融合情感语义与句法结构的中文开放域事理图谱构建研究*

赵又霖[1,2]　林怡妮[1]　石燕青[3]
（1. 河海大学商学院，南京 211100；2. 南京大学信息管理学院，南京 210023；
3. 南京农业大学信息管理学院，南京 210095）

摘要：为解决大规模开放域事理图谱构建过程中缺少标注数据以及事件类型未知导致的限定域事理图谱构建方法难以迁移的问题，利用规则匹配方法高效识别开放域文本中包含的多种事件逻辑关系，融合情感语义与句法结构信息分析提高事件抽取准确性，以更好完成事理图谱的构建任务。首先，总结并扩展因果、顺承、条件、转折等多种逻辑关系抽取模板，并基于规则模板、依存句法信息筛选逻辑关系事件句；其次，创新性地引入情感语义分析方法，在句法结构信息的基础上，通过捕获事件及事件间关系的情感语义精准识别事件类型，进而抽取事件论元；再次，计算语义相似度，进行事件融合，构建<前序事件，事件逻辑关系，后序事件>三元组，得到事件事理图谱，并进一步进行事件泛化以构建抽象事理图谱；最后，以事件发展较完整的“2022年猴痘事件”为数据源，通过实证分析证明开放域事理图谱构建方法可以实现不同类型事件的识别、事件间逻辑关系的揭露，其有效性、可行性得到验证。研究不仅弥补了现有事理图谱构建理论的不足，也为决策支持、事件发展预测等提供有力的数据支持。

关键词：开放域；事理图谱；依存句法分析；语义依存分析；情感分析
中图分类号：G254　DOI：10.3772/j.issn.1673-2286.2024.03.002
引文格式：赵又霖，林怡妮，石燕青. 融合情感语义与句法结构的中文开放域事理图谱构建研究[J]. 数字图书馆论坛，2024，20（3）：12-24.

事理图谱的构建目前已经取得一定进展，现有研究多以限定域数据为研究对象，且集中探索特定领域事件的因果和时序演化逻辑。随着互联网和社交媒体的发展，大量的开放域数据，例如网络新闻、社交媒体数据等得以产生和积累，形成广泛、多样的文本数据集群，给研究者和决策者提供了极其丰富的信息资源。然而，开放域数据往往存在缺少人工标注、数据规模大、数据质量不高的问题，这也导致开放域事理图谱的构建面临多方面挑战：一是在事件抽取过程中未充分利用句子中的语义信息，存在触发词与事件类型一对多的问题，降低事件类型分类准确性；二是在事件论元抽取过程中未充分利用论元间的相关关系，导致论元信息缺失；三是在事理图谱中未对顺承、因果、转折、条件等多种事理逻辑关系进行全方位梳理，未能充分体现事件间复杂逻辑关系。

使用自然语言处理技术，从大规模的开放域数据中抽取事件、事件关系等要素并构建事理图谱，是当前自然语言处理及知识图谱等技术的热点之一。鉴于此，

收稿日期：2023-12-18
*本研究得到江苏省社会科学基金青年项目“社会感知数据驱动下的公共卫生事件时空演化研判机制研究”（编号：20TQC001）、中国博士后科学基金特别资助“面向应急管理的时空数据语义模型构建及创新应用机理研究”（编号：2021T140311）、中国博士后科学基金面上项目“环境污染突发事件的时空数据挖掘及协同治理机制研究”（编号：2019M650108）资助。

本文提出开放域事理图谱构建框架，利用事件间因果、顺承、条件、转折等逻辑关系模板全面抽取事件间逻辑关系，并创新性地引入情感语义分析方法，融合情感语义与句法结构信息抽取事件信息，从而直观呈现事件及事件间逻辑关系，实现从海量的开放域数据中提取重要信息，以便人们更好地理解和分析事件间关联关系以及事件演变规律，为决策提供参考。

1　面向开放域的事理图谱构建现状

1.1　面向开放域的事件关系抽取

事件关系抽取的目标是从数据文本中抽取事件的逻辑关系，事理图谱注重描述事件之间深层次的变化逻辑与规律关系，准确的事件关系抽取结果是事理图谱规模和质量的保证。面向开放域的关系抽取对于事件关系类型和文本数据来源不加限制，因此其在跨领域关系抽取上具有不可比拟的优势。现有开放域关系抽取研究大多基于语句进行三元组抽取，以下两种方法应用较多。①基于句法分析与模式匹配的关系抽取法。其对规则模板的质量和覆盖度有较高要求，主要结合实体上下文信息、利用依存句法分析抽取得到全要素三元组[1]，准确率较高，但随着数据规模扩大，该方法不能很好地完成任务。②基于神经网络的事件关系抽取模型。其使用神经网络自动提取特征，较为简明方便。例如：徐康等[2]提出一种自监督学习框架，在语言学知识的基础上，利用图卷积网络模型增强文本表征效果；Du等[3]将双向门控循环单元（BiGRU）与事件内部结构特征、事件语意特征相结合以提取隐式因果关系；Xie等[4]提出一种基于交叉注意融合的图形卷积网络的开放关系抽取方法；Li等[5]提出一个基于tBERT、K-BERT模型和少样本学习的远程监督开放域关系抽取系统，解决了传统关系抽取方法在新事件关系抽取任务中表现不佳的问题；Wang等[6]设计了基于混合监督学习的事件关系抽取模型，可用于不同领域的开放域关系抽取。

1.2　面向开放域的事件抽取

事件抽取是事理图谱构建的关键环节，即从非结构化或结构化的自然语言文本中抽取并呈现事件信息，主要包括触发词识别、事件类型识别、事件论元抽取等。面向开放域的事件抽取是指在事件类型未知、事件场景不固定等情况下，从不含标注数据的开放文本数据中基于统计的思想或无监督方法对事件进行分析、检测[7]，具有较好的事件覆盖率，但难度较大，主要通过以下3类方法实现。①基于传统聚类方法。主要依靠人工提取特征，未充分利用文本语义信息，先识别出触发词和事件论元，再运用LDA模型、层次聚类算法等方法识别事件类型。例如：Sha等[8]结合事件语义关系，借用归一化切割标准的聚类算法对事件论元所属类别进行精确划分；Huang等[9]结合分布语义和符号来检测和表示事件结构，利用分布相似性对事件进行聚类，并通过多领域实验证明该方法的性能可与基于大规模训练数据的监督模型媲美。②基于句法规则匹配的方法。通过人工构建规则与模板，运用正则、语法树等技术对语料进行事件抽取，对句法分析、语义分析工具有较大的依赖性。例如：Jia等[10]总结中文的特征，研究无监督的中文开放关系提取，提出基于依存语义范式的开放域大规模实体和关系知识库，并应用于句子级事件抽取；单晓红等[11]通过构建规则模板、设定匹配规则等完成事件抽取；高李政等[12]基于Zipf's词频分布定律提出一种基于FrameNet的开放域元事件抽取模板；Gao等[13]设定自适应语义模板、扩展触发词集，在事件抽取任务中取得良好效果。③基于神经网络的方法。可以有效结合语义上下文信息，实现对特征的自动提取。陈箫箫等[14]运用LDA模型和条件随机场，提出一种基于微博数据的开放域事件抽取系统，按照时间的顺序自动抽取微博数据中被提及的事件；Wang等[15]提出一种对抗神经网络的开放域事件抽取模型，利用狄利克雷分布对事件建模，捕获潜在事件模式，从不同形式的文本中抽取事件；He等[16]结合BiLSTM-BERT模型，从新闻事件中抽取出未预先定义类型的事件；Yi等[17]提出一个基于图注意力神经网络的事件抽取模型，并应用于电影场景；Du等[18]将事件抽取视为二分类问题，结合零样本学习、神经网络框架完成事件抽取；Song等[19]将提示学习（Prompt Learning）运用于事件抽取实践。

综上所述，国内外已有研究重点关注开放域事件抽取和事件关系抽取，但针对中文开放域事件抽取和事件关系抽取的研究比较少，相应地也缺少中文开放域事理图谱的构建研究。为了实现针对中文语料的开放域事理图谱的构建，本文结合中文开放域数据特征，借鉴已有研究成果，总结并扩展出一套更为全面的包含因果、顺承、条件、转折等多种逻辑关系的抽取模板，

从而基于模式匹配方法高效抽取事件关系，并在综合运用事件语义信息、依存句法信息的基础上，引入情感分析方法，识别事件句的情感倾向，辅助区分事件类型，提高事件关系抽取准确性，从而实现并优化中文开放域事理图谱的构建。

2 面向开放域的事理图谱构建

面向开放域的事理图谱构建流程（见图1）主要包括开放域数据处理、事件关系抽取、事件抽取、事件融合、图谱可视化5个环节[20]。首先，以开放域文本数据为起点，通过数据预处理得到开放域语料库；其次，在现有因果、顺承关系规则模板的基础上，借助同义词扩展算法总结并扩展得到包含因果、顺承、条件、转折等多种关系触发词的逻辑关系抽取模板，高效抽取逻辑事件关系对；再次，在利用现有依存句法分析方法识别事件触发词、事件类型的基础上，引入情感语义分析方法，结合事件情感倾向信息提高事件类型识别准确率，进而提高触发词-事件类型对照关系识别准确率，并结合句法语义信息进一步抽取事件；最后，通过事件融合及事件泛化操作，借助Neo4j图数据库，将＜前序事件，事件逻辑关系，后序事件＞的三元组存储并可视化，得到事件事理图谱、抽象事理图谱。

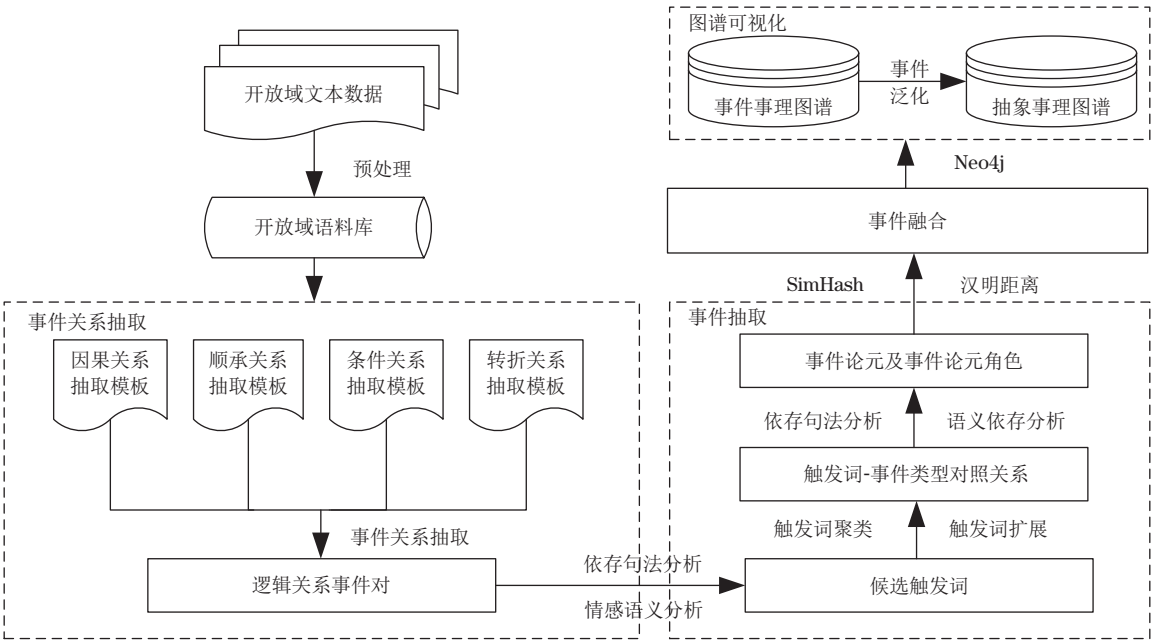

图1 面向开放域的事理图谱构建流程

2.1 基于规则匹配的事件关系抽取

现有通过深度学习方法抽取事件关系的技术依赖强有力的标注数据集，而大规模中文开放域文本存在标注不完备的情况，因此很难构造出一个没有先验知识的高级深度学习模型。同时，隐式逻辑关系成分模糊、不易提取，显式逻辑关系具有特殊的提示词，利用逻辑关系提示词生成匹配规则模板，可高效判定逻辑关系句。因此，总结现有关系抽取规则方案[21-22]，并通过同义词扩展进一步得到包含因果、顺承、条件、转折等逻辑关系的抽取模板，如表1所示。结合依存句法分析结果，基于逻辑关系抽取模板识别出事件句中的显式逻辑关系，进一步过滤否定逻辑关系，提取出前序事件、后序事件，并以＜前序语句，事件逻辑关系提示词，后序语句＞的三元组形式表示。

（1）语料预处理。对开放域语料的预处理包括：①分句，以终结符号为依据对文本进行划分，并进行短句处理，即将过短的句子与前一句进行合并或结合上下文补齐句子成分，以避免短句导致的意义不明或产生歧义等问题；②分词及词性标注，利用pkuseg工具对文本进行分词，运用哈工大语言技术平台（Language Technology Platform，LTP）对分词结果进行词性标注，词性含义如表2所示。

表1　逻辑关系抽取模板

逻辑关系类型	句法模式	匹配规则模板	逻辑关系提示词（部分）
因果关系	由果溯因配套式	＜Conj＞{Effect}，＜Conj＞{Cause}	＜[之]所以，因为|由于|缘于＞
	由因到果配套式	＜Conj＞{Cause}，＜Conj＞{Effect}	＜因[为]|由于，从而|为此|所以|因此|以至[于]|以致[于]|因而|于是|故|故而|致使＞、＜既[然]，所以|却|就|因此＞、＜只有|除非，才＞、＜如果，那么|则＞、＜如果|只要，就＞
	由因到果居中式明确	{Cause}，＜Conj＞{Effect}	于是、所以、故、致使、以致[于]、因此、以至[于]、从而、因而
	由因到果居中式精确	{Cause}，＜Verb|Adverb＞{Effect}	波及、促[进|发|成|使]、触发、牵动、产生、带来、导[向|致]、勾起、指引、引[出|发|导|入|起|来|致]、使[得|动]、渗[入|透]、予以、酿成、推[进|动]、影响、滋生、归于、决定、攸关、令人、浸染、挟带、关系、诱[发|惑|使|导|致]、造[成|就]、招致、致使、作用
	由因到果前端式模糊	＜Prep＞{Cause}，{Effect}	为[了]、依据、按照、因[为]、按、依赖、照、比、凭借、由于
	由因到果居中式模糊	{Cause}，＜Verb|Adverb＞{Effect}	以免、以便、为此、才
	由因到果前端式精确	＜Conj＞{Cause}，{Effect}	既[然]、因[为]、如果、由于、只要
	由果溯因居中式模糊	{Effect}＜Prep＞{Cause}	[在|出|缘]于、[出|来|发]自、[根|来|起|发]源于、取决[于]、立足[于]
	由果溯因居端式精确	{Effect}＜Conj＞{Cause}	因为、由于
顺承关系	顺承词配套式	＜Conj＞{Event1}，＜Conj＞{Event2}	＜又|再|才|并，进而＞、＜首先|第一|先，其次|然后＞、＜首先|先是，再|又|还|才＞、＜一方面，另一方面|又|也|还＞、＜在，后|之后＞、＜起先，随后＞
	顺承词居端式明确	＜Conj＞{Event1}，{Event2}	首先、起先
	顺承词居中式明确	{Event1}＜Conj＞，{Event2}	就、才、便、于是、后来、其次、[从|继]而、[而|随|之|然]后、接下来
	顺承词居中式精确	{Event1}＜Verb|Adverb＞{Event2}	跟着、接着、可见、意味、标志
条件关系	条件词配套式	＜Conj＞{Event1}，＜Conj＞{Event2}	＜除非，否则|才|不然|要不|否则的话＞、＜还是|无论|不管，还是|都|总＞、＜既然，又|且|也|亦|那么＞、＜一旦，就＞、＜只有，才|还＞、＜万一，那么|就＞、＜只要，就|便|都|总＞、＜即使|就是，也|还是＞、＜尽管|假若|假使|如果|要是|假如|既然|既|如|假设，那么|就|那|则|便|的话|也|还＞
转折关系	转折词配套式	＜Conj＞{Event1}，＜Conj＞{Event2}	＜虽然|纵然|即使|尽管|就算是|即便|固然|虽说，倒|还是|但是|但|可是|可|却|但却|但也|但是也|但是还|但是却|也|还|然而|仍然＞、＜无论|不管|不论|即使，都|也|还|仍然|总|始终|一直＞、＜与其，宁可|不如|宁肯|宁愿＞、＜与其|宁可|宁肯|宁愿，决不|也不|也要＞、＜不是，而是＞

表2　哈工大LTP词性列表

标记	词性	标记	词性	标记	词性
a	形容词	n	名词	nz	其他专名
b	名词修饰词	nh	人物名词	p	介词
c	连词	ni	机构名词	q	量词
d	副词	ns	地理名词	v	动词
m	数词	nt	时间名词	wp	标点符号

利用肯德尔协调系数验证分词及词性标注结果的一致性。将标注好的样本词语集定义为集合X，寻找并修改集合X中分词结果或词性标注结果存在歧义的词语，将修改后的词语集定义为集合Y，利用式（1）计算肯德尔协调系数W。

$$W=\frac{2(C-D)}{N(N-1)} \quad (1)$$

式中：N为集合X和集合Y的元素个数；C为集合X和集合Y中一致的元素对数（两个元素为一对）；D为集合X和集合Y中不一致的元素对数。

（2）事件逻辑关系抽取。根据设计的事件逻辑关系抽取规则模板，以正则匹配的方式对事件文本语料进行逻辑关系判断与提取。利用规则模板判断句子中是否包含逻辑关系提示词，若包含则利用该逻辑关系提示词区分前序事件和后序事件。在逻辑关系抽取输出结果的基础上，对抽取结果进行清洗，例如过滤否定逻辑关系、剔除字数低于8个字或表意不完整的逻辑关系、补充指代不明的逻辑关系等，以保证抽取出来的句

子在具有逻辑关系的同时，符合一定的表达规范[23]。

对事件逻辑关系清洗结果进行逐句判断：①是否含有逻辑关系提示词；②在含有逻辑关系提示词的前提下，是否存在真实事件逻辑关系；③是否存在关系子句，若存在则进一步抽取子句中的事件逻辑关系。

进一步进行准确率（Precision）、召回率（Recall）、F1值计算，判断事件逻辑关系抽取效果。如表3所示：TP指句子中存在事件逻辑关系，且正确抽取出该事件逻辑关系；FN指句子中存在事件逻辑关系，但未正确抽取出该逻辑关系；FP指句子中不存在事件逻辑关系，却抽取出错误的逻辑关系；TN指句子中不存在事件逻辑关系，且未抽取出逻辑关系。

表3　准确率、召回率计算规则

类型	存在事件逻辑关系	不存在事件逻辑关系
抽取出事件逻辑关系	TP（True Positive）	FP（False Positive）
未抽取出事件逻辑关系	FN（False Negative）	TN（True Negative）

准确率P是指在所有抽取出事件逻辑关系的句子中真实存在逻辑关系的句子所占的比例，如式（2）所示。

$$P=\frac{n_{TP}}{n_{TP}+n_{FP}} \tag{2}$$

式中：n为句子数量。

召回率R是指在所有存在事件逻辑关系的句子中，抽取出逻辑关系的句子所占的比例，如式（3）所示。

$$R=\frac{n_{TP}}{n_{TP}+n_{FN}} \tag{3}$$

F1值S_{F1}是对准确率和召回率的综合分析，如式（4）所示。

$$S_{F1}=\frac{2\times P\times R}{P+R} \tag{4}$$

2.2　融合情感语义与句法结构的开放域事件抽取

事件抽取的主要任务是确定并提取事件的构成要素，包括触发词、事件类型、论元、论元角色。本节将基于事件逻辑关系对，重点关注开放域事件触发词和事件论元抽取。①构建种子触发词库，利用哈工大LTP对事件句进行依存句法分析，根据分析结果抽取出事件句中的触发词。②识别事件类型，根据事件句情感倾向与触发词聚类结果，对事件进行准确分类，从而构建事件触发词-事件类型对照表。③抽取事件论元，根据触发词对应的事件类型，进一步通过依存句法分析、语义依存分析，确定并提取事件论元及论元角色。

（1）基于依存句法分析的事件种子触发词库构建。利用哈工大LTP对事件句进行依存句法分析，依存关系标签如表4所示。其中，SBV为主谓关系，VOB为动宾关系，通过触发词抽取代码可判断句子主谓关系中的动词、动宾关系中的动词与当前判断的动词是否一致，若一致则抽取当前判断的动词作为候选触发词[24]。根据依存句法分析结果，抽取出符合的动词作为候选触发词。

表4　哈工大LTP依存关系标签

标签	描述	关系类型
SBV	Subject-Verb	主谓关系
VOB	Verb-Object	动宾关系
ATT	Attribute	定中关系
CMP	Complement	动补关系
POB	Preposition-Object	介宾关系
ADV	Adverbial	状中关系

获得的候选触发词中可能存在较多无意义的词，影响事件类型判断的正确性[24]。①出现频率较低，不足以说明某类事件的发生，例如能够、感到、讨论、持续等。②动词不足以成为句子的依存结构，例如有、是、应该、作为、包括等。根据上述两个触发词过滤规则对候选触发词进行过滤，得到种子触发词库。

（2）结合种子触发词聚类与情感语义分析的事件类型识别。基于K-means算法对种子触发词进行聚类，综合考虑肘部法则、轮廓系数结果以确定最佳聚类簇数k，达到最佳的聚类效果。考虑到特定领域内相似事件类型中可能包含较多相同或相似的触发词，仅基于触发词识别不同事件类型可能存在难以准确分类、易产生歧义等问题，而事件句的情感倾向、触发词的直接主语或直接宾语对事件描述也存在一定的影响，因此在分析种子触发词与句子的主谓宾结构词义相似度的基础上，引入情感分析方法挖掘事件句情感倾向信息。将事件句的情感倾向融入聚类过程，捕捉事件句情感特征，辅助提高聚类精度，从而优化事件类型识别准确性。百度AI开放平台的情感倾向分析接口结合了情感词典和机器学习方法，文本情感分类通用准确率高达92%[25]，因此利用该接口对预处理后的文本进行自动

情感分类（负向、中性、正向），并给出分类置信度。

触发词是指最能够代表事件发生、表征事件重要特征的词[26]，种子触发词包含的触发词不足以覆盖所有语料，因此借助同义词扩展算法扩展种子触发词，可以提高事件抽取的准确性。为避免在扩展过程中引入较多相关程度低的词语[27]，通过计算词语相似度将相似词语纳入触发词库。此处将相似度大于0.7的词语（词义相似度取值为[0，1]）扩展到种子触发词库中，形成事件触发词库。扩展的触发词所对应的事件类型与该种子触发词所对应的事件类型一致，构建得到最终的触发词-事件类型对照关系。

（3）结合依存句法分析与语义依存分析的事件论元抽取。开放域数据一般为非结构化数据，文本结构不规则或不完整，且部分事件之间相互嵌套、成分共用，只提取其中的主谓宾成分作为事件论元可能导致事件表示不完整，甚至产生歧义。此外，开放域数据中缺乏有效的人工标记数据，这极大地影响了机器学习、深度学习事件抽取方法的效果。因此，借助哈工大讯飞语言云和哈工大LTP，利用其依存句法分析功能捕获事件句核心成分，运用其语义依存分析功能补全事件语义、增强词间固有的联系，得到更全面完整的事件论元。语义依存关系标签及释义如表5所示。

表5　哈工大LTP语义依存关系标签

标签	标签释义	标签	标签释义	标签	标签释义	标签	标签释义
AGT	施事关系	LINK	系事关系	REAS	缘由角色	FEAT	修饰角色
EXP	当事关系	TOOL	工具角色	TIME	时间角色	mPUNC	标点标记
PAT	受事关系	MATL	材料角色	LOC	空间角色	mNEG	否定标记
CONT	客事关系	MANN	方式角色	MEAS	度量角色	mRELA	关系标记
DATV	涉事关系	SCO	范围角色	STAT	状态角色	mDEPD	依附标记

依存句法分析从整体角度研究事件句中分词单位间的依赖关系，不受分词单位间物理位置的影响。谓语动词依赖于根节点，是句子的核心成分，其他句子成分与其有直接或间接关联，并通过成分间的关联性形成一棵依存树。利用依存句法分析抽取事件句核心成分[28]，其相应的抽取规则如表6所示，利用该抽取规则可有效解决事件句成分缺失问题。①若句中存在主谓宾关系，则以谓语动词为核心成分，通过识别主谓关系、动宾关系得到主谓宾成分。②若句中存在动宾关系，则以谓语动词为核心，通过识别定中关系得到主谓宾成分。③若句中存在主谓动补关系，则以谓语动词为核心，通过主谓关系、动补关系识别前序论元，通过介宾关系识别后序论元。

表6　基于依存句法关系的句子成分抽取规则

序号	依存句法关系	抽取规则
1	主谓宾关系	postags[index]= ‘v’ ‘SBV’ and ‘VOB’ in dict（e1，r，e2）
2	动宾关系，定语后置	postags[index]= ‘v’ relation= ‘ATT’ ‘VOB’ in dict（e1，r，e2）
3	主谓动补关系	postags[index]= ‘v’ ‘SBV’ and ‘CMP’ in dict（e1，r） ‘POB’ in dict（e2）

语义依存分析主要刻画词与词之间的语义依存，可以简单明了地给出论元之间的关联，从而使语句中的语义信息得到更为完整和全面的描述。语义依存分析能够为依存句法分析结果提供补充语义分析，但与依存句法分析结果可能存在矛盾[29]，因此，重点考虑利用依存句法分析确定事件句中的核心成分，即主语、谓语、宾语等成分，并基于语义依存关系识别各核心成分的边界、补全事件语义，得到更加完整的事件论元。

2.3　基于SimHash和汉明距离的事件融合

事件融合又称事件互指融合，是指通过合并互补信息、补充缺失内容、剔除冗余信息等方式，将反映同一事件的多条数据融合成一条数据。由于数据来源多样、事件组成要素多样、文字形式多样等，所抽取的事件可能存在事件论元要素不完整、事件信息相互补充、事件信息互相矛盾、事件信息重复等问题，必须通过事件融合去除冲突、错误或冗余等信息，降低事件抽取的不确定性，从而获得更为丰富完备的事件论元结构信息。

考虑到开放域数据规模大、需要快速计算相似度，并且重要的文本处理在事件抽取部分已经完成，选择计算效率高、资源消耗低的SimHash及汉明距离（Hamming Distance）方法进行事件融合。事件融合主要分为3个阶段：①事件编码，事件融合通常在同一

类型事件间进行，利用SimHash算法可对同一类型事件群内的事件文本进行编码[30]；②相似度计算，计算事件编码后的汉明距离，通过比较候选事件间的汉明距离与预先设定的相似度阈值，判断事件间的相似性；③事件融合，将相似度超过阈值的事件合并为其中发生频率最高的事件。

2.4 开放域事理图谱可视化

（1）事件事理图谱的构建。经过事件逻辑关系抽取、事件抽取、事件融合后，可以得到事件三元组<pre_event，relation，post_event>，即<前序事件，事件逻辑关系，后序事件>。使用py2neo库进行图数据库操作以将三元组存储于Neo4j图数据库中：①调用Graph函数配置Neo4j图数据库的相关参数及用户信息；②利用load csv指令，将构建的事件逻辑关系三元组CSV文件数据导入Neo4j存储；③如图2所示，调用Node函数和Relationship函数依次将各个三元组中的前序事件、后序事件数据转换为Neo4j的节点数据，将事件逻辑关系转换为Neo4j的边数据；④利用Cypher语句MATCH（n）RETURN n实现节点和边在Neo4j图数据库中的图谱展示。

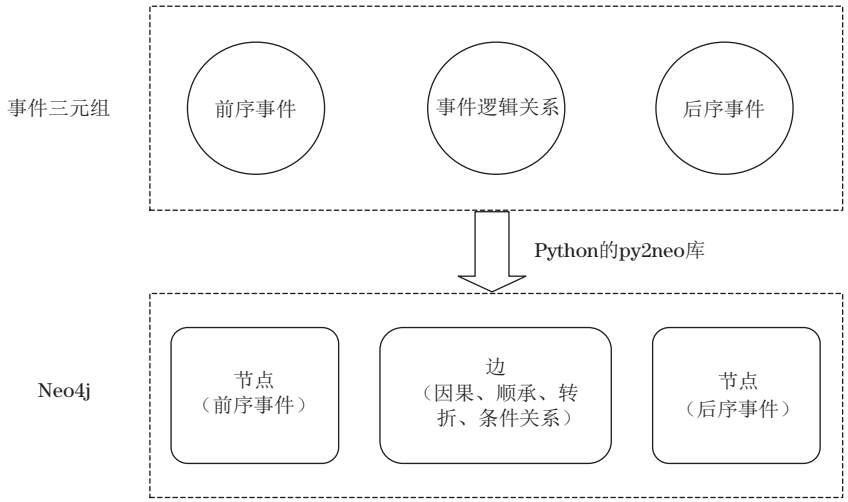

图2 三元组转换

构建得到的事件事理图谱根据其内容与涵盖的领域，可应用于事件相关信息的查询与分析，例如追溯事件源头、追踪事件的发展演变与影响，分析事件的传播路径、发展规律、影响范围以及事件之间的潜在关联关系，为事件未来发展提供预测、为应对策略的制定提供支持。

（2）抽象事理图谱的构建。事件泛化是构建抽象事理图谱的关键，通过抽取具体事件的共性或关键特征，可以把复杂逻辑关系由具体事件信息抽象为更一般、更通用的形式，将事件事理图谱泛化为抽象事理图谱，进而揭示出更具有普遍性的复杂逻辑关系演变规律，提高逻辑关系的代表性。具体事件泛化后得到抽象事件，即事件类型。通过事件抽取过程中聚类得到的事件类型，获取事件句中前序事件、后序事件所对应的原因类型。

事件逻辑关系的泛化通过归纳完成，如事件a和事件b构成三元组<a，因果关系，b>，则a对应的抽象事件A和b对应的抽象事件B之间构成因果关系三元组<A，因果关系，B>。为计算抽象事件间的概率，在泛化归纳时还需统计各类事件逻辑关系的数量：①基于具体事件三元组，将所有前序语句、后序语句替换为其所属事件类型，得到抽象事件三元组；②删除前序语句、后序语句所属事件类型相同的抽象事件三元组；③统计抽象事件三元组事件逻辑关系数量作为抽象事理图谱中边的权重，进而计算该事件逻辑关系的概率，为事件的预测提供数据支持，如式（5）所示。

$$c=\frac{\mathrm{Count}\left(e_i,r,e_j\right)}{\mathrm{Count}\left(e_i\right)} \tag{5}$$

式中：c为逻辑关系的概率；e_i为泛化后的前序事件；e_j为泛化后的后序事件；r为事件间逻辑关系；$\mathrm{Count}(e_i,r,e_j)$为事件$e_i$、$e_j$同时出现且存在关系$r$的次数；$\mathrm{Count}(e_i)$为$e_i$作为前序事件出现的次数。

3　实证研究

3.1　数据来源与处理

猴痘事件（2022年5月—2023年7月）发展较为完整，且在新闻平台、社交媒体平台上引发广泛关注，因此以其为研究事例具有一定代表性。以“猴痘”为检索关键词，利用Scrapy爬虫框架爬取中国新闻网、人民网、新华网等影响力较高平台关于2022年猴痘事件的官方媒体报道。爬取时间为2023年7月31日，事件时间跨度为2022年5月7日—2023年7月30日，得到共计944篇新闻文本，经筛选去重后保留有效新闻845篇。对采集到的语料集进行筛选并分句，得到共13 299个句子。鉴于pkuseg工具的分词结果比jieba和哈工大LTP更准确、对领域专有词汇的划分和识别能力更好，利用pkuseg数据进行分词，并通过人工额外构建自定义猴痘疫情词典，对领域专有词汇进行特殊标记，增强歧义纠错能力。利用式（1）计算得到样本数据W=0.923，在0.8~1.0范围内，说明分词及词性标注的一致性较强。进一步根据歧义修改自定义词典，并将该标注过程应用于所有数据，得到所有数据的词性标注结果，为后续实验做好准备工作。

3.2　猴痘事件关系抽取与事件抽取

（1）猴痘事件关系抽取。根据表1的复合逻辑关系抽取模板抽取并筛选抽取结果，最终抽取到1 447个复合逻辑关系，其中因果关系824个、顺承关系31个、转折关系431个、条件关系161个，抽取到的部分逻辑关系如表7所示。根据式（2）~（4），计算得到猴痘事件逻辑关系抽取的准确率、召回率、F1值，如表8所示，结果显示关系抽取效果较好。

表7　事件逻辑关系抽取结果（部分）

原文本	前序语句	事件逻辑关系	后序语句
事实上，迄今为止，全球变暖导致的物种全球迁徙表明，病毒溢出可能已经在进行中	全球变暖	因果关系	物种全球迁徙，病毒溢出可能已经在进行中
按照合同，该公司将把根据与美方先前合同已经生产的疫苗，转化为冻干疫苗，从而延长其保质期	把已经生产的疫苗，转化为冻干疫苗	顺承关系	延长其保质期
目前，我国尚无猴痘病例报道，也未在野生动物或入境检疫中检出猴痘病毒，但随着全球旅行及贸易的恢复，不排除我国发生输入病例的可能性	尚无猴痘病例报道，也未在野生动物或入境检疫中检出猴痘病毒	转折关系	随着全球旅行及贸易的恢复，不排除我国发生输入病例的可能性
6月15日，圣湘生物相关负责人对第一财经记者表示，目前国内还未开通针对猴痘核酸检测试剂的注册通道，如果注册通道开通了，公司不排除会申请相关核酸试剂在国内上市	针对猴痘核酸检测试剂的注册通道开通	条件关系	公司不排除会申请相关核酸试剂在国内上市

表8　事件逻辑关系抽取效果

单位：%

类型	准确率	召回率	F1值
因果关系	90.86	95.08	92.26
顺承关系	91.92	93.53	92.72
转折关系	90.31	93.73	91.99
条件关系	88.70	90.28	89.48
总体	90.48	94.11	92.26

（2）猴痘事件抽取。在复合关系抽取得到的事件句的基础上构建触发词库，获得309个候选触发词。剔除猴痘疫情相关领域出现频率较低的候选词，例如能够、感到、讨论、持续等，最终得到257个种子触发词，例如：储备、帮助、保护、报道、报告、避免、变化、变异、表明、表现、病变、波动、补充、采集、出现、处理、传播、传染、担忧、登记、低估、调查、调整、订货、动力、短缺、恶化、发热、发生、发现、发展、反应、防范、防护、放缓、分发、服务、改变、损伤、感染、干扰、隔离、公开、关联、关注、观察、忽视、患病、加大、加剧、假设、监测、监督、检测、检查、建议、教训、接触、接近、接种、结痂、介绍、警惕、距离、考试、恐慌、扩散、了解、落实、蔓延、排除、排队、判断、喷洒、批评、批准、匹配、评估。

根据触发词、句子主谓宾结构及情感倾向，确定K-means算法的聚类簇数k时综合考虑轮廓系数和簇内误方差（SSE），如图3~4所示。当k=6时，簇内误方差的下降速度减缓且轮廓系数较大，因此，将猴痘事件归并为六大类抽象事件，分别为疫情检测、社会影响、医疗资助、调查研究、政策发布、经济影响。分析聚类准确度时发现疫情检测、调查研究、经济影响三大类仍可进一步聚类细分，结果如表9所示。

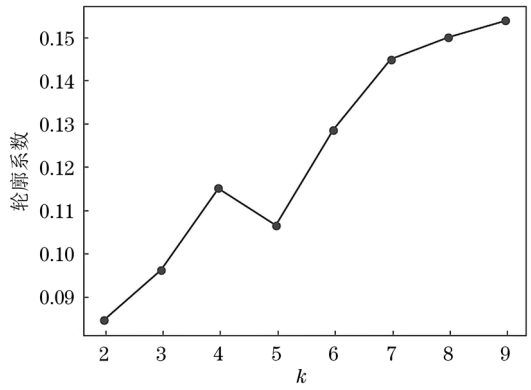

图3　聚类簇数与轮廓系数的关系

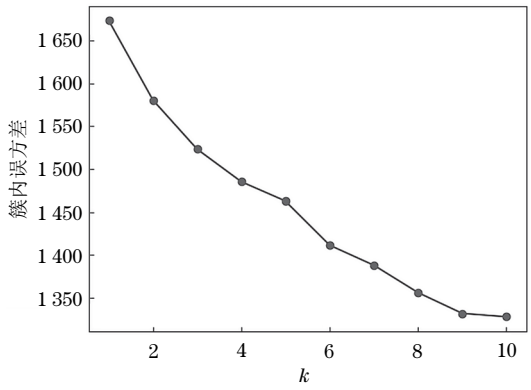

图4　聚类簇数与簇内误方差的关系

表9　事件类型聚类准确度实验结果

单位: %

类别	事件类型	F1值	
		依存句法分析	依存句法（主谓宾结构）+情感倾向
疫情检测	疫情发生	73.6	80.3
疫情检测	疫情传播	77.3	86.6
疫情检测	治疗控制	78.1	82.2
社会影响	社会影响	74.8	77.0
医疗资助	医疗资助	75.5	80.3
调查研究	疫源调查	72.9	75.0
调查研究	信息报告	72.4	74.3
政策发布	政策发布	71.5	74.9
经济影响	股价波动	83.1	79.2
经济影响	产品研发/储备	78.8	79.6
经济影响	产品采购	79.4	80.1

表9显示，引入情感倾向的句法语义分析方法的聚类准确度普遍较高，最高F1值达到86.6%，说明结合依存句法信息、情感倾向信息的方法可提高事件类型识别准确性和事件抽取准确性。同时利用同义词扩展算法对触发词进行扩展，得到猴痘触发词-事件类型对照表，如表10所示。在得到触发词的基础上，利用依存句法分析和语义依存分析，抽取＜前序事件，事件逻辑关系，后序事件＞三元组，并存储于Neo4j数据库。

表10　事件类型及对应触发词

序号	类别	事件类型	事件触发词（部分）
1	疫情检测	疫情发生	确诊；增加；发展；出现；感染；爆发；突发；累计；平息；已有
2	疫情检测	疫情传播	波及；传播；接触；导致；蔓延；扩散；密接；流调；追踪；发酵
3	疫情检测	治疗控制	隔离；诊治；注射；控制；检测；转阳；转阴；接种；观察；预防
4	社会影响	社会影响	扰乱；报道；引发；登上
5	医疗资助	医疗资助	支援；抽调；调转；增派；互派；发放；补给；捐赠；调度；分发
6	调查研究	疫源调查	确认；鉴别；测定；操作；搭配
7	调查研究	信息报告	介绍；称；显示；宣布；确认；强调；发布；回应；回复；发稿
8	政策发布	政策发布	调整；发布；公布；批转；转发；颁发；印发
9	经济影响	股价波动	下跌；受创；亏损；紧缩；涨；触及；开盘；投资；看好；投资
10	经济影响	产品研发/储备	储备；构建；注册；推进；取得；实现；供应；组织；研发；生产
11	经济影响	产品采购	采购；订购；销售

3.3　事理图谱可视化

依循上文事件融合、事理图谱可视化步骤，得到猴痘事件事理图谱，该图谱中共有994个节点、970条边，事件事理图谱局部如图5所示。

图5中，“猴痘疫情爆发”“猴痘病毒传播”“应对猴痘疫情”3个中心节点是猴痘事件发展过程中较为关键的事件，边代表事件之间的逻辑关系，包括因果关系、顺承关系、转折关系、条件关系。其中，“猴痘疫情爆发”和“应对猴痘疫情”属于重要的原因事件，该事件向外扩散导致后续其他事件的发生。造成“猴痘病毒传播”事件的原因事件较多，包括“国际传播”“抗疫躺平”“密切接触”等。相关部门应针对这些原因事件提前准备应急措施，防止应对措施缺乏造成的公众恐慌。在联系较为紧密的逻辑关系事件对中，前序事件的发生在很大概率上会引起后一个事件。通过Cypher语句MATCH（n：‘感染猴痘病毒’）RETURN n，可以得到“感染猴痘病毒”子事件事理图谱（见图6）。由

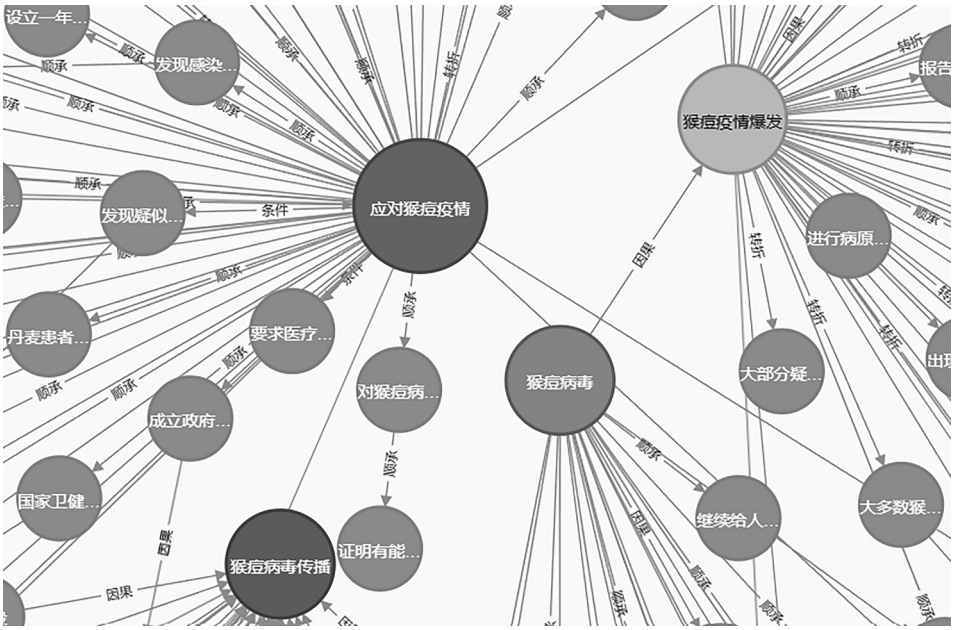

图5　猴痘事件事理图谱（局部）

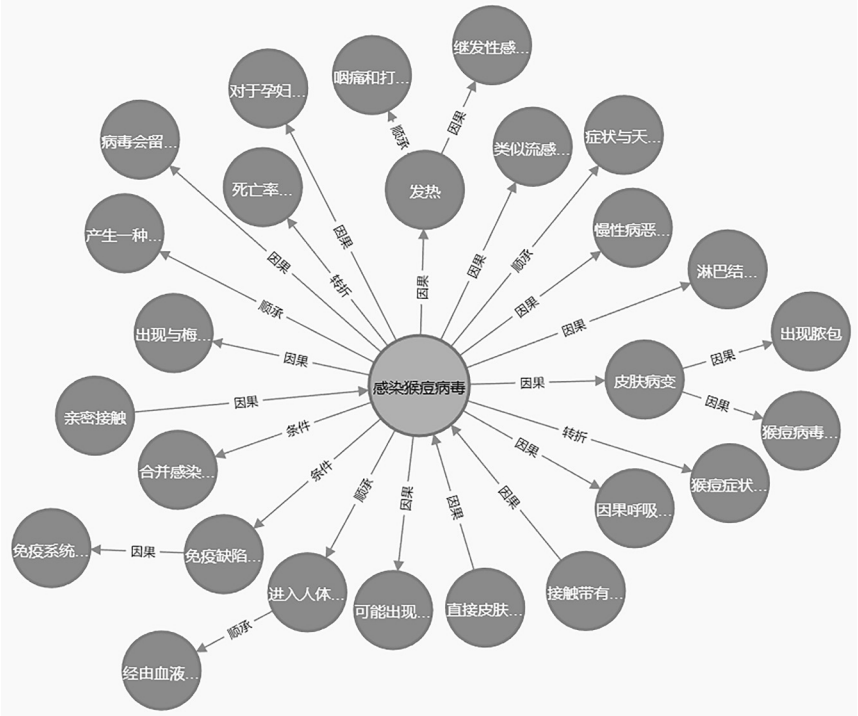

图6　“感染猴痘病毒”节点事件事理图谱

该图可知，“感染猴痘病毒”的后序事件大多为感染病毒后出现的症状，例如“引发呼吸道疾病”“淋巴结肿大”“皮肤病变”等。“继发性感染”和“出现脓包”属于结果事件，结果事件的发生预示着事件链即将结束。相关部门需提前制定有效的应对措施，尽量避免事件朝负面方向发展。

事件事理图谱包含较多事件及事件间逻辑关系，难以揭示普遍性的事件演变规律，通过事件泛化将猴痘事件抽象为疫情发生、疫情传播、治疗控制、社会影响、医疗资助、疫源调查、信息报告、政策发布、股价波动、产品研发/储备、产品采购11类事件。以泛化事件为节点，按式（5）计算事件转移概率。由于各类型事件间

均存在多方向交叉的关系，泛化处理后各泛化事件之间均存在一定程度上的联系。为便于分析，省略转移概率小于0.12的逻辑关系，得到猴痘事件抽象事理图谱（见图7）。

事件关系转移概率揭示了猴痘疫情发展和防控过程中各个环节之间的相互作用和影响。例如，“社会影响”事件对“股价波动”事件的影响显著，存在0.20的因果关系转移概率和0.20的转折关系转移概率，说明社会舆论或公众对事件的看法可以直接影响股市，造成股价的上涨或下跌。此外，转折关系转移概率的存在说明“社会影响”在某些情况下也可能对“股价波动”造成意料之外的影响。“医疗资助”事件对“社会影响”“疫情传播”和“疫情发生”事件均有0.14的因果关系转移概率，凸显出医疗资助在疫情防控中的重要作用，比如医疗资助可以控制疫情的传播和发生，同时也可以加强公众对防控工作的信心，有效遏制疫情的蔓延。综合事件转移概率来看，通过加强信息报告、医疗资助、政策发布、产品采购和研发等环节的工作，可以有效提高治疗控制的效率，进而控制疫情的传播和发生。同时，这些分析结果也可为决策者提供参考，帮助制定更科学有效的疫情防控策略，以控制疫情的蔓延并减轻其对社会和经济的影响。

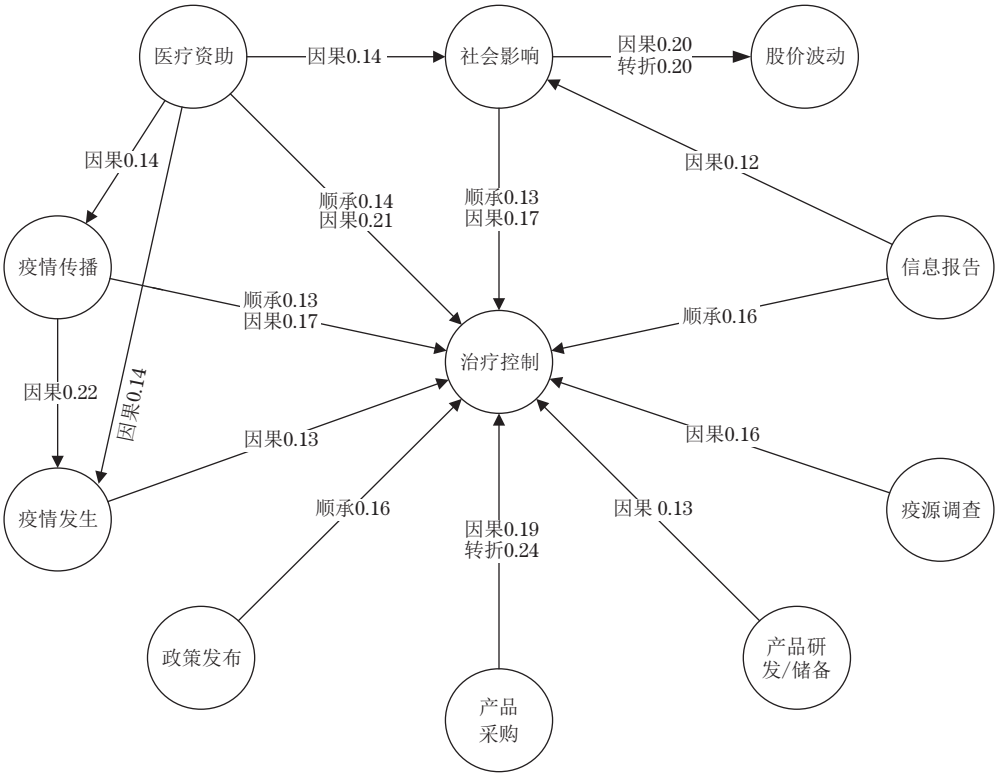

图7　猴痘事件抽象事理图谱

4　结语

本文提出了一种面向开放域的事理图谱构建方法，该方法结合中文开放域文本特征，通过一系列步骤实现了面向开放域的、全面高效的事件关系抽取、事件抽取以及事理图谱构建。首先，聚焦于开放域文本数据中的因果、顺承、条件、转折等多种显式逻辑关系，构建了逻辑关系抽取模板库，从而高效识别事件关系并生成事件三元组；其次，引入情感倾向分析优化事件分类准确性，并结合依存句法信息、语义依存信息抽取完整的事件信息；再次，借助SimHash和汉明距离对同类型中相似度高的事件群做事件融合处理，进一步提升图谱质量；最后，利用Neo4j图数据库实现事件事理图谱的可视化，并通过事件泛化得到抽象事理图谱。通过实证研究，验证了该开放域事理图谱构建方法的可行性，并展示了其在猴痘事件中的应用效果。该中文开放域事理图谱构建框架的搭建方法突破了以往限定域事理图谱构建方法的局限性，有助于更全面、深入地理解和

分析中文开放域数据中的事件及事件关系，丰富了事理图谱构建理论体系。同时，通过构建全面、准确的开放域事理图谱，可以直观地展示事件的演变过程，有助于更好地掌握事件发展脉络和演化模式，为决策支持、事件预测、风险管理等提供有力的信息支持。

未来可在以下方面继续深入研究：①深入挖掘隐式事件关系，基于规则模板抽取的逻辑关系多为显式关系，未充分抽取文本中存在的隐式关系；②根据历史事件数据计算在前序事件发生基础上后序事件的发生概率，可帮助预测事件未来的发展方向，从而管控、引导事件的演变。

参考文献

[1] 温清华，朱洪银，侯磊，等. 多策略中文开放关系抽取方法[J]. 中文信息学报，2023，37（1）：88-96.

[2] 徐康，余胜男，陈蕾，等. 基于语言学知识增强的自监督式图卷积网络的事件关系抽取方法[J]. 数据分析与知识发现，2023，7（5）：92-104.

[3] DU J W，ZHAO H R，YU Y Y，et al. A method to extract causality for safety events in chemical accidents from fault trees and accident reports[J]. Computational Intelligence and Neuroscience，2020，2020：7132072.

[4] XIE B H，LI Y，ZHAO H Y，et al. A cross-attention fusion based graph convolution auto-encoder for open relation extraction[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2022，31：476-485.

[5] LI H G，LIU B. An open relation extraction system for web text information[J]. Applied Sciences，2022，12（11）：5718.

[6] WANG X X，HU J P. An open relation extraction method for domain text based on hybrid supervised learning[J]. Applied Sciences，2023，13（5）：2962.

[7] 胡瑞娟，周会娟，刘海砚，等. 基于深度学习的篇章级事件抽取研究综述[J]. 计算机工程与应用，2022，58（24）：47-60.

[8] SHA L，LI S J，CHANG B B，et al. Joint learning templates and slots for event schema induction[EB/OL]. [2023-11-12]. https://arxiv.org/abs/1603.01333.

[9] HUANG L F，CASSIDY T，FENG X C，et al. Liberal event extraction and event schema induction[EB/OL]. [2023-11-12]. https://aclanthology.org/P16-1025.pdf.

[10] JIA S B，SHIJIA E，LI M Z，et al. Chinese open relation extraction and knowledge base establishment[J]. ACM Transactions on Asian and Low-Resource Language Information Processing，17（3）：15.

[11] 单晓红，庞世红，刘晓燕，等. 基于事理图谱的网络舆情事件预测方法研究[J]. 情报理论与实践，2020，43（10）：165-170.

[12] 高李政，周刚，黄永忠，等. 基于Zipf's共生矩阵分解的开放域事件向量计算方法[J]. 计算机科学，2020，47（10）：207-214.

[13] GAO J Q，LUO X F，WANG H. An uncertain future：predicting events using conditional event evolutionary graph[J]. Concurrency and Computation：Practice and Experience，2021，33（9）：e6164.

[14] 陈箫箫，刘波. 微博中的开放域事件抽取[J]. 计算机应用与软件，2016，33（8）：18-22，109.

[15] WANG R，ZHOU D Y，HE Y L. Open event extraction from online text using a generative adversarial network[EB/OL]. [2023-11-12]. https://arxiv.org/abs/1908.09246.

[16] HE L，ZHANG Q，DUAN J Y，et al. An open-domain event extraction method incorporating semantic and dependent syntactic information[J]. Applied Sciences，2023，13（13）：7942.

[17] YI Q，ZHANG G X，LIU J，et al. Movie scene event extraction with graph attention network based on argument correlation information[J]. Sensors，2023，23（4）：2285.

[18] DU X，ZHANG Z，LI S，et al. Resin-11：schema-guided event prediction for 11 newsworthy scenarios[C]//Conference of the North-American-Chapter-of-the-Association-for-Computational-Linguistics（NAAACL）-Human Language Technologies，2022：54-63.

[19] SONG C Y，CAI F，ZHENG J M，et al. AugPrompt：knowledgeable augmented-trigger prompt for few-shot event classification[J]. Information Processing & Management，2023，60（4）：103153.

[20] 王毅，沈喆，姚毅凡，等. 领域事件图谱构建方法综述[J]. 数据分析与知识发现，2020，4（10）：1-13.

[21] QIU J N，DU Y J，WANG Y Z. Extraction and representation of feature events based on a knowledge model[C]//2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology，2008：219-222.

[22] 李彭伟，李亚钊. 面向事件画像的事理图谱构建方法[J]. 指挥信息系统与技术，2021，12（5）：54-60，69.

[23] 刘政昊，曾曦，张志剑. 面向应急管理的金融突发事件事理知识图谱构建与分析研究[J]. 信息资源管理学报，2022，12（3）：137-151.

[24] DING X，QIN B，LIU T. Building Chinese event type paradigm based on trigger clustering[C]//Proceedings of the Sixth International Joint Conference on Natural Language Processing，2013：311-319.

[25] 谭春辉，陈晓琪，梁远亮，等. 隐私泄露事件中社交媒体围观者情感分析[J]. 情报科学，2023，41（3）：8-18.

[26] 刘雅姝. 多维视角的重大突发事件演变机理及应对策略研究[D]. 长春：吉林大学，2021.

[27] 丁效. 句子级中文事件抽取关键技术研究[D]. 哈尔滨：哈尔滨工业大学，2011.

[28] 刘政昊. 基于知识关联的多层本体立方体设计与实现：以金融证券领域为例[J]. 现代情报，2022，42（1）：72-86.

[29] 万齐智，万常选，胡蓉，等. 基于句法语义依存分析的中文金融事件抽取[J]. 计算机学报，2021，44（3）：508-530.

[30] 赖佳敏. 基于事理图谱的意图识别方法研究[D]. 上海：华东师范大学，2022.

作者简介

赵又霖，女，博士，副教授，研究方向：数据分析与挖掘、知识组织研究，E-mail：sobzyl@hhu.edu.cn。
林怡妮，女，硕士研究生，研究方向：数据分析与挖掘、知识组织研究。
石燕青，女，博士，副教授，研究方向：复杂网络、社交媒体数据挖掘。

Construction of Chinese Open Domain Event Graph Integrating Sentiment Semantics and Syntactic Structure

ZHAO YouLin[1,2] LIN YiNi[1] SHI YanQing[3]
(1. Business School of Hohai University, Nanjing 211100, P. R. China; 2. School of Information Management, Nanjing University, Nanjing 210023, P. R. China; 3. College of Information Management, Nanjing Agricultural University, Nanjing 210095, P. R. China)

Abstract: In the construction process of large-scale open domain event graph, lack of annotation data and unknown event types cause difficulties in the transfer of limited domain event graph construction method. To solve this problem, we utilize rule matching methods to efficiently identify multiple event logical relationships contained in open domain texts, and integrate sentiment semantics and syntactic structure information analysis to improve the accuracy of event extraction, in order to better complete the task of constructing event graphs. Firstly, we summarize and expand various logical relationship extraction templates such as cause and effect, succession, condition, transition, etc., and screen logical relationship event sentences based on rule templates and dependency parsing information. Secondly, we innovatively introduce the sentiment semantic analysis method to accurately identify event types by capturing the sentiment semantics of events and inter-event relations on the basis of syntactic structural information, and then extract event arguments. Then, the semantic similarity is computed for event fusion, and the ＜preceding event, event logical relation, subsequent event＞ ternary is constructed to get the event graph, and further event generalization is performed to construct the abstract event graph. Finally, taking the “2022 Mpox Incident” event data as the data source, empirical analysis proves that the open domain event graph construction method can realize the identification of different types of events and reveal the logical relationships between events. Its effectiveness and feasibility are verified. The construction of the Chinese open domain event graph not only fills the gaps in the existing theories of event graph construction, but also provides powerful data support for decision-making and event development prediction.

Keywords: Open Domain; Event Graph; Dependency Parsing; Semantic Dependency Parsing; Sentiment Analysis

（责任编辑：王玮）