Computer Science and Application 计算机科学与应用, 2017, 7(9), 858-868
Published Online September 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.79098

Hans汉斯

Packaging Industry Information Inquiry Technology Architecture Based on Knowledge Graph

Wenqiu Zhu, Yuan Si

School of Computer Science, Hunan University of Technology, Zhuzhou Hunan
Email: wenqiu_zhu@126.com

Received: Aug. 30th, 2017; accepted: Sep. 10th, 2017; published: Sep. 18th, 2017

Abstract

The packaging industry information inquiry technology architecture based on knowledge graph is proposed, and the knowledge hierarchy construction, the knowledge extraction, the knowledge fusion, and the knowledge application are described. In this paper, we introduce a new method of classifying questions with the help of domain-specific ontology and obtain structural semantic information for the question. Given a seed pattern, relevant pattern can be learned automatically from large-scale training corpus. The packaging industry database search method based on the knowledge graph for handling query natural language query is proposed, and the constructing procedures of the packaging industry information inquiry system based on knowledge graph are provided.

Keywords

Packaging Industry Information, Knowledge Graph, Ontology Knowledge Base, Information Inquiry

基于知识图谱的包装产业信息查询技术架构

朱文球，司　元

湖南工业大学计算机学院，湖南　株洲
Email: wenqiu_zhu@126.com

收稿日期：2017年8月30日；录用日期：2017年9月10日；发布日期：2017年9月18日

文章引用: 朱文球, 司元. 基于知识图谱的包装产业信息查询技术架构[J]. 计算机科学与应用, 2017, 7(9): 858-868.
DOI: 10.12677/csa.2017.79098

摘　要

提出了基于知识图谱的包装产业信息查询技术架构，并对知识体系构建、知识抽取、知识融合和知识应用等核心技术进行了阐述。提出一种基于领域本体的问题分类方法和结构化语义信息提取方法，根据给定的种子模板，从大规模训练数据中可以自动学习相关的模板。以中国包装产业数据库搜索为例，提出一种能处理自然语言查询的基于知识图谱的中国包装产业数据库查询方法，给出了基于知识图谱的中国包装产业数据库查询系统的具体构建步骤。

关键词

包装产业信息，知识图谱，本体知识库，信息查询

1. 引言

随着信息化时代日新月异的发展，以及对数据的生产、存储、处理等能力的进一步提高，传统的互联网技术正由原来的计算科学转换为数据科学，在大数据、“互联网+”和“工业 4.0”迅速发展的浪潮中，谁拥有数据就拥有了开启未来大门的钥匙。我国包装产业从改革开放以来快速发展，已建成涵盖设计、生产、检测、流通、回收循环利用等产品全生命周期的较为完善的体系，分为包装材料、包装制品、包装装备三大类别和纸包装、塑料包装、金属包装、玻璃包装、竹木包装五大子行业。总产值从 1980 年的 72 亿元跃升至 2015 年的 1.8 万亿元，初步形成长三角、珠三角以及环渤海三大包装产业带。但是，包装产业“大而不强”的矛盾十分突出。我国包装企业总数达 30 万个，其中规模以上企业只有 2 万多家，90%左右为中小企业。大数据时代的到来使得包装行业不仅仅是自己这个企业的数据统计，还有国内国外整个行业以及行业上下游的行业数据，因此，这个数据要是统计下来可是一个非常庞大的数据库。包装产业决策人员面对互联网产生的包装产业海量数据，如何进行信息挖掘，并将信息优势转换为决策优势，改善企业的前途，改良行业的状态，已成为一大难题。以知识图谱(knowledge graph)为代表的知识工程技术应用是解决该难题的方法之一。相对于信息，知识能更直接地指导人的决策和行动，从而弥补信息优势向决策优势转换中的缺失，即信息优势首先转换为知识优势，然后再由知识优势转换为决策优势。

本文提出了基于知识图谱的包装产业信息搜索技术架构[1]并阐述了相关核心技术，同时提出基于领域本体[2]的问题分类方法和结构化语义信息提取方法[3] [4] [5]，在此基础上，实现了一种能处理自然语言查询的包装产业数据库搜索方法。

2. 知识图谱

知识图谱，也被称为科学知识图谱、知识域可视化或知识域映射地图，是科学知识的发展进程与结构关系的一系列各种不同的图形。产业知识图谱是聚焦在产业和金融这个垂直领域，以企业为核心，建立起相关经济要素之间的相互联系，然后用大数据对关系量化，最后用机器学习寻找要素之间的隐含影响和传导效应，最终梳理出一条完整的商业逻辑链条。包装产业知识图谱主要描述了包装企业、包装人物、包装产品等，将人、产品、企业等进行关联起来。

随着大数据时代的到来，知识图谱已经在其他领域有所应用[6]。国内已有百度“知心”、搜狗“知立方”等通用领域知识图谱[7]，已有医疗、歌曲、电影等专业领域知识图谱。Google 在 2012 年发布了“知识图谱”，利用知识图谱将 Google 的搜索结果进行知识系统化。2013 年 2 月，百度也推出了自己的知识图谱。不同于基于关键词的传统搜索引擎，知识图谱可用来更好地查询复杂的关联信息，从语义层面理解用户意图，改进搜索质量。例如在百度的搜索框里输入“马云”的时候，搜索结果页面的右侧还会出现“阿里巴巴创始人成员”等与“马云”相关的人物，如图 1(a)所示；另外，对于包含逻辑关系的搜索语句例如“马云妻子”，百度能准确返回他的妻子“张瑛”，如图 1(b)所示。这就说明搜索引擎通过知识图谱真正理解了用户的意图。

3. 中国包装产业大数据知识图谱技术体系架构

3.1. 构建流程

由于中国包装产业大数据知识图谱强调知识的深度和整体的层次结构，因此在构建时通常采用自顶向下和自底向上相结合的方式。其中，自顶向下的方式是指通过本体编辑器或手工构建的方法预先构建知识图谱的模式图，进而构建数据图。而自底向上的方式指在构建数据图时，利用多种抽取技术获得知识源中的实体、属性和关系，并将这些置信度高的抽取结果合并到知识图谱中。知识图谱 G 由模式图 G_s、数据图 G_d 及二者之间的关系 R 组成。本文在已经构建了包装产业大数据知识图谱模式图 G_s 的前提下，从数据源出发，采用自底向上的方式说明构建知识图谱数据图 G_d 和关系 R 的过程。如图 2 所示。

(a)

(b)

Figure 1. Application of knowledge graph of Baidu
图 1. 百度搜索知识图谱应用

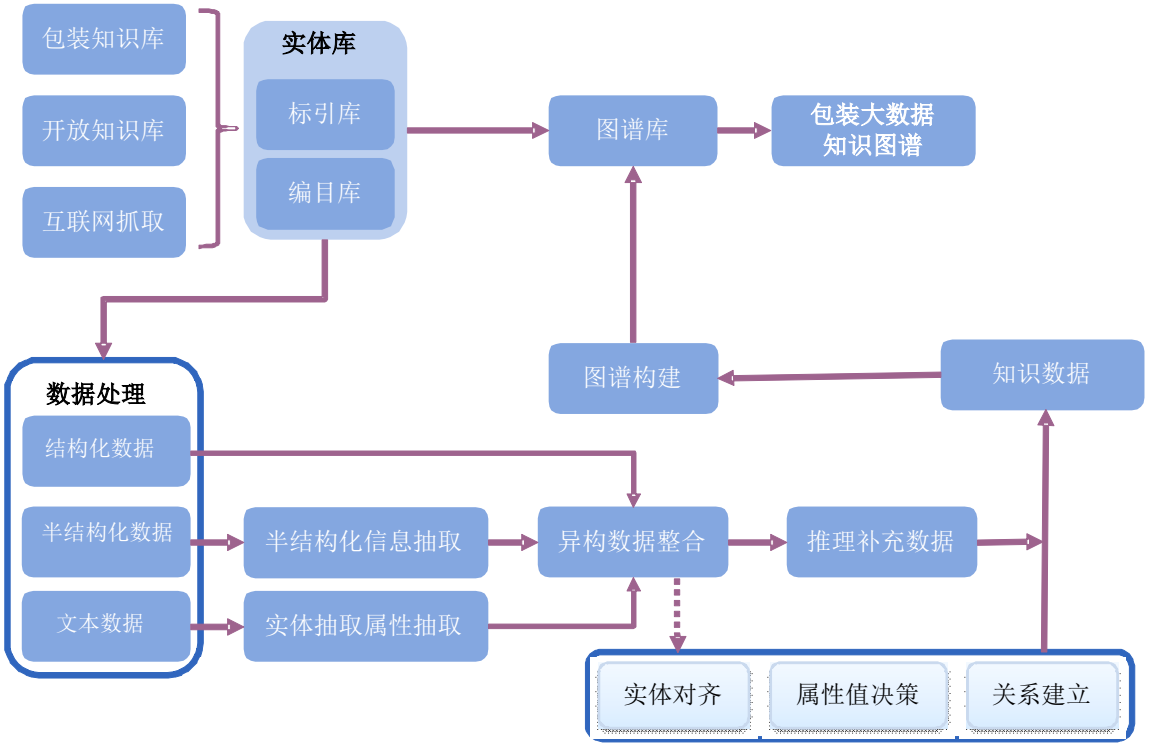


Figure 2. The construction process of China packaging knowledge graph
图 2. 中国包装大数据知识图谱构建流程

3.2. 知识存储

知识图谱中的知识存储在它的知识库中，是一个规模庞大的关联集合。对知识进行抽象和约束，是建立知识图谱的基础，主要包括本体库和知识分类。本体库指以某种方式有序组织的本体的集合，其中本体描述了概念的属性和相互关系，如包装企业应包含企业名称、企业基本信息、企业股东、企业高管、近年财报等属性。包装材料应包含材料类别、特点、作用和产品性能等属性，包装企业和包装材料间的关系为包装企业生产包装材料。知识分类描述了不同概念和实体的分类以及上下位关系，如 PE 薄膜、珍珠棉、气泡袋均属于内防护包材；按包装材料的成分可以分为含金属包材、塑胶类包材、纸质类包材、玻璃类包材、木质包材等；按用途又可分为隔热包装、抗菌包装、绝缘包装等。

3.3. 知识库构建

知识图谱技术的核心，可分为知识抽取和知识融合 2 个层次[8] [9] [10]。

3.3.1. 多策略学习的知识抽取方法

在知识体系约束和引导下，从结构化、半结构化和非结构化数据中使用人工规则和自动实体抽取算法相结合的方法自动抽取包装企业、包装产品和包装人物等包装类实体，同时可以针对包装竞争关系、包装产品从属关系等多种关系进行实体关系抽取；另外，对于包装产业大数据知识图谱中的基本实体类可以从开放数据中进行属性信息的自动补充，包括的数据类型如字符串类型、时间类型、范围类型、集合类型、对象类型等，并导入知识图谱中。例如，从某新闻报道中“……湖工大党委书记唐未兵、校长谭益民一行来我校考察交流……”抽取出“湖工大”、“唐未兵”、“谭益民”等实体，以及“湖工大的党委书记是唐未兵”、“湖工大的校长是谭益民”的实体关系。

本文提出多策略学习的方法进行知识获取。多策略学习是指利用不同知识源之间的冗余信息，使用较易抽取的信息来辅助抽取那些不易抽取的信息。结构化知识和半结构化知识由于具有显式的结构和固定的格式，属于易抽取的信息，而无结构的文本知识属于较难抽取的信息。

如图 3 所示，对于结构化知识中的关系数据库数据，通过 D2R (Relational Database to RDF)映射的方法将其转化成知识图谱中的链接数据。对于百科数据中的信息框(Infobox)、表格等半结构化知识，则使用基于封装器(HTML Wrapper)的抽取方法。对以上两种数据来源的知识进行抽取，并且将抽取的结果加入到种子集中。

对于广泛存在于中国包装联合会、各包装企业、有包装学科的大学等的非结构化知识，则采用远程监督和基于模式的方法相结合的增量迭代抽取方式。所谓远程监督是一种基于假设“如果两个实体存在某种关系，那么任何包含这对实体的句子都很有可能表达相同的关系”、利用已知的实体关系对自动标注文本的方法。这里就可以利用种子集自动标注文本数据，然后根据标注结果自动地生成高质量的模式。利用这些模式到文本中学习新的知识，并加入到种子集中。这一过程不断迭代，直至没有新的知识被学习出来。

以上三类知识经过抽取，并将抽取结果加入种子集中。

3.3.2. 知识融合

对不同来源知识进行整合和优化的过程，包括实体对齐、实体属性值判定、实体消歧以及实体关系补全等。实体对齐将描述同一实体的不同描述方式映射到同一实体，如将“湖工大”和“株洲唯一一所本科学校”映射到“湖南工业大学”；实体属性值判定是在进行实体融合时，不同来源的知识对实体的同一属性描述不一致，需确定该属性的取值，如不同来源的知识在描述“湖南工业大学”时，给出了不同的建校时间值，此时可根据知识来源的可信程度以及属性值被不同来源提及的次数进行判定；实体消歧是指同一个词汇可能代表不同实体，因此需根据上下文信息推测当前词汇究竟指向哪个实体，如企业高管名字可能出现同名，此时可以根据上下文描述的企业名称、企业所在地区等信息推测该名字具体指哪位企业高管；实体关系补全利用多个来源的知识推测实体间关系，如在学术会议相关知识中发现某 2 位教授参加的学术会议存在大量交集，则可推测这 2 位教授相互认识。本文的系统所采用的知识融合方法如图 4 所示。

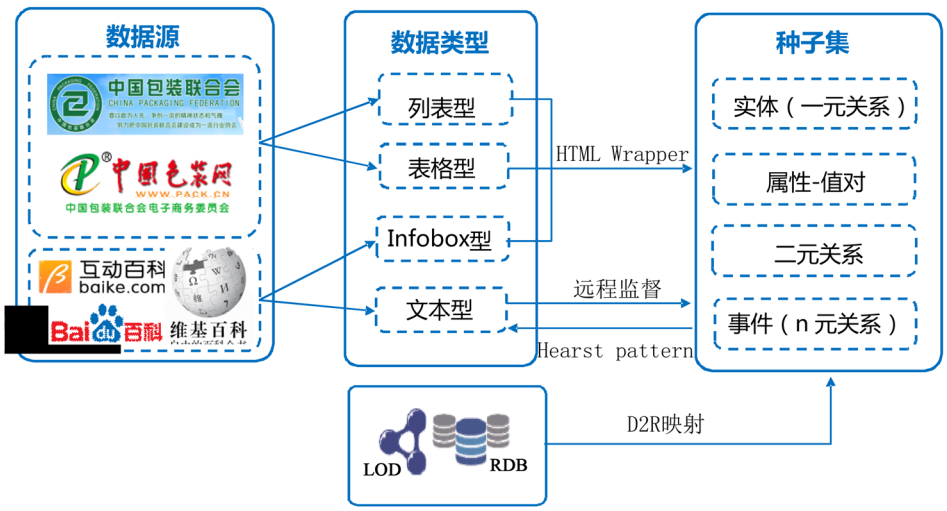


Figure 3. Multi strategy learning method
图 3. 多策略学习方法

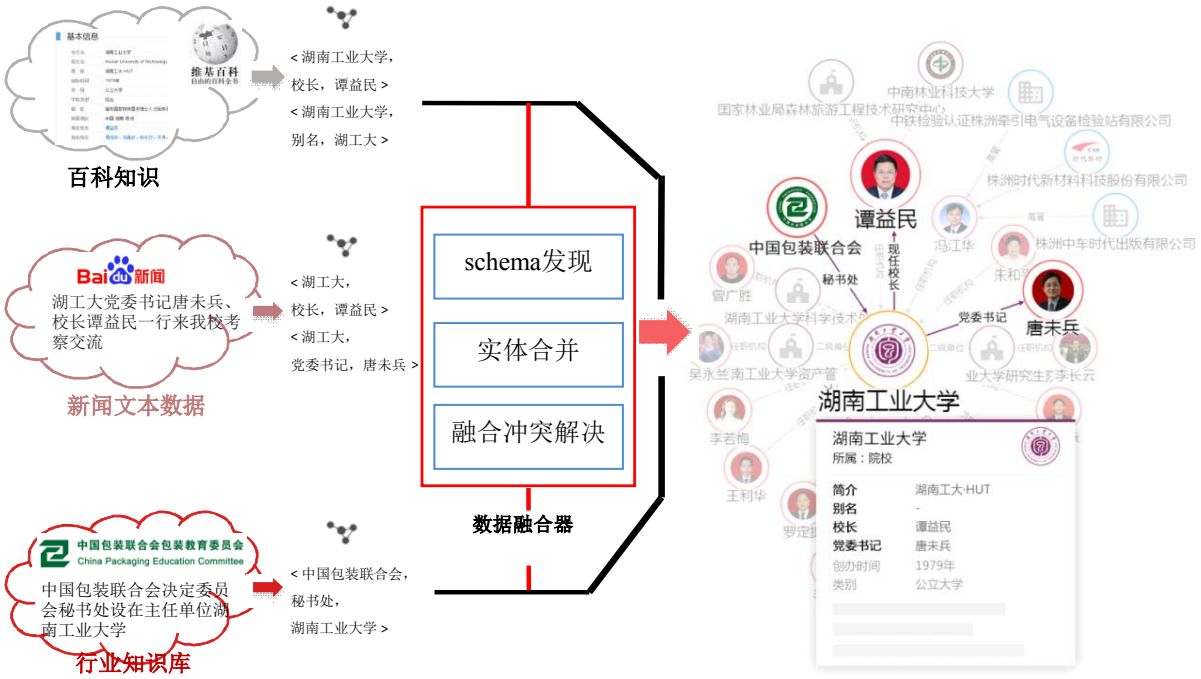


Figure 4. Knowledge fusion method
图 4. 知识融合方法

4. 基于模板匹配的结构化查询

知识图谱构建后，通过知识图谱存储和查询引擎，可使用简洁灵活的查询语句对知识进行高效查询。但是需要用户将信息需求写成特定格式的查询语句，以便获得相应知识，故要求用户具有专业知识并经过训练。由于 RDF 数据的复杂性以及 SPARQL、Cypher 查询语言的复杂性，而自然语言查询语句是一种无结构化的查询语句，为此本文提出一种基于规则模板匹配的查询语句自动转换算法，从而支持用户利用自然语言进行查询。

提出的方法主要包括自然语句分词、问题分类、模板匹配、查询生成等步骤[1]。自然语言查询语句的转换过程如图 5 所示。

4.1. 对输入的语句进行分词处理

自然语句分词采用中国科学院计算技术研究所的汉语词法分析系统 ICTCLAS。在处理特定领域文本时，因为大量的专业术语没有被收录到系统库中，造成分词的结果不准确，甚至是错误。本文的系统分词时底层依赖于两类词库：非专业领域词库和包装知识库。这样系统进行分词时就会优先按照用户自定义词典中的词条进行分词，从而将领域专业术语正确的识别出来。

非专业词库收录了日常用语、流行词汇等来源于生活和网络的词汇，主要用于对句子中常用人称、动词、名词等的识别。包装知识库是从专家撰写的知识体系文档、互联网(这里主要是指中国包装网)以及部分包装资料库数据文档中抽取得到的，使用团队之前已经翻译的部分 RDF 文档作为实验数据来源，抽取文档中包装工程知识的中文名称，构建用于实验的包装知识词库。

4.2. 要素检测

要素检测的一个重要任务就是将问题中的关键词映射为本体中对应的元素，这样才能从语义层次上

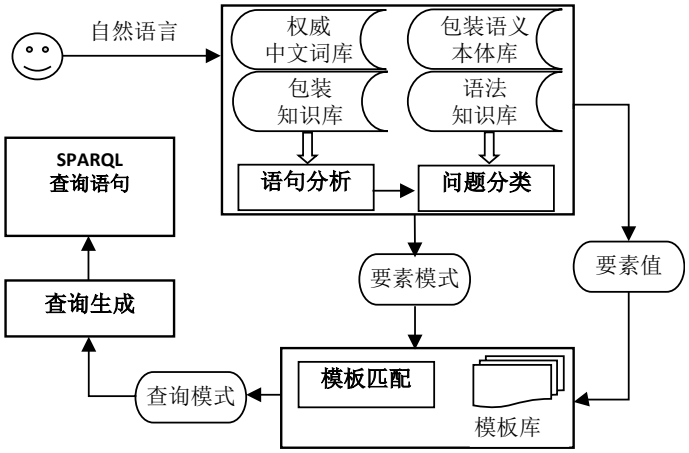


Figure 5. The flow of natural language change into query language
图 5. 自然语言转换为查询语言的处理流程

对问题进行理解。为了方便关键词向本体元素的映射，我们结合本体知识表达的方式，定义了几个本体要素，它们分别是 C、R、V、E。其中 C 代表本体中的类，例如，“大学”就是一个类。R 代表对象属性，它将本体中的两个类或者实例关联起来，例如，“开设”就是一个 R，<湖南工业大学，开设，包装印刷专业>就是它的一个例子。V 代表数据属性，它用来描述类的内在属性，它将一个实例和字符串或数值关联起来，例如，“哪些”就是一个 V。E 代表本体中的具体实例，例如，“玻璃瓶”、“玻璃罐”、“玻璃盒”就是玻璃包装材料的实例。

将知识图谱中概念、实体和关系的名称按照上述要素分别建立词库。当用户输入自然语言查询时，结合上述词库使用自然语言分词算法，将自然语言查询分割成要素和要素值。例如，“哪些大学开设包装工程专业”的分词结果为：<哪些，V> <大学，C> <开设，R> <包装工程专业，E>。

4.3. 模式匹配

根据上两个步骤得到的分词结果以及自然语言的要素模式，通过查询构建的模板库，可实现要素模式到查询模式的映射。通过分析包装领域用户查询的问题，可将用户的查询需求拆解为以下一些基本类型及由这些基本问题类型组合而成的复杂问题类型。

1) 查询某一类事物，例如，“什么是木包装”；

2) 查询某一类事物的分类，例如，“木箱包装有哪些种类”；

3) 查询某一类事物的某一方面属性，例如，“什么是瓦楞纸的克重”；

4) 查询某一类事物中实例的最值，例如，“国内最大的包装机械制造公司是哪个”；

5) 查询两个实例是否存在特定关系，例如，“湖南工业大学在株洲吗”；

6) 查询某一实例一其他实例之间的关系，例如，“谁是湖南工业大学的校长”、“湖南工业大学校长是谁”；

7) 比较两个实例的某一方面的属性，例如，“瓦楞纸和箱板纸的克重谁的高”；

8) 查询某一实例与某一类事物实例的关系，例如，“湖南工业大学的包装学科带头人有哪些”；

9) 查询属性值为某一给定值的一类事物，例如，“哪些大学开设了包装工程专业”、“国内哪些高校有包装工程专业”；

10) “国内包装工程专业生源数在 200 以上的大学有哪些”，则是(8)和(9)类型问题组成的；

假设用户查询问题的语义可以用一个或者多个三元组近似表示，三元组表示形式为<主语，谓语，宾

语>。如，“湖南工业大学开设了哪些包装类专业”的三元组表示为：<湖南工业大学，开设，包装类专业>。用本体要素 C 和 E 表示三元组的主语和宾语，R 和 V 表示谓语，也是问题中的已知信息；用“?”或者“? + 本体要素”表示待查询对象；用符号 m 表示数值，Max 表示最值类型问题，Comp 表示比较类型问题。表 1 给出了包装知识图应用中一些基本查询问题的要素模式到查询模式的映射模板库。

4.4. 查询生成

将 3.2 节中得到的要素值填入 3.3 节得到的三元组查询模式，生成知识图谱查询语句。例如，“哪些大学开设了包装工程专业”，用上面提出的方法，得到的查询为<?大学，开设，包装工程专业>。

4.5. 构造 SPARQL 查询

对于受控短语的转换，需要定义相应的模板。如“哪些大学开设了包装工程专业”，确定模板为“select A where B”类型。这里需要确定是 A 和 B 的内容，将对应的语句内容替换出查询语句中的 A 和 B 就能够得到相应的 SPARQL 查询。

经过处理后的语句，将其转换为 SPARQL 查询语句的关键点是将语义类型节点替换成拥有该类型的变量。语句中“大学”首先被映射为要查询的语义类型节点，将其选择出来作为变量“?x”替换语句中的 A。根据之前步骤语义映射及分析的结果，提取语句中的三元模式集为(?x type 大学.?x 开设包装工程专业)。将三元模式作为查询语句中 B 的生成源，将(?x type 大学.?x 开设 包装工程专业)按照规范替换 B，生成完整的 SPARQL 查询语句为：

“select ?x where (?x type 大学.?x 开设包装工程专业)”。

上述基于规则模板匹配的方法适用于简单的自然语言查询，如果需处理语义更复杂的查询，应使用基于组合范畴语法，以及基于机器学习等其他算法。

5. 包装大数据知识图谱构建及应用

5.1. 图谱构建

构建知识图谱首先从中国包装联合会、中国包装网、维基百科、百度百科等资源中提取所需内容。利用爬虫技术从包装互联网空间中抓取的文本包含 HTML 标签等杂讯，需要进行数据清洗。数据准备

Table 1. Mapping template library from element schema to query schema
表 1. 要素模式到查询模式的映射模板库

问题要素模式	三元组查询模式
<C>	<?>
<C> <V>	<C, V, ?V> Max <C, V, ?V>
<E> <R> <E>	?<E1, R, E2>
<E> <R>	<E, R, ?>
<R> <E>	<?, R, E>
<E> <E> <V>	Comp (<E1, V, ?V>, <E2, V, ?V>)
<E> <R> <C>	<E, R, ?C>
<V> <E> <R> <C>	<?C, R, E>
<E> <R> <V> <m> <C>	<E, R, ?C>, <?C, V, m>

完成之后，我们通过统计机器学习算法提取文本中的命名实体，继而通过特殊的正则模式匹配找出实体之间的关系，并将其持久化为 csv 文件。最后，我们将所有命名实体及实体关系导入 Neo4j 图数据库。构建的知识图谱如图 6 所示，图谱中的圆节点为为实体。为了显示效果，实体各属性未在图中绘制。

5.2. 中国包装知识图谱应用实例

建立在知识库上，在知识库存储查询引擎、自然人机交互以及语义分析等技术支撑下，实现了面向包装产业任务的信息搜索服务。如，对于包装企业，根据行业问答“国内排名前 10 的包装企业有哪些”，智能解析问题，在知识图谱中寻找答案，并可视化给出答案：胜达集团、上海界龙集团有限公司、上峰SALFO……。对于技术人员，根据科研问答“国内哪些高校有包装工程专业”，智能解析问题，在知识图谱中寻找答案，并可视化给出答案：湖南工业大学、西安理工大学、武汉大学、北京印刷学院……。又例如，对知识图谱任意两个(或多个)实体节点，发现它们之间可达的路径，如图 7 所示，左图表示湖南工业大学、杨军、冯江华之间的关系，右图表示湖南工业大学与株洲时代机电设备有限责任公司的关系。

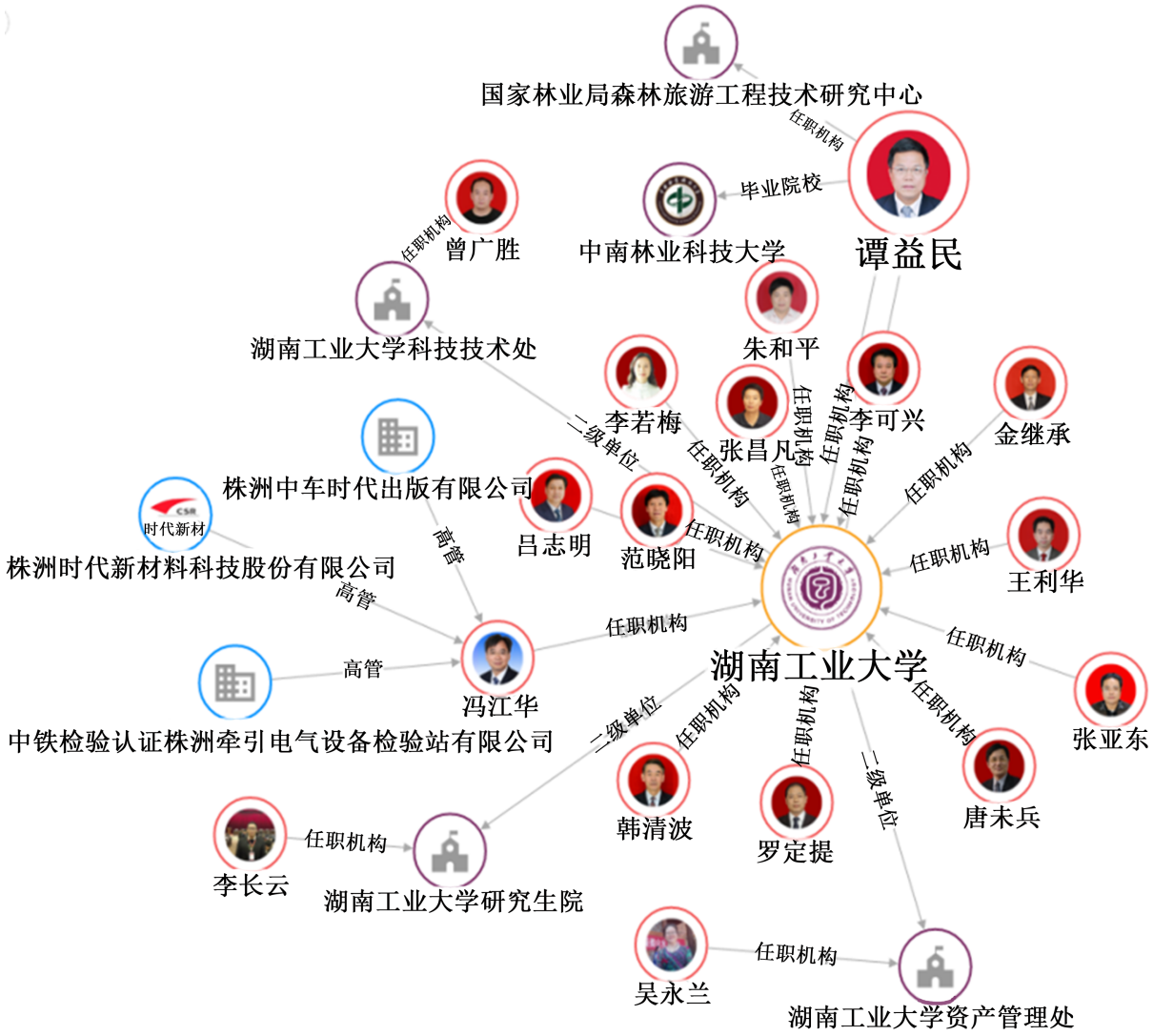
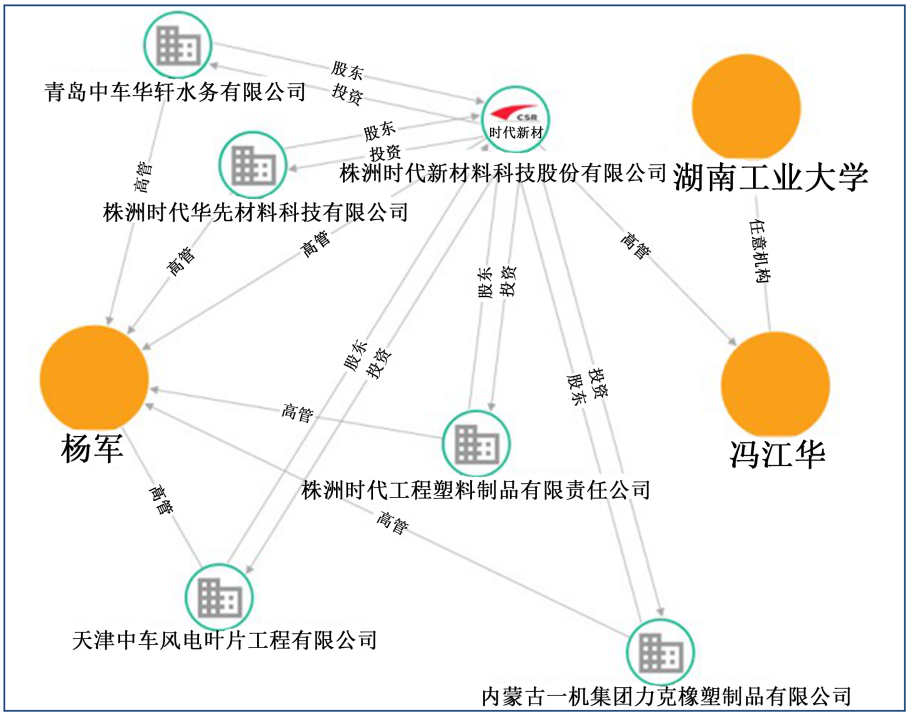


Figure 6. A knowledge graph based on sample database
图 6. 样例数据库构建的知识图谱



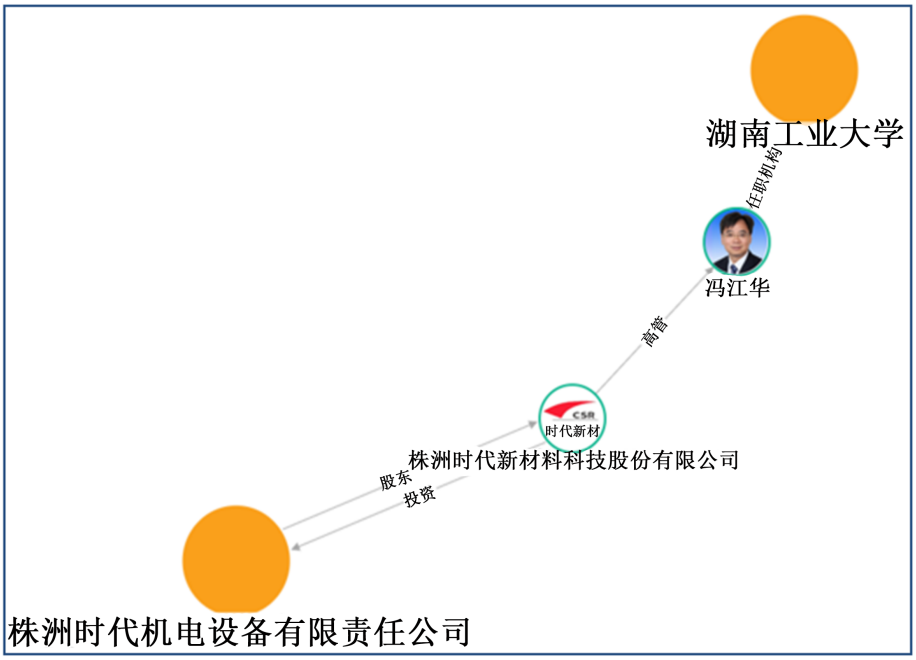


Figure 7. Sketch map of path and correlation analysis
图 7. 路径、关联分析示意图

6. 结束语

本文详细阐述了中国包装大数据知识图谱的技术架构，提出了基于知识图谱的包装产业信息查询技术架构，并对知识体系构建、知识抽取、知识融合和知识应用等核心技术进行了阐述。提出了基于知识图谱的包装产业信息搜索技术架构，并结合包装产业数据库搜索给出了具体步骤。采用面向特定领域本

体的问题分类方法，将包装产业知识图谱上的查询归结为 10 种基本类型问题及由基本类型问题组合而成的复合问题。采用结构化语义分析方法，提出一种能处理自然语言查询的基于知识图谱的中国包装产业数据库查询方法。最后，给出了基于知识图谱的中国包装产业数据库查询系统的具体构建步骤和应用实例。

基于知识图谱的包装产业信息搜索具有广泛的应用场景，能支撑中国包装产业大数据的行业数据情报服务、招聘求职服务、市场咨询服务、创新与众包服务、智慧包装服务、包装知识服务等多种应用，后续将在异构数据知识图谱构建和用户搜索意图理解等方面展开进一步研究。

基金项目

湖南省科技厅重点研发计划项目“工业装备智能监测与健康管理平台研发及应用(2016GK2017)”、中国包装联合会项目“中国包装产业大数据知识图谱的构建与应用(17ZBLWT001KT006)”。

参考文献 (References)

[1] 蒋锴, 钱夔, 郑玄, 等. 基于知识图谱的军事信息搜索技术架构[J]. 指挥信息系统与技术, 2016, 7(1): 47-52.

[2] Haveliwala, T.H. (2003) Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. *IEEE Transactions on Knowledge & Data Engineering*, **15**, 784-796. https://doi.org/10.1109/TKDE.2003.1208999

[3] 贾真, 杨宇飞, 何大可, 等. 面向中文络百科的属性和属性值抽取[J]. 北京大学学报: 自然科学版, 2014, 50(1): 41-47.

[4] 怀宝兴, 宝腾飞, 祝恒书, 等. 一种基于概率主题模型的命名实体链接方法[J]. 软件学报, 2014, 25(9): 2076-2087.

[5] 金贵阳, 吕福在, 项占琴. 基于知识图谱和语义网技术的企业信息集成方法[J]. 东南大学学报(自然科学版), 2014, 44(2): 250-255.

[6] Xin, R.S., Gonzalez, J.E., Franklin, M.J., *et al.* (2013) GraphX: A Resilient Distributed Graph System on Spark. *International Workshop on Graph Data Management Experiences and Systems*, New York, 22-27 June 2013, 1-6.

[7] 陆晓华, 张宇, 钱进. 基于图数据库的电影知识图谱应用研究[J]. 现代计算机, 2016(7): 76-83.

[8] 袁旭萍. 基于深度学习的商业领域知识图谱构建[D]: [硕士学位论文]. 上海: 华东师范大学, 2015.

[9] 王仁武, 袁毅, 袁旭萍. 基于深度学习与图数据库构建中文商业知识图谱的探索研究[J]. 图书与情报, 2016(1): 110-117.

[10] Zheng, W.G., Zou, L., *et al.* (2015) How to Build Templates for RDF Question/Answering: An Uncertain Graph Similarity Join Approach. *SIGMOD Conference*, Melbourne, 31 May-4 June, 2015, 1809-1824.