

产业集群知识图谱构建方法研究

范存庆 余军合 战洪飞 王瑞¹

(宁波大学 机械工程与力学学院, 浙江 宁波 315211)

【摘要】: 通过探索知识图谱在产业集群的应用, 提出了产业集群知识图谱整体设计框架, 设计了产业集群知识图谱中实体及其关系本体模型。选取宁波市注塑机产业集群为例, 通过互联网平台获取企业数据, 经过数据预处理、知识抽取和知识融合, 依据本体模型构建了宁波市注塑机产业集群知识图谱, 将所构建的所有三元组存储于 Neo4j 图数据库中。根据所构建的知识图谱, 分析注塑机集群整体特性, 通过分析发现产业集群在地理位置临近关系上和业务往来有着密切的关系, 验证本文提出方法的可行性。

【关键词】: 知识图谱 产业集群 知识抽取 Neo4j

产业集群是推动区域经济发展的重要形式, 产业集群中包含企业、人员、机构等众多主体, 各主体之间都有直接或间接的复杂关系, 构成庞大的网络结构。随着大数据技术的发展, 通过获取产业集群相关数据, 构建产业集群网络结构, 从微观的角度研究产业集群的发展规律, 更能详细研究产业集群的内部演化规律。知识图谱是研究网络结构数据的有效工具, 其本质是一个能对现实世界中事物之间的关系进行直观映射的语义网络, 可以利用知识图谱理论方法来研究产业集群的内部结构。知识图谱作为大数据技术的重要组成部分, 现在已经被广泛应用^[1]。

车金立等构建了军事装备知识图谱, 用于实现军事装备领域的知识问答^[2]; 在煤矿安全领域, 刘鹏等将知识图谱结合 LatticeLSTM 模型和语义相似度计算, 提出了一种自然语言知识查询方法, 从而提高了煤矿安全信息资源整合^[3]; 杜志强等围绕自然灾害事件、灾害应急任务等 4 个要素, 结合本体建模方法和条件随机场模型构建了洪涝灾害应急知识图谱, 根据数据关联, 实现对相关数据节点的推荐应用, 为提高自然灾害应急响应研究提供了理论方法基础^[4]; Rotmensch 等提出了一种使用基本概念从大规模电子病历中提取医学信息并自动构建高质量健康知识图谱的方法^[5]; Fang 等将计算机视觉算法和本体模型相结合, 开发出建筑安全知识图谱, 可依照安全法规自动识别建筑工地的安全隐患^[6]; Xiao 等采用 BiLSTM+CRF 模型从大量文献中提取气象模拟知识并结合 Neo4j 图数据库构建气象模拟知识图谱, 实现气象模拟知识的结构化存储和集成^[7]。

本文选取宁波地区的注塑机产业集群作为研究对象, 提出构建产业集群知识图谱的整体框架, 构建产业集群本体模型, 并从互联网平台获取数据, 构建注塑机产业集群知识图谱, 探索产业集群知识图谱的应用。

1 产业集群知识图谱整体框架设计

产业集群相关数据包括企业基本信息、专利信息、产品信息和人员需求信息等, 这些数据包括结构化数据和非结构化数据, 良莠不齐且数据量大, 并非所有数据都可以用于构建知识图谱。因此构建产业集群知识图谱需要先明确产业集群的本体模型, 根

¹**作者简介:** 范存庆, 宁波大学机械工程与力学学院硕士研究生, 研究方向: 数据分析及制造系统工程; 余军合, 宁波大学机械工程与力学学院副教授, 研究方向: 智能制造和知识管理; 战洪飞, 宁波大学机械工程与力学学院教授, 研究方向: 智能制造和知识管理; 王瑞, 宁波大学机械工程与力学学院讲师, 研究方向: 智能制造和知识管理。

基金项目: 国家自然科学基金项目——“面向产业集群广义制造系统的知识模块化机理研究”(项目编号: 71671097; 项目负责人: 战洪飞) 成果之一; 国家重点研发项目——“注塑机和模具产业网络协同制造发展模式及关键技术研究”(项目编号: 2019YFB1707101; 项目负责人: 战洪飞) 成果之一

据建模结果从海量的数据中抽取相应的实体和关系，构建产业集群知识图谱。本文提出的整体框架如图 1 所示。

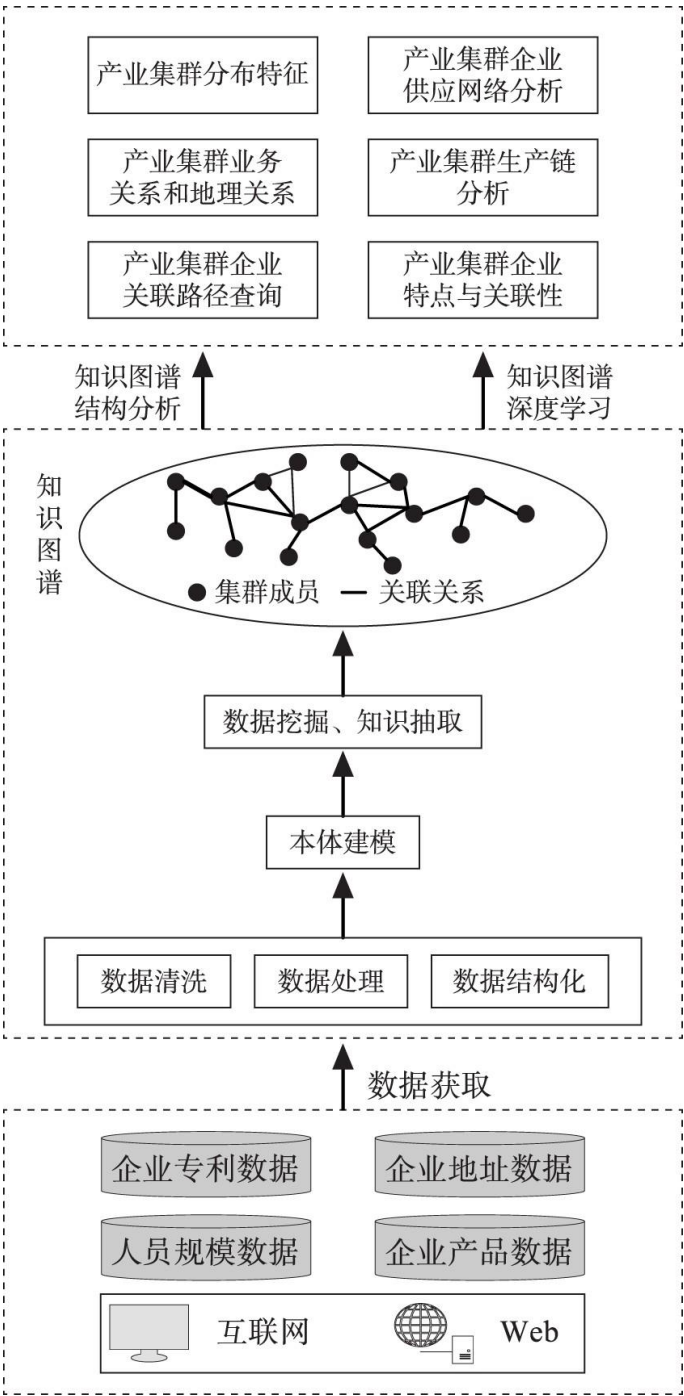


图 1 产业集群知识图谱整体框架图

数据获取与预处理：数据的来源主要有天眼查之类的企业信息网站、企业专利库及招聘网站等，通过数据挖掘等手段进行获取，并对获取数据进行预处理，为信息抽取做准备。

本体建模和知识抽取：明确集群中企业、人员、产品等实体的属性以及各个实体之间的关系，构建产业集群的本体模型；根

据本体模型从已有数据中抽取相关的企业实体、产品实体，同时抽取实体关系，构建知识图谱。

知识存储和可视化：在将数据转化成结构化数据基础上，抽取实体及其属性以及关系，构造“实体-关系-实体”三元组。传统的关系型数据库难以直观描述这种三元组关系，因此本文选用图数据库 Neo4j 存储知识图谱三元组。Neo4j 可以直观反应实体之间的关系，并且利用 Cypher 图数据库查询语句可以对知识图谱进行高效的增删改查。

知识图谱应用：产业集群知识图谱从数据层面描述了产业集群的物理关系。通过知识图谱图结构分析方法和深度学习方法，对产业集群相关问题进行研究。

2 产业集群知识图谱构建

研究选择自顶向下的构建方式，先构建产业集群本体模型，再根据模型从数据中抽取相关实体和关系。

2.1 产业集群知识图谱本体建模

产业集群知识图谱建模侧重于构建知识图谱的本体模型，明确产业集群知识图谱中出现的实体概念和关系，实体概念可以理解实体的类，关系是指对象之间的二元关系。其中实体概念一共包括 6 个：企业、人员、产品、地区、知识资源和科研院校。

企业类实体概念。企业是产业集群的主体，定义企业实体概念为 {label;name;size;address;industry;time}，中括号中是该实体的属性。其中 label 代表此类实体类别，比如“企业”；name 代表企业名称，比如“富”机械制造有限公司”；size 代表企业规模，选取企业参保人数作为企业规模的衡量标准；address 代表企业所属地区，比如“北仑区”；industry 代表企业所属行业，比如“通用设备制造业”；time 代表企业成立时间，单位是年，如“2013”。

产品类实体概念。产品是产业集群经营活动的体现，也是产业集群的主体之一，定义产品实体概念为 {label;name;class;frequency;function;price}。其中 label 代表此类实体的类别，如“产品”；name 代表产品名称，如“注塑机”；class 代表产品类别，如“生产设备”；frequency 代表产品在集群中出现的频次；function 代表产品功能，如“加工塑料制品”；price 代表产品价格。

人员类实体概念。人员是产业集群中知识资源的载体，将人员实体概念定义为 {label;age;sex;education;skill;post}。其中 label 代表此类实体的类别；age 代表人员年龄；sex 代表人员性别；education 代表人员学历；skill 代表人员技能；post 代表人员岗位。人员类概念分为法人、研究人员及职工。

其他实体概念。地区类实体概念主要描述空间信息，用于关联企业业务活动。知识资源类实体概念主要描述产业集群的技术层面信息，包括专利、论文和技能等。科研院校类实体概念是集群中区别于企业的另一类机构实体，是论文等知识资源的主要输出单元。

在对产业集群知识图谱中的实体概念进行建模基础上，需要对实体间关系进行建模。关系包括实体间的二元关系以及实体和属性之间的关系。由于产业集群数据中有大部分结构化数据，其中的关联关系比较明确，方便对关系进行建模。结合前述步骤所建模的实体类概念，使用 protégé 工具构建出完整的产业集群知识图谱模式层及其实体类概念、实体关系和实体属性如图 2、图 3 所示。

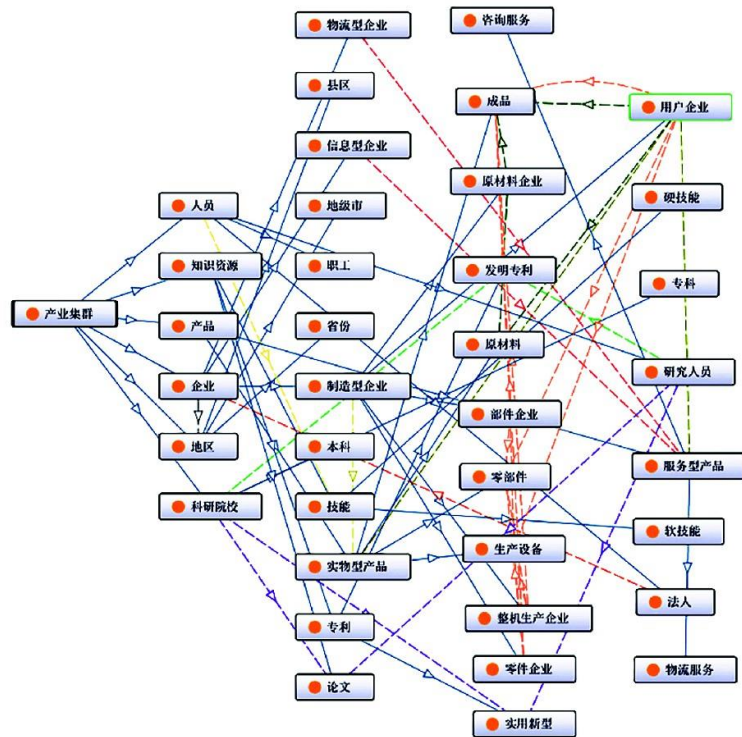


图 2 产业集群知识图谱模式层概念关系

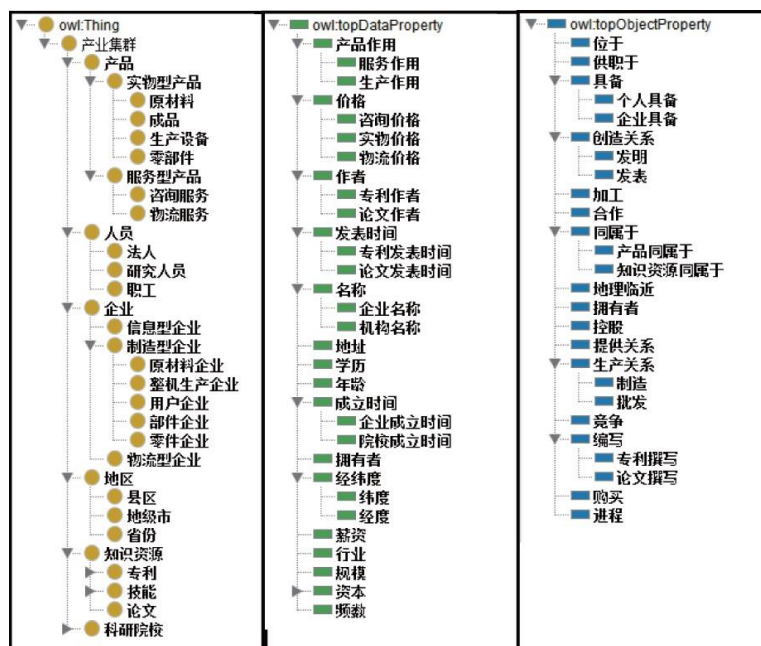


图 3 产业集群知识图谱模式层概念及其关系

2.2 实体抽取和关系抽取

在本体建模基础上，可以从已有的数据中抽取相关的实体和关系，组成三元组，构建知识图谱。

企业实体及其关系抽取：企业基本信息主要存储在结构化数据中，由于结构化数据质量较高，具有规范的模式，对于结构化数据采用直接映射和基于规则的抽取方式。根据知识图谱建模部分所构建的企业本体模型，从数据库中提取“企业名称”字段作为 name 属性，“参保人数”字段作为 size 属性，“所属区县”字段作为 address 属性，“所属行业”字段作为 industry 属性，“成立日期”字段作为 time 属性，一共抽取到 1958 个企业实体，部分“企业”实体数据如表 1 所示。

表 1 企业实体数据(部分)

Id	Name	Size	Address	Industry	Time
131	先*机械有限公司	1	江北区	通用设备制造业	2013
132	速*包装制品有限公司	1	江北区	造纸和纸制品业	2013
134	海*工贸有限公司	70	奉化区	电气机械和器材制造业	2012
135	富**机械制造有限公司	305	江北区	通用设备制造业	2001
136	创*机械有限公司	268	余姚市	通用设备制造业	2004

企业-企业关系的地理临近由企业地址计算得出，经过数据预处理阶段，每一个企业的地址都是唯一的，可以在地图上准确定位。结合百度地图开发平台 API, 将企业地址转换为经纬度进而计算出企业之间的相对距离，计算公式为：

$$\begin{aligned} \text{haversin}\left(\frac{d}{R}\right) &= \text{haverln}\left(\varphi_2 - \varphi_1\right) + \cos\left(\varphi_1\right) \cos \\ &\left(\varphi_2\right) \text{haversin}\left(\Delta \lambda\right) \end{aligned} \tag{1}$$

其中， $\text{haversin}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = (1 - \cos(\theta))/2$, R 为地球半径,可取平均值 6371km; φ_1, φ_2 分别为两点的纬度;
 $\Delta \lambda$ 是两点经度的差值。

计算出所有企业两两之间的距离之后，将距离小于 500m 的两个企业设定为具有“地理临近”关系，并且将实际距离设置为该关系的权重。表 2 为部分企业关系数据。

表 2 企业间关系数据(部分)

CompanyA	CompanyB	Relation	Weights
富**机械制造有限公司	猛*消防设备有限公司	地理临近	0.51723
宜*物流有限公司	隆*电子有限公司	地理临近	0.49436
意*机械有限公司	兴*机电机械厂	地理临近	0.22345
亚*塑料有限公司	北*物流有限公司	地理临近	0.34006

产品实体抽取：产品信息主要存储在非结构化文本中，然而由于这部分数据没有规范模式，因此使用 BERT+BiLSTM+CRF 模型进行实体识别。该模型有 3 个模块组成，第一个模块是 BERT 模块，负责将输入的文本转换成字向量输入到下一模块；第二层是 BiLSTM 模块，负责提取上下文语义特征并将结果输入第三模块；第三模块是 CRF，负责对第二模块输出结果进行解码，对输入文本进行序列标注。

将文本输入该模型之前，需要先对文本进行标注，目前，常见的中文标注体系有三种：BMES、BIO 和 BIOES，本文采用的是 BIO 标注体系。标注对象是文本中的产品名称，B-PRO 代表命名实体开始的字符，I-PRO 代表命名实体剩余的字符，O 代表非命名实体的字符。

选取了 1000 家企业的经营范围文本数据进行标注，训练该模型。标注方法采用自动化标注和人工标注相结合的方式，最终一共标注了 3698 个句子，将已标注的数据集按照 7：3 的比例划分为训练集和测试集。实验评价指标采用准确率 P，召回率 R 和 F1 值进行实体识别的效果评价。为了对比该模型的效果，选取 BiLSTM+CRF 模型和 Word2Vec+BiLSTM+CRF 模型进行对比实验，结果如表 3 所示。

表 3 实验结果对比

Model	P (%)	R (%)	F1 (%)
BiLSTM+CRF	82.83	84.28	83.55
Word2Vec+BiLSTM+CRF	85.85	86.69	86.27
BERT+BiLSTM+CRF	91.28	94.83	93.02

结果证明 BERT+BiLSTM+CRF 在实体识别中效果较好。最终通过该模型抽取产品实体 859 个，以企业为单位将抽取到的产品名称存储在结构化数据中，并统计所有产品在所有企业出现的频次，基于构建的产品实体模型创建产品节点，其中产品名称作为 name 属性，产品频次作为 frequency 属性。

企业-产品和产品-产品关系抽取：通过分词发现，企业和产品之间的关系包括“制造”“批发”“提供”3 种。因此，产品和企业的关系采用直接映射的方式构建，其中“制造”“批发”两个关键词在对企业经营范围文本进行分词时保留，然后直接与抽取到的产品进行匹配，构建“企业-制造/批发-产品”三元组。对于服务类型的产品，比如“仓储”“货物运输”等，和企业的关系为“提供”，构建“企业-提供-产品”三元组。

将两种产品出现在同一家企业的共现关系定义为产品-产品关系，关系名称为“同属于”。此外，构建产品的共现矩阵，比如“注塑机”和“塑料制品”在一家企业同时出现则计数 1，如果在另一家企业又同时出现则计数加 1，将两种产品的共现频次经过归一化后的数值作为“同属于”关系的权重，权重越大，产品关联越强，反之越弱。

2.3 知识融合

由于数据来源多样，存在同一个实体在不同数据源中表述不一致的问题，造成实体冗余，降低了知识图谱的质量。为这一问题，需要进行实体对齐。针对不同实体使用基于匹配规则和基于相似度两种方式进行实体对齐。对于企业实体，使用基于匹配规则的知识融合，即建立企业名称库，将抽取的实体统一匹配到名称库中，将表述统一。

对于产品实体，由于产品数量多，而且同一产品的表述远不止两种，因此提出基于语义相似度的知识融合。使用 BERT 输出的产品词向量，计算两个产品向量的余弦相似度，计算公式为：

$$\cos(x, y) = \frac{x \times y}{\|x\| \times \|y\|} \quad (2)$$

其中 x, y 为两个词的词向量，计算出的余弦相似度越接近 1，两个词越相似，越接近 0，两个词越不相似。本文设置语义相似度的阈值为 0.7，即余弦相似度大于等于 0.7 的词判定为相似词，将两者对应的实体进行融合。

经过知识融合，最终构建的图谱一共有实体 2596 个，包括“企业”和“产品”；关系 38965 个，包括“地理临近”“制造”“批发”“提供”“同属于”。将所有三元组全部存入 Neo4j 数据库中，其部分结果可视化结果如图 4 所示。

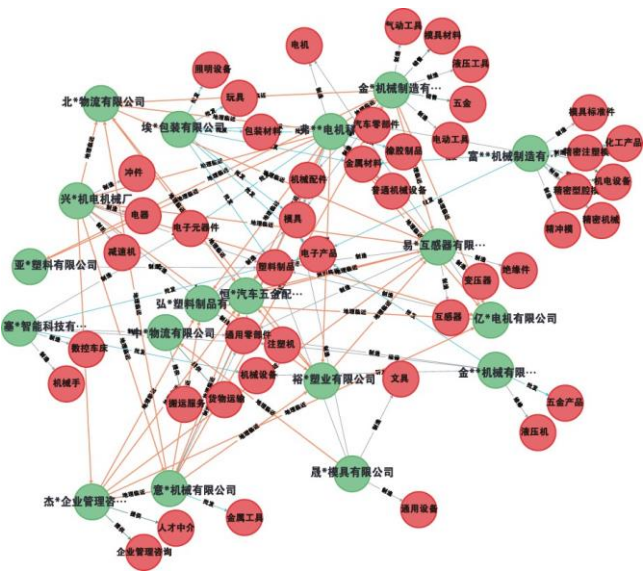


图 4 产业集群知识图谱(局部)

3 产业集群知识图谱分析

企业之间的业务配套往往和地理位置的临近紧密关联，即在地理位置上临近的企业之间有更多的业务往来，反之，有较多业务往来的企业地理位置较临近。企业之间的业务往来可以从产品体现。

注塑机生产企业涉及的行业较多，上游企业包括钢材、冶金铸造等机械类零部件生产加工行业，以及液压零件、电子仪器仪表、电子元器件等传动类和控制类零部件加工行业。下游企业则包括塑料建材、汽车配件、家用电器和物流、包装材料以及其他普通塑料制品行业。

使用 Cypher 图数据库查询语言可以精确查询节点和关系信息，如图 5 所示。选取“富**机械制造公司”为例，查询其 2 跳关系的节点及关系，可以发现作为主机厂，“富**机械制造公司”周边分布着物流公司，下游业务公司比如塑料制品企业，以及上游公司比如机械零件制造企业，整个集群依托地理位置临近形成，可以看出它们的相对位置充分利用了地理环境优势，方便开展企业业务，各个企业之间的业务相互配套，形成制造系统。

computer vision with ontology[J].Automation in Construction,2020,119:103-109.

[7]XIAO Z,ZHANG C.Construction of Meteorological Simulation Knowledge Graph Based on Deep Learning Method[J].Sustainability,2021,13(3):131-140.