科技政策扩散路径生成模型构建

在对科技政策资源全解析的基础上，探讨了科技政策扩散路径生成的理论方法与流程。通过构建政策文本的知识网络结构，将科技政策扩散路径的生成分为科技政策显性扩散路径生成和科技政策隐性扩散路径生成，借助引用政策对显性扩散路径的生成进行建模；依据政策内容和结构，对隐性扩散路径的生成结果进行展示与分析。

2.1 科技政策知识网络模型构建

知识网络的构成要素主要包括网络节点、网络资源、网络活动和支撑环境。知识网络的节点是行为主体，包括企业、高校、机构、论文或词语等，根据本文的研究内容，知识网络的节点是从政策内容汇总抽取的概念；知识网络中节点的活动是节点与节点之间的联系，如在引证网络、词网络是引证关系，在共现网络是共现关系，本文的知识网络节点活动是科技政策中包含的语义关系。因此，本文将提取的概念与概念间的关系定义为知识网络。

知识网络模型的构建主要是概念集合的构建和概念间关系集合的构建。构建概念集合就是通过概念模型抽取出科技政策篇章文本中的所有概念，收集所有不重复的概念，形成概念集合；概念间关系集合是从所有的概念间关系中抽取概念，并将概念保存在相应的标签信息类别下。因此，本文将知识网络表示为 $N=(V, E, D, L)$。$V=\{v_1, v_2, \cdots, v_N\}$ 表示节点，即各个概念。$e_{i,j}=(v_i, v_j) \in E$ 表示节点之间的边，即概念之间的关系。$D=\{w_1, w_2, \cdots, w_N\}$ 表示每个节点的文本信息。$L=\{L_v, L_r\}$ 表示概念和关系标签的集合，其中 L_v 表示概念节点的类别标签，而 L_r 表示概念间关系的类别标签。如图 1 显示出了“科学技术”这个概念与其他概念之间关系。

```
"科学技术": {
        "Constellation": ["纳米科学", "科技服务"],
        "Forward": ["工程领域", "科学技术", "信息学", "科技发展政策", "基础设施", "科学技术竞争力", "核心技术", "新型工程"],
        "Inclusion": ["公共技术设施系统", "绿色技术", "信息通信技术"],
        "Reason": []
}
```


图 1　知识网络 json 格式

在构建知识网络之后，如何合理的表示网络中的特征信息是网络分析研究的关键问题之一。传统的网络建模方法在大规模网络上不适应高效地进行网络分析，如节点聚类、节点分类、链路预测和重要节点发现等，那如何利用知识网络中的信息有效地对网络中的节点表示仍是进行网络分析的关键问题。传统的邻接矩阵仅对知识网络的边进行了表示，无法加入节点的属性。随着机器学习的快速发展，网络节点作为机器学习算法输入的低维表示学习日益重要，网络表示学习已经成为数据深度挖掘的研究热点。

TriDNR 模型通过利用知识网络的节点结构、节点内容和节点标签等，借助于深度网络表示模型学习知识网络中最优节点表示。网络的信息表示为 $G=(V, E, D, L)$。$V=\{v_1, v_2, \cdots, v_n\}$ 表示节点的集合。$e_{i,j}=(v_i, v_j) \in E$ 表示节点之间的关系。$d_i \in D$ 表示一个文本文档与节点之间的关系。$C=L \cup U$ 表示的是网络中类标签信息，其中，L 表示有标记的节点，U 表示无标记的节点。

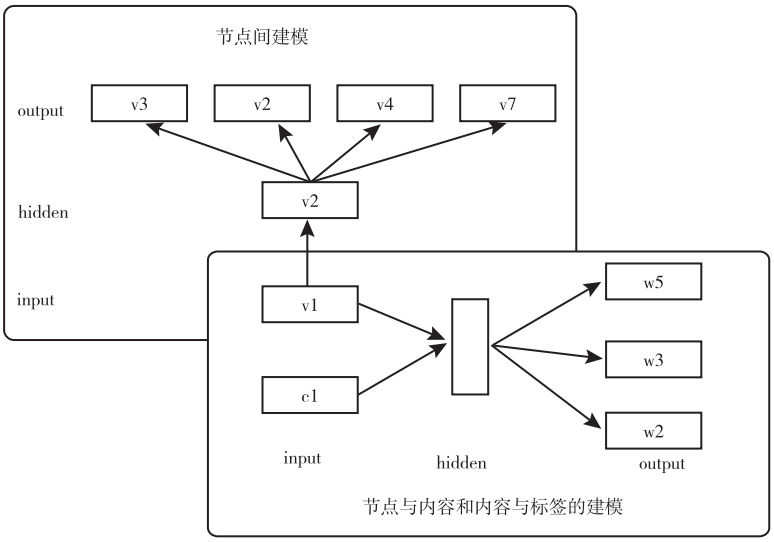

图 2　TriDNR 模型结构图

TriDNR 算法模型主要由两部分组成，分别是随机游走序列生成的网络结构；耦合神经网络模型学习嵌入的每个节点。通过节点间关系建模，上层在假设连通的情况下，学习结构关系，统计相应的节点，借助于耦合神经网络模型获取节点间的关系，从而评估节点内容的相关性。下层是对文档中词的上下文信息进行建模，主要是对节点内容的相关进行建模，每个节点的标签作为

输入，同时学习输入标签的向量和输出词的向量，标签信息不用于节点间关系建模。下层结构的目标函数为：

$$L=\sum_{i=1}^{|L|}\log\ \mathbb{P}\ (w_{-b}:w_b\mid c_i)+\sum_{i=1}^{|N|}\log\ \mathbb{P}\ (\mathrm{w}_{-b}:w_b\mid v_i) \tag{1}$$

通过计算上下文中一系列词的长度，以及节点的类标签，利用文本信息学习将每个文档进行向量表示。整体模型的目标函数是公式（2）的最大似然估计。

$$\begin{aligned}L=&(1-\alpha)\sum_{i=1}^{N}\sum_{s\in S}\sum_{-b\leqslant j\leqslant b,j\neq o}\log\ \mathbb{P}\ (v_{i+j}\mid v_{\mathrm{i}})\\&+\alpha\sum_{i=1}^{N}\sum_{-b\leqslant j\leqslant b}\log\ \mathbb{P}\ (w_j\mid v_i)+\sum_{i=1}^{|L|}\sum_{-b\leqslant j\leqslant b}\log\ \mathbb{P}\ (w_j\mid c_i)\end{aligned} \tag{2}$$

式中，α 是平衡节点拓扑结构、节点内容和节点标签信息的权重，b 是序列的窗口大小。其中第一个子式是计算给定一个节点，出现在这个节点周围的其他节点，最终实现了将节点的拓扑结构、文本内容和标签三者信息共同融合。

2.2 科技政策显性扩散路径生成模型构建

政策扩散是对政策知识、信息、经验等在不同政策主体间的扩散。政策扩散路径中路径节点的引用政策参照的是文献引用网络分析。文献引用网络主要包括专利文献、科技期刊、科技报告、会议论文集以及学位论文等文献的引用和被引用关系。对于政策研究来说，政策与政策之间的扩散或者引用参照，与文献引用相类似，引用政策标题通常会在政策内容中采用书名号的方式将被引用政策法规的名称列出，政策的显性扩散路径主要包括国家立法机关和行政机构制定的全国性政策引用其他同级别的中央文件、地方政府引用中央文件和地方政府间同级别扩散三种。例如：生态环境部在 2021 年 5 月 30 日颁布的《关于加强高耗能、高排放建设项目生态环境源头防控的指导意见》中提到“新建‘两高’项目应按照《关于加强重点行业建设项目区域削减措施监督管理的通知》要求……”，河北省人民政府办公厅印发《关于加快医学教育创新发展实施方案的通知》（冀政办字〔2020〕220 号）提到“为认真贯彻落实《国务院办公厅关于加快医学教育创新发展的指导意见》（国办发〔2020〕34 号）精神，加快……”，浙江省科学技术厅印发的《浙江省科研诚信信息管理办法（试行）》提到“根据中共中央办公厅 国务院办公厅《关于进一步加强科研诚信建设的若干意见》、《浙江省公共信用信息管理条例》、省委办公厅、省政府办公厅《关于进一步加强科研诚信建设弘扬科学家精神的实施意见》……”。上述政策内容中体现了政策文献之间的扩散关系，即《关于加强重点行业建设项目区域削减措施监督管理的通知》《国务院办公厅关于加快医学教育创新发展的指导意见》和《浙江省公共信用信息管理条例》等都是引用政策或者扩散政策，分别对应了政策的显性扩散路径各种情况。

科技政策在形式结构方面具有特定的体式、严格的规范和统一的要求，以及公文的特点。

科技政策的标题是对政策内容的高度概括，明确了对象、文件的来源等。政策一般分为目录形式和总分总形式。目录形式的政策一般是国家立法机关和行政机构制定的全国性政策或者地方政府制定的文件。总分总形式的政策一般包括开头、主体内容和总结三个部分，其中在政策的开头部分简要的说明了发文的依据或者实现目标，引用政策一般会出现在科技政策的开头部分。通过对政策内容的阅读分析及引用政策总结，发现引用政策或者路径节点一般是依据固定的词汇来表达对引用关系对的说明。例如："依据《中华人民共和国民办教育促进法》《中华人民共和国消费者权益保护法》《中华人民共和国未成年人保护法》……"、"全面落实《国务院办公厅关于加快医学教育创新发展的指导意见》（国办发〔2020〕34 号）……"、"为贯彻落实《国务院办公厅关于加快医学教育创新发展的指导意见》（国办发〔2020〕34 号）精神……"等。通过分析大量政策文本，发现"依据""落实"等路径词加上书名号的形式，可以依据参照关系构建扩散路径网络。本文确定了 11 个路径词，并依据路径词从政策文本中抽取对应的引用政策，从而构建政策扩散路径网络，其中节点表示的是引用政策。如表 1 所示。

表 1　显性扩散路径词

序号	路径词	示例
1	落实	进一步落实《中共中央 国务院关于促进中医药传承创新发展的意见》和全国中医药大会部署……
2	适用	适用《中华人民共和国招标投标法》的政府采购工程建设项目……
3	贯彻落实要求	深入贯彻落实习近平总书记考察浙江重要讲话精神和《长江三角洲区域一体化发展规划纲要》《国家职业教育改革实施方案》《深化新时代教育评价改革总体方案》要求……
4	根据	根据《国家自然科学基金条例》《内蒙古自治区科学技术进步条例》……
5	实施	加快实施《云南省企业公共服务平台云化改造升级方案》，提升完善我省中小企业公共服务体系……
6	制定	研究制定《陕西省落实中外"快捷通道"实施方案》，为符合条件……
7	按照	按照《关于深化项目评审、人才评价、机构评估改革的意见》《山东省科技计划项目科研诚信管理办法》（鲁科字〔2020〕105 号）等相关规定……
8	贯彻	深入贯彻《山东省人民政府办公厅关于推进省级财政科技创新资金整合的实施意见》（鲁政办字〔2020〕64 号）……
9	发布	每年滚动发布《工业企业智能改造指南》，各市制定地区智能改造实施方案……
10	修订	推动修订《吉林省促进科技成果转化条例》，加快新一代……
11	参照	参照《公务员法》管理的事业人员……

基于以上发现的路径词，构建政策扩散路径节点的语言模型，即从目前已获取的所有政策中抽取政策内容中包含的引用政策，将抽取的结果保存在政策库中。从政策的标题开始查询政策内容中包含的引用政策标题，接着对抽取到的引用政策标题进行匹配，匹配其对应的标题，循环查询引用政策标题，直到无引用政策为止。对于从政策内容中查询其包含的引用政策标题来说，通

过该政策的知识网络结构相似度值为前 10 调用政策扩散路径节点的语言模型，从政策内容中抽取对应的结果，从而循环查询。如图 3 所示。

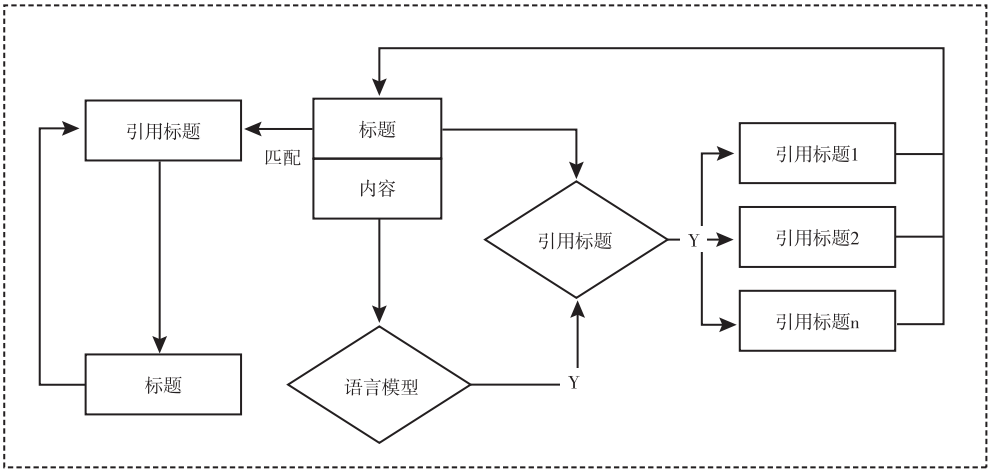

图 3　显性扩散路径查询流程

2.3　科技政策隐性扩散路径生成模型构建

科技政策篇章文本内容包含丰富的知识，从不同角度挖掘政策的关键内容，可以有效的建立政策与政策之间的知识网络联系。政策文本结构复杂，是显性的组织结构和隐性的语义结构的结合体。政策的组织结构体现了政策自身的内容组织，政策语义结构体现了整个科技政策体系下政策内容的联系。各级政府在制定政策的过程中，对创新政策有意识的进行理解和选择性的吸收，由于政策文本的特殊性，被参照的政策的主题和关键词往往表达了政策的重要内容，体现了政策与创新政策之间的区别，在制定创新政策过程中需要重点学习和理解。裴雷[15]对信息化政策扩散中的主题承继和主题创新进行检测验证，发现政策扩散主题在扩散时，承继了扩散性政策的重要政策主题并保持同样的政策重要性或比例。

通过抽取科技政策篇章文本中的概念实体和关系，组成多个三元组，当三元组之间有公共概念时，就会形成 { 概念 1，语义关系，概念 2}、{ 概念 2，语义关系，概念 3} 等多个三元组，那么概念 1 与概念 3 之间就建立了联系。例如：“国家高新区要加大基础和应用研究投入，加强关键共性技术、前沿引领技术、现代工程技术、颠覆性技术联合攻关和产业化应用，推动技术创新、标准化、知识产权和产业化深度融合。”其中，“关键共性技术”、“前沿引领技术”、“现代工程技术”和“颠覆性技术”概念之间会形成一种联系。通过建立概念之间的关联关系，需要抽取科技政策篇章文本中的概念实体和关系，从而生成政策网络拓扑结构。

科技政策篇章文本内容广泛，存在一定的缩略词语，但科技政策内容撰写通常与公文保持一致，用词严谨正式，同时会承继描述科技政策的特征词，运用命名实体识别的方法将科技政策篇章文本中的概念抽取出来。对政策与政策之间无法建立有效联系的问题，本文提出了将科技政策篇章文本中的概念、概念间关系共同组成知识网络，将政策与政策之间的关联与计算问题转化为

计算政策文本的结构相似度度量问题，即通过构建科技政策篇章文本的知识网络结构，两两计算政策知识网络结构相似度。

隐性扩散路径生成具体步骤如下所示：

步骤一：基于科技政策篇章文本构建概念、概念间关联关系的知识网络结构；

步骤二：使用网络表示学习中的 Tri-Party 模型将科技政策篇章文本中概念与概念间关系抽取模型关联得到的知识网络向量化，保留网络信息、节点内容信息和节点标签信息；

步骤三：将网络表示学习得到的向量表示输入相似度模型；

步骤四：计算政策知识网络结构相似度，生成科技政策隐性扩散路径。

3　科技政策扩散路径生成结果及评价

政策扩散路径生成的过程需要从原始政策资源中凝练出各类重要的信息，为了满足生成政策扩散路径所需资源，本文构建了地方政策库、政策资源库、科技政策显性扩散路径库、科技政策隐性扩散路径库和再生资源库等科技政策资源库，为科技政策资源解析、存储和展示奠定了基础。政策扩散路径生成模型包括显性扩散路径、隐性扩散路径。政策与政策之间存在多种扩散关系，如一篇政策的内容来源于多篇政策、一篇政策被多篇政策借鉴和引用以及单篇政策之间的扩散。对于单篇科技政策文本，借助于语言模型判断政策文本中是否包含生成显性扩散路径的依据，从该政策文本中抽取概念，标引概念间关系，构建隐性知识网络结构，为政策扩散路径生成提供指导。

3.1　科技政策资源获取与解析

3.1.1　资源获取

本文资源获取从 2020 年 10 月开始，截至 2021 年 12 月，主要收集中央人民政府网站、科技部网站、省（区、市）人民政府网站、中国科技情报网公开的、具有正式文号的规范政策文本，政策的类型主要包括法律、条例、纲要、规划、计划、办法、决定、意见、细则、通知等，不包括回函、批示、领导讲话以及名单、行业标准等。本研究收集了中央人民政府网站和 31 个省（区、市）人民政府网站公布的科技政策 58241 篇，利用 MongoDB 数据库对这些资源进行存储，科技政策具体来源分布如表 2 所示。

表 2　政策来源分布

省（区、市）	数量（篇）	省（区、市）	数量（篇）
北京	7135	辽宁	831
福建	5435	湖南	822
江西	3559	西藏自治区	794
河南	3206	广西壮族自治区	723
陕西	2771	山西	500

续表

省（区、市）	数量（篇）	省（区、市）	数量（篇）
河北	2642	广东	463
安徽	2591	海南	457
吉林	2117	四川	384
湖北	2025	重庆	356
黑龙江	1793	新疆维吾尔自治区	278
贵州	1503	浙江	167
宁夏回族自治区	1071	山东	104
天津	1001	江苏	91
青海	998	内蒙古自治区	56
云南	988	甘肃	15
上海	850		
中央人民政府网站		12515	
总计（篇）		58241	

注：按照 31 个省（区、市）人民政府网站科技政策数量排序。

3.1.2　资源解析

已获取的政策资源是构建地方政策库的重要基础。地方政策库是对政策资源进行初步的结构化解析，主要包括构建 Schema 和元信息提取。由于政策资源结构的特殊性，对语料进行初步的存储，方便中间资源库和再生资源库的调用。因此，本文构建了地方政策库、中间资源库和再生资源库的 Schema 结构。

地方资源库是将获取到的政策内容，未经删减，直接保存的政策资源集合。通过爬虫将政策内容、发布时间、标题等保存到 MongoDB 数据库中。下面是关于地方资源库 Schema 字段的描述。

Title：政策标题。标引了政策的标题。

Data：政策发布时间。标引政策的发布时间，如果该网站只有发文日期而没有发布日期，则以发文日期为准。

Url：政策内容原始网站链接。标引了政策来源网站，可以通过链接直接查看原网站内容。

TaskName：政策来源或者是任务名。标引政策的来源省（区、市）等。

Content：政策内容。将政策内容全部存入 solr 库，供以后调用。

```
{
  "id": "d7f998388da6e21d",
  "type": "地方政策库",
  "_root_": "d7f998388da6e21d",
  "resourceType": "LocalPolicyBank.xsd",
  "from": "本地",
  "title": "北京市医疗保障局 北京市卫生健康委员会 北京市财政局 北京市人力资源和社会保障局关于调整新型冠状病毒核酸检测相关政策的通知",
  "content": "京医保发〔2020〕22号\n各区医疗保障局、卫生健康委员会、财政局、人力资源和社会保障局，北京经济技术开发区社会事业局、财政审计局，各有关医
  "language": "CH",
  "fileId": "3f9a3562582c966e",
  "level": "0",
  "localPolicyBank_info_content": "京医保发〔2020〕22号\n各区医疗保障局、卫生健康委员会、财政局、人力资源和社会保障局，北京经济技术开发区社会事业J
  "localPolicyBank_info_taskName": "LocalPolicyBank_beijing",
  "localPolicyBank_info_url": "http://www.beijing.gov.cn/zhengce/zhengcefagui/202006/t20200629_1934023.html",
  "localPolicyBank_info_title": "北京市医疗保障局 北京市卫生健康委员会 北京市财政局 北京市人力资源和社会保障局关于调整新型冠状病毒核酸检测相关政策
  "localPolicyBank_body": "",
  "localPolicyBank_info_date": "2020-06-24",
  "uploaderSearch": [
    "admin"
```


图 4　地方政策库

中间资源库存储的资源来源于地方政策库，是对地方政策库资源的二次解析，是解析过程中产生的中间结果，可以作为完整的资源按需检索与展示。中间资源库包括资源库、显性政策路径库和隐性政策知识网络路径库。下面是关于中间资源库 Schema 字段的描述。

Explicit：显性路径。依据路径词从政策内容中抽取相关的引用政策，作为构建科技政策显性路径的资源。如从《浙江省科研诚信信息管理办法（试行）》中提取出的内容，涉及《浙江省科研诚信信息管理办法（试行）》《关于进一步加强科研诚信建设的若干意见》《浙江省公共信用信息管理条例》《关于进一步加强科研诚信建设弘扬科学家精神的实施意见》《科研诚信案件调查处理规则（试行）》《国家科技计划（专项、基金等）严重失信行为记录暂行规定》。

Cat：分类。采用分类模型对政策内容进行分类。如科技政策、财金审计政策等共分为 18 类。

PublicDate：发布日期。如 20200623。

Title：政策标题。

Content：政策内容。如果政策格式符合上述描述的，content 字段只保存了按照政策结构提取之后没有结构的政策内容。

SiteName：网站名称。在展示政策内容时，可以直观展示政策的来源网站。


```
{
  "id": "f10714e3b2f8171a",
  "type": "政策",
  "from": "本地",
  "resourceType": "policy.xsd",
  "title": "（六）发挥行业组织作用。",
  "content": "发挥行业组织熟悉行业、贴近企业的优势，为政府和行业提供双向服务。行业组织应加强数据统计、成果鉴定、检验检测、标准制订等能力建设，提高为
  "language": "CH",
  "fileId": "7b19d063ff32bebb",
  "uploaderSearch": [
    "admin"
  ],
  "parentTitle": "工业和信息化部 发展改革委 科技部关于印发《汽车产业中长期发展规划》的通知_四、保障措施",
  "parentTitleId": "4d3870d88ef29e39",
  "level": "2",
  "order": 1,
  "wholeTitle": "工业和信息化部 发展改革委 科技部关于印发《汽车产业中长期发展规划》的通知_四、保障措施_（六）发挥行业组织作用。",
  "_version_": 1601963235759620000
```


图 5　中间资源库

ParentTitle：政策父标题。如《浙江省科研诚信信息管理办法（试行）》。

Title：政策子标题。如《关于进一步加强科研诚信建设的若干意见》《浙江省公共信用信息管理条例》。

ParentId：政策父标题索引。保存的是该标题来源的索引，每一条数据都有唯一的索引 id，通过索引 id 可以查到该政策的所有内容。

Id：政策子标题索引。


```
"id": "68d458d05b7a5a9d",
"type": "显性政策路径",
"_root_": "68d458d05b7a5a9d",
"resourceType": "explicitpath.xsd",
"from": "本地",
"title": "ed0b1f1cdfa9e96f",
"content": "ed0b1f1cdfa9e96f\n",
"language": "CH",
"fileId": "d1425f534dab35ee",
"level": "0",
"explicitpath_info_parentTitle": "江西省人民政府办公厅印发关于完善政府性融资担保体系切实支持小微企业和“三农”发展若干措施的通知",
"explicitpath_info_parentId": "ed0b1f1cdfa9e96f",
"explicitpath_info_title": "国务院关于促进融资担保行业加快发展的意见",
"explicitpath_info_id": "39f64f33-75b8-45e3-9452-409f8d1b625c",
"uploaderSearch": [
  "admin"
```


图 6　显性政策路径库

Title1：扩散父标题。如《浙江省科研诚信信息管理办法（试行）》。

Title2：政策子标题。如《关于进一步加强科研诚信建设的若干意见》《浙江省公共信用信息管理条例》。

Id1：政策父标题索引。保存的是该标题来源的索引，每一条数据都有唯一的索引 id，通过索引 id 可以查到该政策的所有内容。

Id2：政策子标题索引。

Result1：知识网络结构 1。通过网络表示学习方法得到的节点向量表示。

Result2：知识网络结构 2。

NetScore：知识网络扩散分数。经过科技政策隐性扩散路径模型计算得到的结果。


```
"id": "2e01b5e62471e03b",
"type": "隐性政策知识网络路径",
"_root_": "2e01b5e62471e03b",
"resourceType": "networksim.xsd",
"from": "本地",
"title": "重庆市人民政府办公厅关于印发重庆市市级高新技术产业开发区认定和管理办法（修订）的通知",
"content": "重庆市人民政府办公厅关于印发重庆市市级高新技术产业开发区认定和管理办法（修订）的通知\n",
"language": "CH",
"fileId": "8baeb783fc37695c",
"level": "0",
"networksim_info_id2": "a08a42fbe920f846",
"networksim_info_netScore": "1.3523868793814442",
"networksim_info_id1": "119336f72162b94b",
"networksim_info_title1": "重庆市人民政府办公厅关于印发重庆市市级高新技术产业开发区认定和管理办法（修订）的通知",
"networksim_info_result1": "result19.bin",
"networksim_info_title2": "重庆市科学技术局 重庆市财政局关于印发《重庆市新型研发机构管理暂行办法》的通知",
"networksim_info_result2": "result24.bin",
"uploaderSearch": [
  "admin"
```


图 7　隐性政策知识网络路径库

再生资源库存储的数据是最终对外服务的资源。对于本研究而言，再生资源库存储的是最终生成的、用于对外服务的政策扩散路径资源。再生资源的数据组织形式是按照扩散路径生成任务具体需求构建的。

3.2 科技政策显性扩散路径生成实验结果与分析

通过从政策内容中提取出引用政策，科技政策显性扩散路径生成的搜索重点是对政策过程中提取到的政策引用标题。本文在构建资源库 Schema 时已经对字段进行了设定，其中字段间的关系是两两对应组成扩散路径对。如《浙江省科研诚信信息管理办法（试行）》，政策 Schema 中的 ParentTitle 与 Title 组成的扩散路径对。通过标题或内容进行匹配，发现两个字段之间的关系。为了提升政策显性路径生成的准确率，研究采用了基于词典的方法抽取了所有政策内容中包含的引用政策，根据政策引用标题词典对已经获取的政策标题进行处理，提升了显性政策路径生成的准确率。如表 3 所示。

3.3 科技政策隐性扩散路径生成实验结果与分析

本文的数据集来源是科技部网站、中国科技情报网以及中央人民政府网站发布的政策中的 500 篇。通过人工标注数据的方法，借助于 brat 标注工具标注数据，共标注 80 篇政策 2130 条概念，对句子中的每一个元素标注一个标签。每个元素是以标签 B 开始，标签 I 用于句子中的字符，标签 O 表示其他且用于标注无关字符。将数据的 80% 作为训练集，剩余的 20% 作为测试集，进行模型训练。本实验评价指标采用准确率（Precision）、召回率（Recall）和 F1 值。

表 3 科技政策引用标题

序号	ParentTitle	Title
1	浙江省科研诚信信息管理办法（试行）	关于进一步加强科研诚信建设的若干意见
		浙江省公共信用信息管理条例
		关于进一步加强科研诚信建设弘扬科学家精神的实施意见
		科研诚信案件调查处理规则（试行）
		国家科技计划（专项、基金等）严重失信行为记录暂行规定
2	国家科技计划（专项、基金等）严重失信行为记录暂行规定	中华人民共和国科学技术进步法
		国务院关于改进加强中央财政科研项目和资金管理的若干意见
		国务院印发关于深化中央财政科技计划（专项、基金等）管理改革方案的通知
		国务院关于印发社会信用体系建设规划纲要（2014–2020 年）的通知
3	国务院印发关于深化中央财政科技计划（专项、基金等）管理改革方案的通知	中共中央国务院关于深化科技体制改革加快国家创新体系建设的意见
		国务院关于改进加强中央财政科研项目和资金管理的若干意见

表 4　科技政策概念抽取对比实验结果

	Precision	Recall	F1
HMM	0.8009	0.7600	0.7765
CRF	0.8239	0.8488	0.8253
BILSTM	0.7962	0.8422	0.7973
CRF+BILSTM	0.8420	0.8492	0.8277

通过对单篇科技政策进行解析，将抽取的概念以及概念间关系，构建科技政策的隐性知识结构，作为网络表示学习算法的输入以及实现对科技政策单篇文本的全解析。本文以《国务院办公厅关于加快医学教育创新发展的指导意见》为例，共提取出概念 1134 个、概念关系对 1004 条，组成多个三元组，作为网络表示学习算法的输入。

生成隐性扩散路径的实验以科技政策全文为例，构建政策与政策之间的知识网络结构。本文在构建资源库 Schema 时，基于规则的方法将政策文本按照政策内容的格式“一、”“二、”等分开存入资源库。为了确保实验效果，本文使用的是政策全文，调用概念抽取与概念间关系抽取模型，关联关系特征，构建知识网络结构等，将知识网络映射成 300 维的向量，如图 8 所示，从科技政策篇章文本抽取的节点的向量表示已经融入了拓扑结构、语义和相应的标签信息，可以直接作为其它机器学习、深度学习模型的输入。

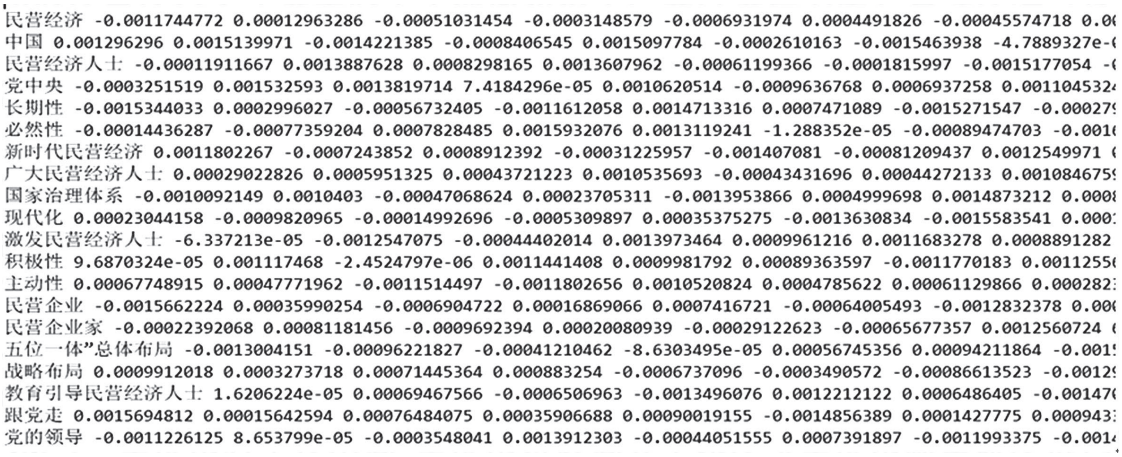

图 8　节点向量表示

科技政策隐性扩散路径的生成是对资源库中的数据借助于余弦相似度两两计算的结果，通过观察路径节点政策原文的相关性，以《浙江省科研诚信信息管理办法（试行）》为例，本文选取了与该政策的知识网络结构相似度值为前 10 的相关政策，结果如表 5 所示。

表 5　主要节点路径对

层级	政策名称（路径对）	分值
1	浙江省科研诚信信息管理办法（试行）	0.8197
2	广东省科研诚信管理办法（试行）	0.5197
3	陕西省人民政府关于进一步做好利用外资工作的实施意见	0.6071
4	国务院办公厅关于服务“六稳”“六保”进一步做好“放管服”改革有关工作的意见	0.5451
5	陕西省人民政府办公厅关于推进人工影响天气工作高质量发展的实施意见	0.5759
6	海南省加快医学教育创新发展实施方案	0.8584
7	宁夏回族自治区人民政府办公厅印发关于加快深化医学教育创新发展实施方案的通知	0.6836
8	山东省人民政府办公厅关于加快山东医学教育创新发展的实施意见	0.715
9	云南省人民政府关于进一步提高上市公司质量的实施意见	0.7749
10	海南省人民政府关于提高上市公司质量促进资本市场发展的若干意见	

从生成的科技政策隐性扩散路径结果可以看出，路径节点“浙江省科研诚信信息管理办法（试行）”与“广东省科研诚信管理办法（试行）”的关联性很强，体现这两篇政策在内容上的语义关系有一定的相关性。但是路径节点“广东省科研诚信管理办法（试行）”与“陕西省人民政府关于进一步做好利用外资工作的实施意见”的关联性很弱，通过观察这两篇政策全文，可以提取政策的主题与政策的引用政策，提升政策与政策之间知识网络结构的重合度。

通过以上方法，用户输入检索词，平台返回与检索词相关的科技政策标题，用户点击返回的科技政策标题后，系统会自动生成该篇政策的显性扩散引用路径，通过树图展示相应的结果，点击节点可以查询引用政策。同时，系统也会自动生成该政策经过科技政策扩散路径生成模型计算的隐性扩散路径关系图，每一层级对检索结果的计算分数进行排序，展示 10 条数据，并且可以查看原始资源。通过查询共有 157 条科技政策隐性扩散路径对，以人工调研《浙江省科研诚信信息管理办法（试行）》为扩散源的政策为例，证明了本文所构建科技政策扩散路径生成模型的有效性，为相关政策研究者提供一定的帮助。

4　总结与启示

本文探讨了科技政策扩散路径生成的理论方法与流程。通过构建政策文本的知识网络结构，将科技政策扩散路径的生成分为科技政策显性扩散路径和科技政策隐性扩散路径，对显性扩散路径的生成进行建模，对隐性扩散路径的生成依赖科技政策类文本的特征，研究并实现了科技政策篇章文本的概念抽取与概念间关系自动标引，提出了构建科技政策篇章文本知识网络构建方法，使用网络表示学习方法得到了科技政策的知识网络结构，在技术层面为科技政策篇章文本解析提

供了方法，为科研人员在政策研究上提供了新思路。

本文存在以下需要进一步研究和改进的地方。科技政策文本的结构化能力有待提高，在进行概念抽取时，训练模型的语料是人工标注的语料，由于对通用政策概念和领域政策概念的理解不充分，获取的概念相对不完整，应该结合先验知识对科技政策文本中的概念进行标引，构建更加完善的、细化的知识网络结构。对科技政策扩散路径的结果展示仅限于展示解析后的数据、显性扩散路径树图和隐性扩散路径关系图，没有将各个资源展示连接在一起，如利用仪表盘展示结果，充分体现政策的相关信息。

【参考文献】

[1] 刘耀 . 图书馆资源组织语义化理论及方法研究 [M]. 北京：科学技术文献出版社 ,2018:68–77.

[2] 刘耀 . 面向专业领域的情报工程技术研究与实现 [M]. 北京：科学技术文献出版社 ,2020:46–67.

[3] 李钢 . 公共政策内容分析方法：理论与应用 [M]. 重庆：重庆大学出版社 ,2007:35–45.

[4] 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.

[5] Grimmer J, Stewart B M. Text as data: the promise and pitfalls of automatic content analysis methods for political texts [J]. Political Analysis,2013, 21(3):267–297.

[6] 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.

[7] 丁洁兰 , 刘细文 , 杨立英 , 等 . 科学计量方法在科技政策研究中应用的实证研究 [J]. 图书情报工作 ,2017,61(24):77–86.

[8] Matthew C. Nowlin. Modeling issue definitions using quantitative text analysis [J]. Policy Studies Journal,2016, 44(3):1–16.

[9] Philip Leifeld, Sebastian Haunss. Political discourse networks and the conflict over software patents in Europe [J]. European Journal of Political Research,2012,51(3):382–409.

[10] Perozzi B,AlRfou R,Skiena S. DeepWalk: Online Learning of Social Representations [J]. New York: ACM,2014:701–710.

[11] Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks [J]. New York:ACM,2016:855–864.

[12] Tang J, Qu M, Wang M Z,et al. LINE: Large–scale information network embedding [J]. International Conference on World Wide Web,2015:1067–1077.

[13] Cao S S, Lu W, Xu Q K. Grarep: learning graph representations with global structural information [J]. In: Proceedings of the 24th International Conference on Information and Knowledge Management. 2015(12):891–900.

[14] Pan S, Jia W,Zhu X,et al. Tri–party deep network representation [C]. International Joint Conference on Artificial Intelligence. Palo Alto:AAAI Press,2016:1895–1901.

[15] 裴雷 , 张奇萍 , 李向举 , 等 . 中国信息化政策扩散中的政策主题跟踪研究 [J]. 图书与情报 ,2016(6): 63–71.

许乾坤，刘耀 . 科技政策扩散路径生成模型关键技术研究［J］. 文献与数据学报，2022，4（1）：035-051.

Research on Key Technologies of Science and Technology Policy Diffusion Path Generation Model

Xu Qiankun　Liu Yao

(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: ［**Purpose/significance**］By analyzing the organizational structure and semantic structure of the science and technology policy text, the potential knowledge network information in the science and technology policy text is found, and the generation model of the diffusion path of science and technology policy is constructed to make the tacit knowledge contained in the text explicit. ［**Method/process**］By analyzing the text structure of science and technology policy, the characteristic words contained in the content of science and technology policy are deeply excavated, thus transforming the unstructured text of science and technology policy into structured data. The generation of explicit diffusion path is to obtain quoted policies by analyzing the characteristics of policy content and structure, and to build the language model of policy diffusion path nodes. Implicit path generation is a vector representation of structured data of science and technology policy combined with network representation learning method, which constructs a knowledge network model of science and technology policy text, discovers potential knowledge network information in science and technology policy text, and constructs an intelligent research model for policy research. ［**Result/conclusion**］Experiments show that the generated diffusion path pair of science and technology policy verifies the effectiveness of the generation model of diffusion path of science and technology policy constructed in this paper, and provide a new research idea for the in-depth research of the content of science and technology policy.

Keywords: Science and technology policy; Diffusion path; Knowledge network structure; Path generation

（本文责编：魏　进）