第9章 知识图谱应用案例

王昊奋 上海乐言信息科技有限公司，丁军 华东理工大学

知识图谱用以描述现实世界中的概念、实体以及它们之间丰富的关联关系。自从2012年谷歌公司利用知识图谱改善搜索体验并提高搜索质量后，引起了社会各界纷纷关注。随着知识图谱应用的深入，作为一种知识表示的新方法和知识管理的新思路，知识图谱不再局限于搜索引擎及智能问答等通用领域应用，而在越来越多的垂直应用领域开始崭露头角，扮演越来越重要的角色。通用知识图谱可以形象地看成一个面向通用领域的“结构化的百科知识库”，其中包含了现实世界中的大量常识，覆盖面极广。领域知识图谱又称为行业知识图谱或垂直知识图谱，通常面向某一特定领域。领域知识图谱基于行业数据构建，通常有着严格而丰富的数据模式，对该领域知识的深度、准确性有着更高的要求。本章重点介绍领域知识图谱的构建方法及系列应用案例。

9.1 领域知识图谱构建的技术流程

由于现实世界的知识丰富多样且极其庞杂，通用知识图谱主要强调知识的广度，通常运用百科数据进行自底向上的方法进行构建。而领域知识图谱面向不同的领域，其数据模式不同，应用需求也各不相同，因此没有一套通用的标准和规范来指导构建，而需要基于特定行业通过工程师与业务专家的不断交互与定制来实现。虽然如此，领域知识图谱与通用知识图谱的构建与应用也并非完全没有互通之处，如图9-1所示，其从无到有的构建过程可分为六个阶段，被称为领域知识图谱的生命周期[1]。本节以生命周期为视角来阐述领域知识图谱构建过程中的关键技术流程。

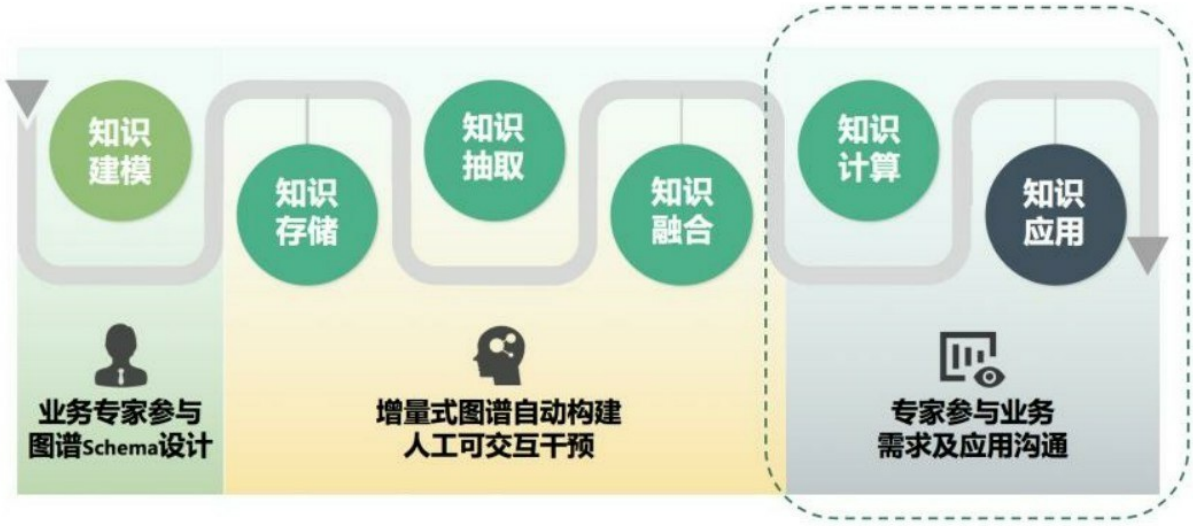

图9-1 领域知识图谱生命周期

9.1.1 领域知识建模

知识建模是建立知识图谱的概念模式的过程，相当于关系数据库的表结构定义。为了对知识进行合理的组织，更好地描述知识本身与知识之间的关联，需要对知识图谱的模式进行良好的定义。一般来说，相同的数据可以有若干种模式定义的方法，设计良好的模式可以减少数据的冗余，提高应用效率。因此，在进行知识建模时，需要结合数据特点与应用特点来完成模式的定义。

知识建模通常采用两种方式：一种是自顶向下（Top-Down）的方法，即首先为知识

图谱定义数据模式，数据模式从最顶层概念构建，逐步向下细化，形成结构良好的分类学层次，然后再将实体添加到概念中。

另一种则是自底向上（Bottom-Up）的方法，即首先对实体进行归纳组织，形成底层概念，然后逐步往上抽象，形成上层概念。该方法可基于行业现有标准转换生成数据模式，也可基于高质量行业数据源映射生成。

为了保证知识图谱质量，通常在建模时需要考虑以下几个关键问题：

1）概念划分的合理性，如何描述知识体系及知识点之间的关联关系[1]；

2）属性定义方式，如何在冗余程度最低的条件下满足应用和可视化展现；

3）事件、时序等复杂知识表示，通过匿名节点的方法还是边属性的方法来进行描述，各自的优缺点是什么[2]；

4）后续的知识扩展难度，能否支持概念体系的变更以及属性的调整。

关于知识建模的详细知识和技术，请参考本书第2章。

9.1.2 知识存储

知识存储，顾名思义为针对构建完成的知识图谱设计底层存储方式，完成各类知识的存储，包括基本属性知识、关联知识、事件知识、时序知识、资源类知识等。知识存储方案的优劣会直接影响查询的效率，同时也需要结合知识应用场景进行良好的设计。

目前，主流的知识存储解决方案包括单一式存储和混合式存储两种。在单一式存储中，可以通过三元组、属性表或者垂直分割等方式进行知识的存储[3]。其中，三元组的存储方式较为直观，但在进行连接查询时开销巨大[4]。属性表指基于主语的类型划分数据表，其缺点是不利于缺失属性的查询[5]。垂直分割指基于谓词进行数据的划分，其缺点是数据表过多，且写操作的代价比较大[6]。

对于知识存储介质的选择，可以分为原生（Neo4j、AllegroGraph 等）和基于现有数据库（MySQL、Mongo 等）两类。原生存储的优点是其本身已经提供了较为完善的图查询语言或算法的支持，但不支持定制，灵活程度不高，对于复杂节点等极端数据情况的表现非常差。因此，有了基于现有数据库的自定义方案，这样做的好处是自由程度高，可以根据数据特点进行知识的划分、索引的构建等，但增加了开发和维护成本。

从上述介绍中可以得知，目前尚没有一个统一的可以实现所有类型知识存储的方式。因此，如何根据自身知识的特点选择知识存储方案，或者进行存储方案的结合，以满足针对知识的应用需要，是知识存储过程中需要解决的关键问题。

关于知识存储的详细知识与技术，请参考本书第3章。

9.1.3 知识抽取

知识抽取是指从不同来源、不同数据中进行知识提取，形成知识并存入知识图谱的过程。由于真实世界中的数据类型及介质多种多样，所以如何高效、稳定地从不同的数据源进行数据接入至关重要，其会直接影响到知识图谱中数据的规模、实时性及有效性。

在现有的数据源中，数据大致可分为三类：一类是结构化的数据，这类数据包括以关系数据库（MySQL、Oracle 等）为介质的关系型数据，以及开放链接数据，如 Yago、Freebase 等；第二类为半结构化数据，如百科数据（Wikipedia、百度百科等），或是垂直网站中的数据，如IMDB、丁香园等；第三类是以文本为代表的非结构化数据。

结构化数据中会存在一些复杂关系，针对这类关系的抽取是此类研究的重点，主要方法包括直接映射或者映射规则定义等；半结构化数据通常采用包装器的方式对网站进行解析，包装器是一个针对目标数据源中的数据制定了抽取规则的计算机程序。包装器的定义、自动生成以及如何对包装器进行更新及维护以应对网站的变更，是当前获取需要考虑的问题；非结构化数据抽取难度最大，如何保证抽取的准确率和覆盖率是这类数据进行知识获取需要考虑的科学问题。

关于知识抽取的详细知识和技术，请参考本书第4章。

9.1.4 知识融合

知识融合指将不同来源的知识进行对齐、合并的工作，形成全局统一的知识标识和关联。知识融合是知识图谱构建中不可缺少的一环，知识融合体现了开放链接数据中互联的思想。良好的融合方法能有效地避免信息孤岛，使得知识的连接更加稠密，提升知识应用价值，因此知识融合是构建知识图谱过程中的核心工作与重点研究方向。

知识图谱中的知识融合包含两个方面，即数据模式层的融合和数据层的融合。数据模式层的融合包含概念合并、概念上下位关系合并以及概念的属性定义合并，通常依靠专家人工构建或从可靠的结构化数据中映射生成。在映射的过程中，一般会通过设置融合规则确保数据的统一。数据层的融合包括实体合并、实体属性融合以及冲突检测与解决。

进行知识融合时需要考虑使用什么方式实现不同来源、不同形态知识的融合；如何对海量知识进行高效融合[7]；如何对新增知识进行实时融合以及如何进行多语言融合等问题[8]。

关于知识融合的详细知识和技术，请参考本书第5章。

9.1.5 知识计算

知识计算是领域知识图谱能力输出的主要方式，通过知识图谱本身能力为传统的应用形态赋能，提升服务质量和效率。其中，图挖掘计算和知识推理是最具代表性的两种能力，如何将这两种能力与传统应用相结合是需要解决的一个关键问题。

知识推理一般运用于知识发现、冲突与异常检测，是知识精细化工作和决策分析的主要实现方式。知识推理又可以分为基于本体的推理和基于规则的推理。一般需要依据行业应用的业务特征进行规则的定义，并基于本体结构与所定义的规则执行推理过程，给出推理结果。知识推理的关键问题包括：大数据量下的快速推理，记忆对于增量知识和规则的快速加载[9]。

知识图谱的挖掘计算与分析指基于图论的相关算法，实现对图谱的探索与挖掘。图计算能力可辅助传统的推荐、搜索类应用。知识图谱中的图算法一般包括图遍历、最短路径、权威节点分析、族群发现最大流算法、相似节点等，大规模图上的算法效率是图算法设计与实现的主要问题。

关于知识推理与分析的详细知识和技术，请参考本书的第6章。

9.1.6 知识应用

知识应用是指将知识图谱特有的应用形态与领域数据和业务场景相结合，助力领域业务转型。知识图谱的典型应用包括语义搜索、智能问答以及可视化决策支持。如何针对业务需求设计实现知识图谱应用，并基于数据特点进行优化调整，是知识图谱应用的关键研究内容。

其中，语义搜索是指基于知识图谱中的知识，解决传统搜索中遇到的关键字语义多样性及语义消歧的难题，通过实体链接实现知识与文档的混合检索。语义检索需要考虑如何解决自然语言输入带来的表达多样性问题，同时需要解决语言中实体的歧义性问题。

而智能问答是指针对用户输入的自然语言进行理解，从知识图谱或目标数据中给出用户问题的答案。智能问答的关键技术及难点包括：

1）准确的语义解析，如何正确理解用户的真实意图。

2）对于返回的答案，如何评分以确定优先级顺序。

可视化决策支持则指通过提供统一的图形接口，结合可视化、推理、检索等，为用户提供信息获取的入口。对于可视化决策支持，需要考虑的关键问题包括：如何通过可视化方式辅助用户快速发现业务模式；如何提升可视化组件的交互友好程度，例如高效地缩放和导航；大规模图环境下底层算法的效率。

关于知识图谱的搜索及问答技术，请参考本书的第7章和第8章。

9.2 领域知识图谱构建的基本方法

不同领域的数据情况不同，有的领域较为成熟，知识体系完备，涵盖面广，单单采用自顶向下的方法进行图谱的构建就足以满足领域的应用。但在一些新兴领域，知识体系欠缺完备性，一部分知识适用于自顶向下构建，但也有很大一部分数据未成体系，这时则需要通过自底向上的方式对这类知识进行基于数据驱动的方式进行构建。因此，通常在领域内，尤其新兴领域，建模时会将自顶向下和自低向上的构建方法相结合。

9.2.1 自顶向下的构建方法

针对特定的行业内有固定知识体系或由该行业专家梳理后可定义模式的数据，大多采用自顶向下的方式构建。国内外现有可借助的建模工具以 Protégé、PlantData 为代表。Protégé[2]是一套基于 RDF（S）、OWL 等语义网规范的开源本体编辑器，拥有图形化界面，适用于原型构建场景。Protégé同时提供在线版本的 WebProtégé，方便在线进行知识图谱语义本体的自动构建。PlantData[3]知识建模工具是一款商用知识图谱智能平台软件。该软件提供了本体概念类、关系、属性和实例的定义和编辑，屏蔽了具体的本体描述语言，用户只需在概念层次上进行领域本体模型的构建，使得建模更加便捷。

为保证可靠性，数据模式的构建基本都经过了人工校验，因此知识融合的关键任务是数据层的融合。工业界在进行知识融合时，通常在知识抽取环节中就对数据进行控制，以减少融合过程中的难度及保证数据的质量。在这些方面，工业界均做了不同角度的尝试，如 DBpedia Mapping[4]采用属性映射的方式进行知识融合。zhishi.me 采用离线融合的方式识别实体间的 sameAs 关系，完成知识融合[10]，并通过双语主题模型，针对中英文下知识体系进行跨语言融合[11]。

接着，需要根据数据源的不同进行知识获取，其方法主要分为三种：第一种是使用D2R工具，该方法主要针对结构化数据，通过D2R工具将关系数据映射为RDF数据。常用的开源D2R工具有D2RQ[5]、D2R Server[6]、DB2triples[7]等。D2RQ通过D2RQ Mapping Language将关系数据转化成RDF数据，同时支持基于该语言在关系数据上直接提供RDF形式的数据访问 API;D2R Server 提供对 RDF 数据的查询访问接口，以供上层的 RDF浏览器、SPARQL 查询客户端以及传统的 HTML 浏览器调用；DB2triples 支持基于 W3C的R2RML和DM的标准将数据映射成RDF形式。

第二种是使用包装器，该方法主要针对半结构化数据，通过使用构建面向站点的包装器解析特定网页、标记语言文本。包装器通常需要根据目标数据源编写特定的程序，因此学者们的研究主要集中于包装器的自动生成。Ion Muslea等人[12]基于层次化信息抽取的思想，提出了一个包装器自动生成算法“STALKER”;Alberto Pan 等人[13]开发了一个名为“Wargo”的半自动生成包装器的工具。

第三种是借助信息抽取的方法，该方法主要针对非结构化的文本。按照抽取范围的不同，文本抽取可分为OpenIE和CloseIE两种。OpenIE面向开放领域抽取信息，是一种基于语言学模式的抽取，无法得知待抽取知识的关系类型，通常抽取规模大、精度较低。典型的工具有 ReVerb[8]、TextRunner[9]等。CloseIE 面向特定领域抽取信息，因其基于领域专业知识进行抽取，可以预先定义好抽取的关系类型，且通常规模小、精度较高。DeepDive是 CloseIE 场景中的典型工具，其基于联合推理的算法让用户只需要关心特征本身，让开发者更多地思考特征而不是算法。

9.2.2 自底向上的构建方法

在领域中部分没有完整知识体系的数据需要采用自底向上的方法进行构建，这与通用知识图谱的构建方法类似，主要依赖开放链接数据集和百科，从这些结构化的知识中进行自动学习，主要分为实体与概念的学习、上下位关系的学习、数据模式的学习。

开放链接数据集和百科中拥有丰富的实体和概念信息，数据通常以一定的结构组织生成，因此从这类数据源中抽取概念和实体较为容易。由于百科的分类体系都是经过了百科管理员或是高级编辑人员的校验，其分类系统中的数据可靠性非常高，因此从百科中抽取概念和实体，通常将标题作为实体的候选，而将百科中的分类系统直接作为概念的候选。对于概念的学习，关键[14]提出了一种基于语言学和基于统计学的多策略概念抽取方法，该方法提高了领域内概念抽取的效果。

实体对齐的目标是将从不同百科中学习到的、描述同一目标的实体或概念进行合并，再将合并后的实体集与开放链接数据集中抽取的实体进行合并。实体对齐过程主要分为六步：

- 从开放链接数据集中抽取同义关系。
- 基于结构化的数据对百科中的实体进行实体对齐。
- 采用自监督的实体对齐方法对百科的文章进行对齐。
- 将百科中的实体与链接数据中的实体进行对齐。
- 基于语言学模式的方法抽取同义关系。
- 实体基于CRF的开放同义关系抽取方法学习同义词关系。

黄峻福[15]提出了一种基于实体属性信息及上下文主题特征相结合进行实体对齐的方法。万静等人[16]提出了一种独立于模式的基于属性语义特征的实体对齐方法。

对于上下位关系，开放链接数据集中拥有明确的描述机制，针对不同的数据集，编写相应的规则直接解析即可获取。百科中描述了两种上下位关系，一种是类别之间的上下位关系，对应概念的层次关系；另一种则是类别与文章之间的上下位关系，对应实体与概念之间的从属关系。实体对齐可从开放链接数据集和百科中抽取上下位关系。WANG 等人[17]引入了弱监督学习框架提取来自用户生成的类别关系，并提出了一种基于模式的关系选择方法，解决学习过程中“语义漂移”问题。

数据模式的学习又称为概念的属性学习，一个属性的定义包含三个部分：属性名、属性的定义域、属性的值域。但概念的属性被定义好，属于该属性的实体则默认具备此属性，填充属性的值即可。概念属性的变更会直接影响到它的实体、其子概念以及这些概念下的实体。因此概念的属性定义十分重要，通常大部分知识库中的概念属性都是采用人工定义等方式生成的，通用知识图谱则可以从开放数据集中获取概念的属性，然后从在线百科中学习实体的属性，并对实体属性进行往上规约从而生成概念的属性。在进行属性往上规约的过程中，需要通过一定的机制保证概念属性的准确性，对于那些无法自动保证准确性的属性，需要进行人工校验。SU[18]提出了一种新的半监督方法，从维基百科页面自动提取属性。Logan I V等人[19]提出了多模态属性提取的任务，用来提取实体的基础属性。

9.3 领域知识图谱的应用案例

典型的通用知识图谱项目有　　DBpedia、WordNet、ConceptNet、YAGO、Wikidata等，本书第1章已有详细介绍。如图9-2所示，领域知识图谱常常用来辅助各种复杂的分析应用或决策支持，在多个领域均有应用，不同领域的构建方案与应用形式则有所不同，本节将以电商、图书情报（以下简称“图情”）、生活娱乐、企业商业、创投、中医临床、金融证券七个领域为例，从图谱构建与知识应用两个方面介绍领域知识图谱的技术构建应用与研究现状。

图9-2 行业知识图谱应用一览[10]

9.3.1 电商知识图谱的构建与应用[11]

当下，电商的交易规模巨大，对每个人的生活都有影响。随着 O2O 和零售行业的发展，电商交易场景不再是单纯的线上交易场景，而是新零售、多语言、线上线下相结合的复杂购物场景，电商企业对数据互联的需求越来越强烈。在此基础上，电商交易逐渐转变为集　B2C、B2B、跨境为一体，覆盖“实物+虚拟”商品，结合跨领域搜索发现、导购、交互多功能的新型电商交易。因而电商知识图谱变得非常重要。相对于通用知识图谱，它有很多不同之处。首先，电商平台是围绕着商品，买卖双方在线上进行交易的平台。故而电

商知识图谱的核心是商品。整个商业活动中有品牌商、平台运营、消费者、国家机构、物流商等多角色参与，相对于网页来说，数据的产生、加工、使用、反馈控制得更加严格，约束性更强。如果电商数据以知识图谱的方法组织，可以从数据的生产端开始，就遵循顶层设计。电商数据的结构化程度相对于通用域来说做得更好。此外，面向不同的消费者和细分市场，不同角色、不同市场、不同平台对商品描述的侧重都不同，使得对同一个实体描述时会有不同的定义。知识融合就变得非常重要。最后，与通用知识图谱比较而言，电商知识图谱有大量的国家标准、行业规则、法律法规对商品描述进行着约束。存在大量的人的经验来描述商品做到跟消费者需求的匹配，知识推理显得更为重要。下面以阿里巴巴知识图谱为例，介绍电商知识图谱的相应技术模块和应用。

在商品知识的表示方面，电商知识图谱以商品为核心，以人、货、场为主要框架。目前共涉及9大类一级本体和27大类二级本体。一级本体分别为人、货、场、百科知识、行业竞争对手、品质、类目、资质和舆情。人、货、场构成了商品信息流通的闭环，其他本体主要给予商品更丰富的信息描述。如图9-3所示为电商知识图谱的数据模型，数据来源包含国内—国外数据、商业—国家数据、线上—线下等多源数据。目前有百亿级的节点和百亿级的关系边。

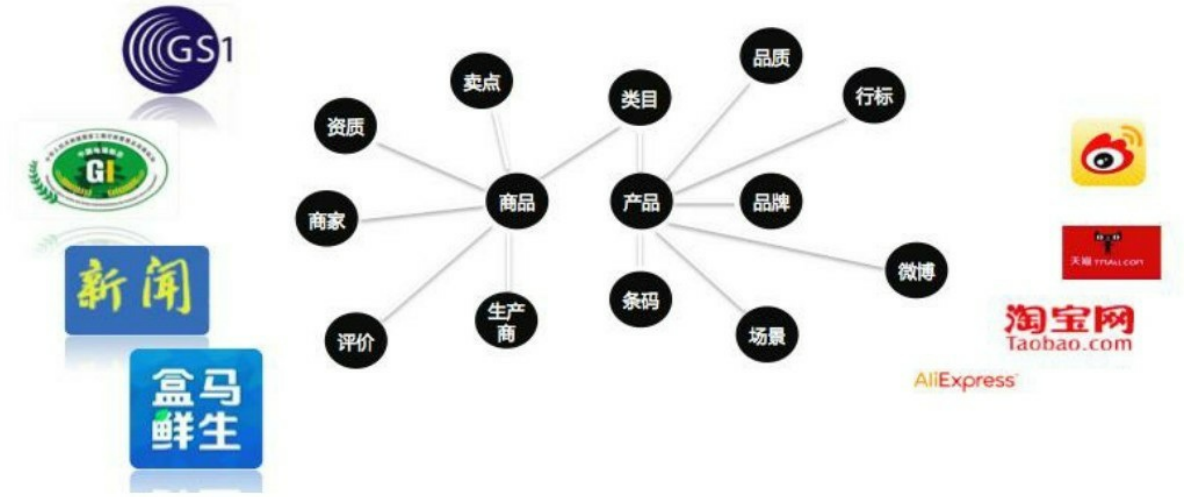

图9-3 电商知识图谱的数据模型

电商知识图谱主要的获取来源为知识众包，这其中的关键就是知识图谱本体设计。在设计上要考虑商品本身，又要考虑消费者需求和便于平台运营管理。另一个核心工作是要开发面向电商各种角色的数据采集工具，例如面向卖家的商品发布端。此外，电商知识的另一个来源是文本数据，例如商品标题、图片、详情、评价，舆情中的品牌、型号、卖点、场景等信息。这就要求命名识别系统具有跨越大规模实体类型的识别能力，能够支持电商域数据、人机语言交互自然语言问题以及更广泛的微博、新闻等舆情域数据的识别，

并且把识别出的实体与知识图谱链接，特别是商品属性和属性值涉及上千类别的实体类型。主要包括：

●商品域：类目、产品词、品牌、商品属性、属性值、标准产品。

●LBS域：小区、超市、商场、写字楼、公司。

●通用域：人物、数字、时间。

最后，对知识图谱实体描述，除了基础的属性与属性值，很多是通过实体标签来实现的。相对来说，标签变化快、易扩展。很大一部分这类知识是通过推理获得的。例如，在食品的标签生成中，知识推理通过食品的配料表数据和国家行业标准，如：

●无糖：碳水化合物含量小于或等于0.5 g /100 g（固体）或0.5g/100 mL（液体）；

●无盐：钠含量小于或等于5 mg /100 g 或5mg/100 mL。

通过推理，可以把配料表数据转化为“无糖”“无盐”等知识点，从而真正地把数据变成了知识标签，并改善消费者的购物体验。

大量的多源异构数据的汇集需要考虑知识的融合，主要涉及商品和产品两个核心节点知识融合。主要利用大规模聚类、大规模实体链指、大规模层次分类等技术，依据商品或产品的图片、文本、属性结构化等数据。图片涉及相似图计算、OCR等技术。

大规模层次分类需要把目标商品或产品归到上千个商品1级和2级类目中去。这里面的难度在于类目的细分和混淆度，以及大规模训练数据的生成和去噪。

大规模聚类的目的是把统一数据源的信息先做一次融合。大规模实体链指的核心是通过知识图谱的候选实体排序，把新的实体与知识图谱目标识别进行关联，从而把新知识融入知识图谱。在新知识融入工程中，涉及不同数据源属性名称和属性值的映射和标准化。这就需要大规模电商词林的建设和挖掘。

通常来说，电商知识图谱的实体量比通用知识图谱的实体量要大很多，选择存储方案时，需要考虑很多因素，例如支持的查询方式、支持的图查询路径长度、响应时间、机器成本等。因此，存储主要采取多种存储方式混合的方案。另一方面，考虑到成本因素，全量的图谱数据通过离线关系数据库存储，共包含实体表、关系表、类目表三种表类型。为了更好地支持在线图查询和逻辑查询，与在线业务相关的知识图谱子图采用在线图数据库来存储。离线关系数据库支持向在线图数据库导入。考虑图数据的查询性能与节点路径长度关系很大，为保证毫秒级的在线响应，部分数据采用在线关系数据库支持查询。

在应用方面，作为商品大脑，电商知识图谱的一个主要应用场景就是智能导购。而所谓导购，就是让消费者更容易找到他们想要的东西，例如说买家输入“我需要一件漂亮的真丝丝巾”，商品大脑会通过语法词法分析来提取语义要点“一”“漂亮”“真丝”“丝巾”这些关键词，从而帮买家搜索到合适的商品。在导购中，为了让发现更简单，商品大脑还学习了大量的行业规范与国家标准，比如说全棉、低糖、低嘌呤等。此外，商品大脑可以从公共

媒体、专业社区的信息中识别出近期热词，跟踪热点词的变化，由运营确认是否成为热点词，这也是为什么买家在输入斩男色、禁忌之吻、流苏风等热词后，出现了自己想要的商品。最后，商品大脑还能通过实时学习构建出场景。例如输入“海边玩买什么”，结果中就会出现泳衣、游泳圈、防晒霜、沙滩裙等商品。

再者，电商平台管控从过去的“巡检”模式升级为发布端实时逐一检查。在海量的商品发布量的挑战下，最大限度地借助大数据和人工智能阻止坏人、问题商品进入电商生态。为了最大限度地保护知识产权，保护消费者权益，电商知识图谱推理引擎技术满足了智能化、自学习、毫秒级响应、可解释等更高的技术要求。例如，上下位和等价推理，检索父类时，通过上下位推理把子类的对象召回，同时利用等价推理（实体的同义词、变异词、同款模型等）扩大召回。以拦截“产地为某核污染区域的食品”为例，推理引擎翻译为“找到产地为该区域，且属性项与‘产地’同义，属性值是该区域下位实体的食品，以及与命中的食品是同款的食品”。

9.3.2 图情知识图谱的构建与应用[12]

图情知识图谱是指聚焦某一特定细分行业，以整合行业内图情资源为目标的知识图谱。提供知识搜索、知识标引、决策支持等形态的知识应用，服务于行业内的从业人员、科研机构及行业决策者。

图情领域与知识图谱的结合由来已久。英国的大英博物馆通过结合语义技术对馆藏品各类数据资源进行语义组织，通过语义细化、多媒体资源标注等方式提供多样化的知识服务形式[13]；英国广播公司 BBC[20]在其音乐、体育野生动物等板块定义了知识本体，将新闻转化为机器可读的信息源（RDF / XML、JSON和XML）进行内容管理与自动生成报道。国内图情领域也越来越重视对知识图谱技术的利用。上海图书馆[14]借鉴美国国会书目框架BibFrame[21]对家谱、名人、手稿等资源构建知识体系，打造家谱服务平台，为研究者们提供古籍循证服务；中国农业科学院[15]则聚焦于水稻细分领域，整合论文、专利、新闻等行业资源，构建水稻知识图谱，为科研工作者提供了行业专业知识服务平台。

图情知识图谱的构建一般采用自顶向下的方式进行知识建模，通常从资源类型数据入手，整理出资源的发表者（人物）、发表机构（机构）、关键词（知识点）、发表载体（刊物）等类型的实体及各自之间的关系，同时通过人物、机构的主页进行实体属性的扩充。如图9-4所示为图情知识图谱Schema模型，展示了概念与概念间的关系以及部分属性。

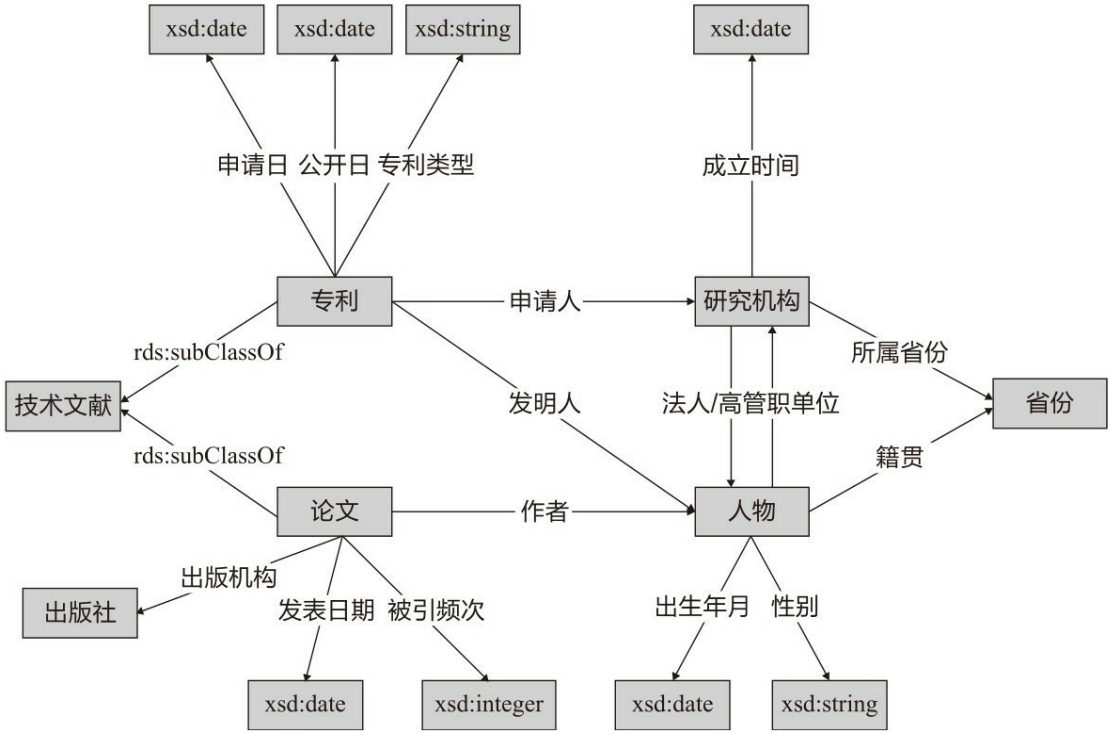

图9-4 图情知识图谱Schema模型

接下来分别对图情领域的数据进行获取，数据源主要包括四类。第一类是知网、专利局等文献类网站，第二类是开放通用数据，包括百科类网站以及DBpedia等的开放链接数据集，第三类是行业垂直的新闻门户，第四类是行业内企业和科研机构内部积累的既有数据。知识获取的方法视数据类型而异，具体可参考本章9.1.3节的介绍。

图情领域的知识融合需要考虑实体层面的融合以及知识体系的融合。对于实体融合，主要解决不同来源实体的属性缺失、冲突等问题，一般采用多数投票的方式进行实体属性的对齐。对于多知识体系的融合，通常确定置信度最高的体系作为基准，如专利的IPC分类，继而将其他来源的知识点进行对齐。由于知识体系的质量影响到了整个知识图谱的知识描述能力与准确性，所以一般允许较多的人工介入来进行体系的融合梳理。

图情知识图谱的存储设计需要兼顾实体、概念等图谱数据与论文、新闻等资源类型数据。对于图谱数据，推荐使用基于 RDF 的存储，如 AllegroGraph、Jena 等，它们对数据中的语义描述有着天然的支持，能更快地实现语义搜索等应用。对于资源数据，则可以使用面向搜索设计的数据库，如Elasticsearch、Solr等，以获得更好的搜索支持。

图情领域中的知识计算主要包括图论算法、知识统计以及知识推理。通过实现基本图论算法来辅助进行各类业务分析。例如，通过图遍历算法进行机构合作的谱系分析；基于社区发现算法寻找学术研究热点；借助图排序算法进行权威分析等。通过统计学方法进行宏观层面的分析，如行业发展趋势、机构研究分布等。通过知识推理完成新知识的补充，

如专家合作关系、公司上下游关系等。

图情知识图谱的典型应用包括知识搜索、知识标引、决策支持等。知识搜索是图情领域的基础性服务，而知识图谱技术可以从准确性和形态上为其赋能。图谱中的实体识别技术能够提高搜索的命中率，同时允许用户通过自然语言的方式进行知识的语义搜索。而通过知识卡片、知识推荐等结果的返回也可以提升用户的交互体验。如图9-5所示为大英博物馆语义搜索。

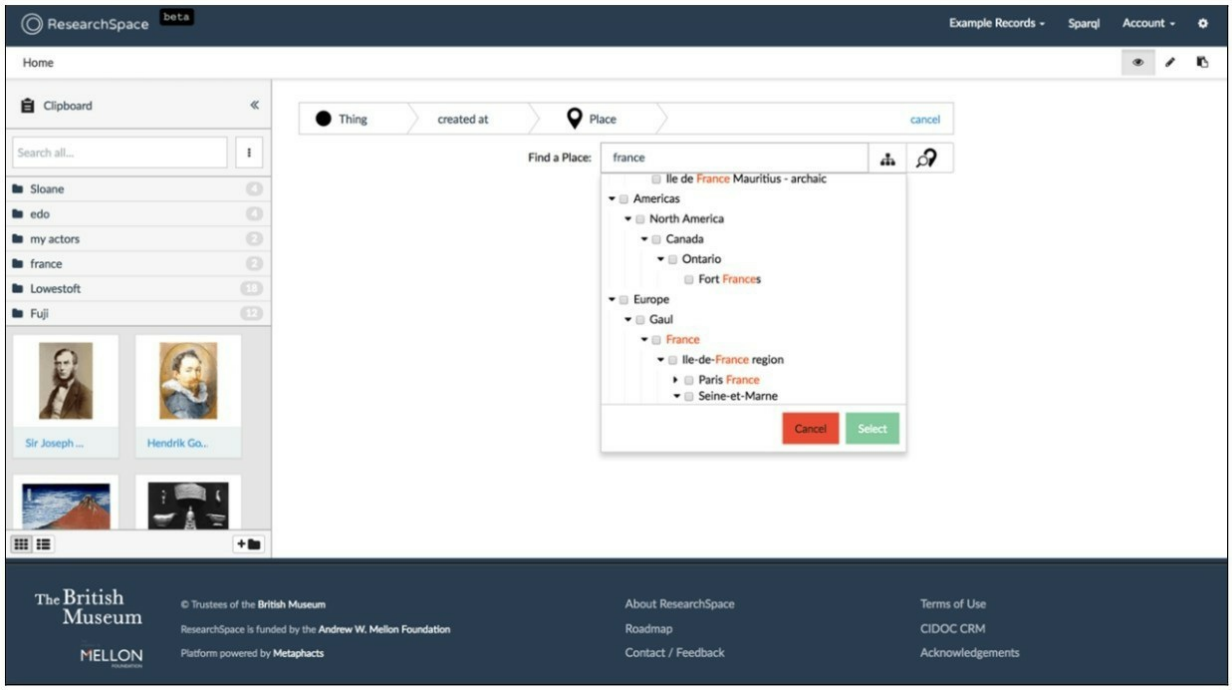

图9-5 大英博馆院语义搜索

知识标引指的是根据构建完成的图情知识图谱，对新闻、文献等文本的内容进行知识标注的过程。知识标引既是图谱构建过程中的重要工作，又是图谱应用的一种形态，可以依托标引技术打造在线的阅读工具，或者集成 Office、PDF reader 等文档类应用，提供知识卡片、知识推荐等服务，辅助终端用户阅读，如图9-6所示。

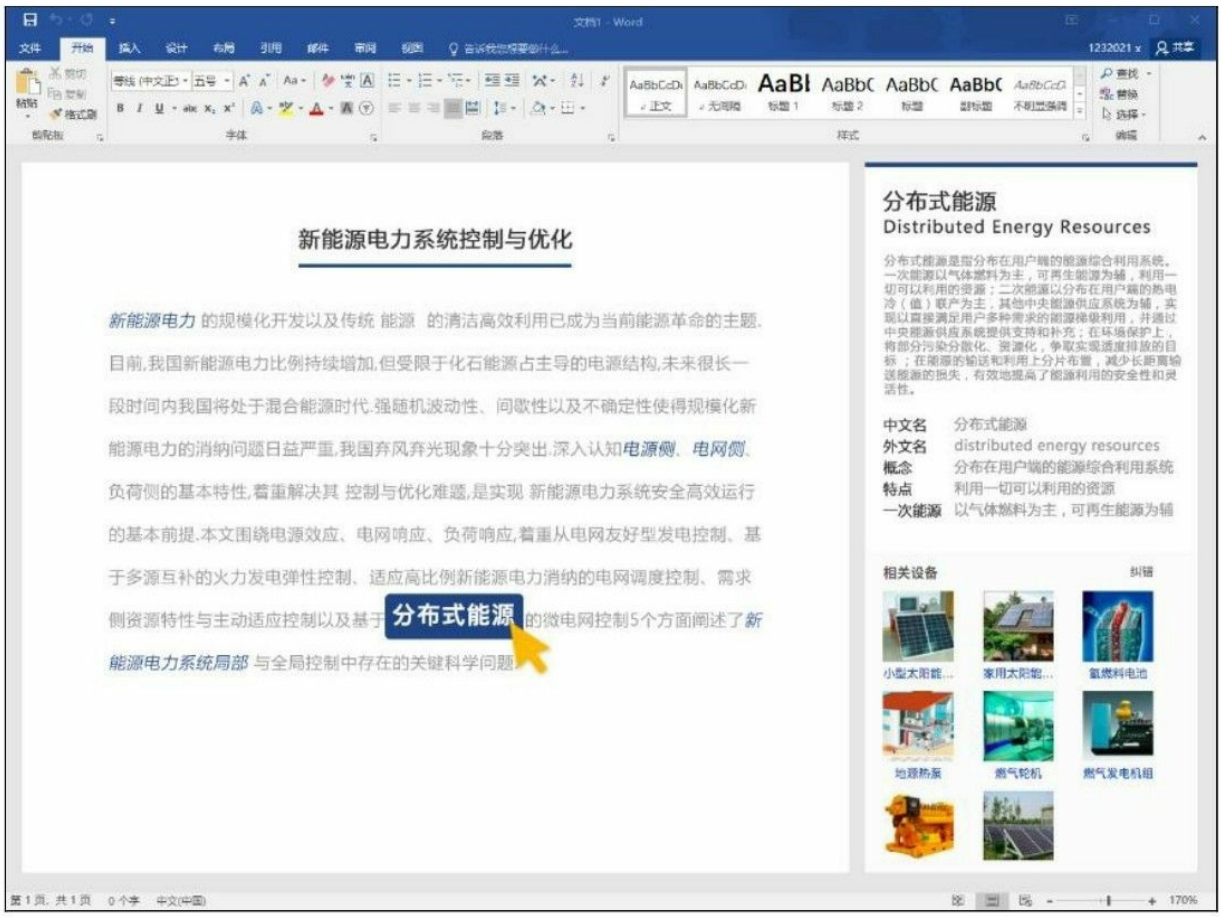

图9-6 基于知识标引的辅助阅读

决策支持基于路径分析、关联分析、节点聚类等图算法进行辅助分析，并通过图谱可视化的方式展示知识间的关联。可以对关联参数，如步长、过滤条件等，以及可视化的形态，如节点颜色、大小、距离等进行定制，从而为可视化决策支持赋予不同的业务含义。如图9-7和图9-8所示为典型的可视化决策支持场景。

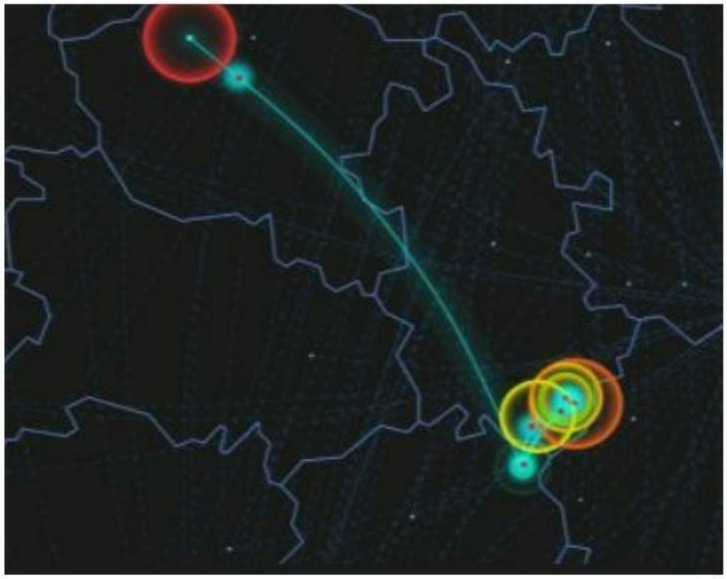

图9-7 上川明经胡氏家族迁徙图

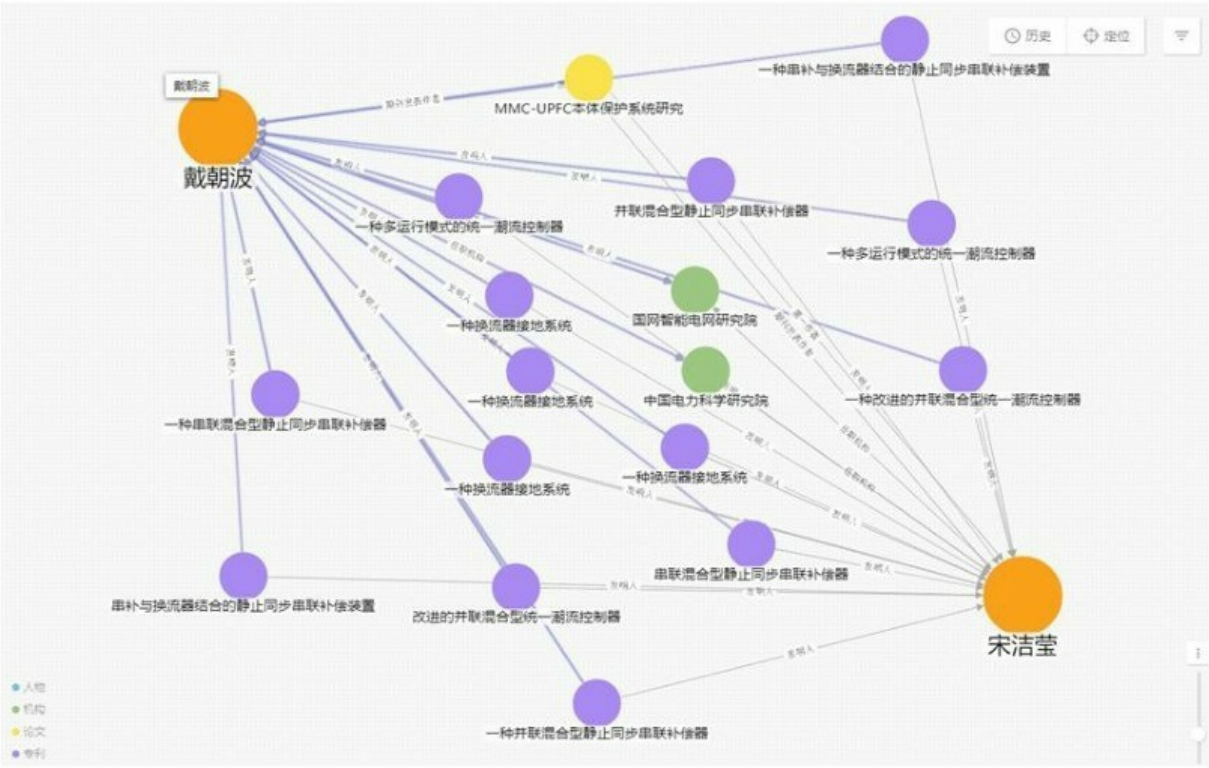

图9-8 专家合作分析

9.3.3 生活娱乐知识图谱的构建与应用：以美团为例[16][17]

1.美团点评知识图谱概述

海量数据和大规模分布式计算力催生了以深度学习为代表的新一代人工智能高潮。深度学习技术在语音、图像领域均取得了突破性的进展。然而，随着深度学习被广泛应用，其局限性也愈发明显。主要表现在以下四个方面：

（1）缺乏可解释性。神经网络端到端学习的“黑箱”特性使得很多模型不具有可解释性，导致很多需要人去参与决策。在这些应用场景中，机器结果无法完全置信而需要谨慎使用，例如医学的疾病诊断、金融的智能投顾等。这些场景属于低容错高风险场景，必须需要显示的证据支持模型结果，从而辅助人进行决策。

（2）常识缺失。人的日常活动需要大量的常识背景知识支持，数据驱动的机器学习和深度学习学习到的是样本空间的特征、表征，而大量的背景常识是隐式且模糊的，很难在样本数据中体现。例如，下雨要打伞，但打伞不一定都是下雨天。

（3）缺乏语义理解。模型并不理解数据中的语义知识，缺乏推理和抽象能力，对于未见数据模型泛化能力差。

（4）依赖大量样本数据。机器学习和深度学习需要大量标注样本数据去训练模型，而数据标注的成本很高，很多场景缺乏标注数据来进行冷启动。

从人工智能整体发展来说，综上的局限性也是机器从感知智能向认知智能迁跃过程中必须解决的问题。认知智能需要机器具备推理和抽象能力，需要模型能够利用先验知识，总结出人可理解、模型可复用的知识。机器计算能力在整体上需要从数据计算转向知识计算，知识图谱就显得必不可少。知识图谱可以组织现实世界中的知识，描述客观概念、实体、关系。这种基于符号语义的计算模型，一方面可以促成人和机器的有效沟通，另一方面可以为深度学习模型提供先验知识，将机器学习结果转化为可复用的符号知识并累积起来。

作为人工智能时代最重要的知识表示方式之一，知识图谱能够打破不同场景下的数据隔离，为搜索、推荐、问答、解释与决策等应用提供基础支撑。美团点评作为在线本地生活服务平台，覆盖了餐饮娱乐领域的众多生活场景，连接了数亿个用户和数千万家商户，积累了宝贵的业务数据，蕴含着丰富的日常生活相关知识。因此，美团点评 NLP 中心开始围绕吃喝玩乐等多种场景，构建了生活娱乐领域超大规模的知识图谱，为用户和商家建立起全方位的链接。通过对应用场景下的用户偏好和商家定位进行更为深度的理解，进而为大众提供更好的智能化服务。目前在建的美团大脑知识图谱有数十类概念、数十亿实体和数百亿三元组，美团大脑的知识关联数量预计在未来一年内将上涨到数千亿的规模。

美团点评积累了40亿的公开评价数据、3450万全球商家数据、1.4亿店菜数据以及10万个性化标签。针对大量的数据，需要从实际业务需求出发，在现有数据表之上抽象出数据模型，以商户、商品、用户等为主要实体，其基本信息作为属性，商户与商品、与用户

的关联为边，将多领域的信息关联起来，同时利用评论数据、互联网数据等，结合知识获取方法，填充图谱信息，从而提供更加多元化的知识。

另一方面，则需要采用Language Model（统计语言模型）、Topic Model（主题生成模型）以及Deep Learning Model（深度学习模型）等各种模型，对商家标签、菜品标签、情感分析进行挖掘。挖掘商户标签，需要先通过机器对用户评论进行阅读，这里采用了无监督模型与有监督的深度学习模型相结合的方式。无监督模型采用了 LDA，其特点是成本比较低，无须标注数据。当然，其他准确性比较不可控，同时对挖掘出来的标签还需要进行人工筛选。有监督的深度学习模型则采用了 LSTM，其特点是需要大量的标注数据。通过这两种模型挖掘出来的标签，再加上知识图谱里面的一些推理，最终构建出商户的标签。

其次，进行评论标签聚合，主要采用知识图谱推理技术与标签排序相结合的方式。举例来说，如果某商户的用户评价都围绕着宝宝椅、带娃吃饭、儿童套餐等话题，就可以得出很多关于这家商户的标签，如图9-9所示。例如可以知道它是一个亲子餐厅，环境比较别致，服务也比较热情等，这些新的标签可以基于知识图谱的推理来进行扩展。

图9-9 商户标签挖掘示意图

接下来，为了更精确地匹配菜品，丰富商户信息，需要对菜品标签进行挖掘。这需要对用户评论进行分析，提取菜品的描述信息。主要采用Bi-LSTM以及CRF模型。例如从某些评论里面可以抽取出一些实体，再通过与其他的一些菜谱网站做一些关联，建立关联更加丰富的菜品知识图谱，就可以得到它的食材、烹饪方法、口味等信息，这样就为每一个店菜挖掘出了非常丰富的口味标签、食材标签等各种各样的标签，如图9-10所示。

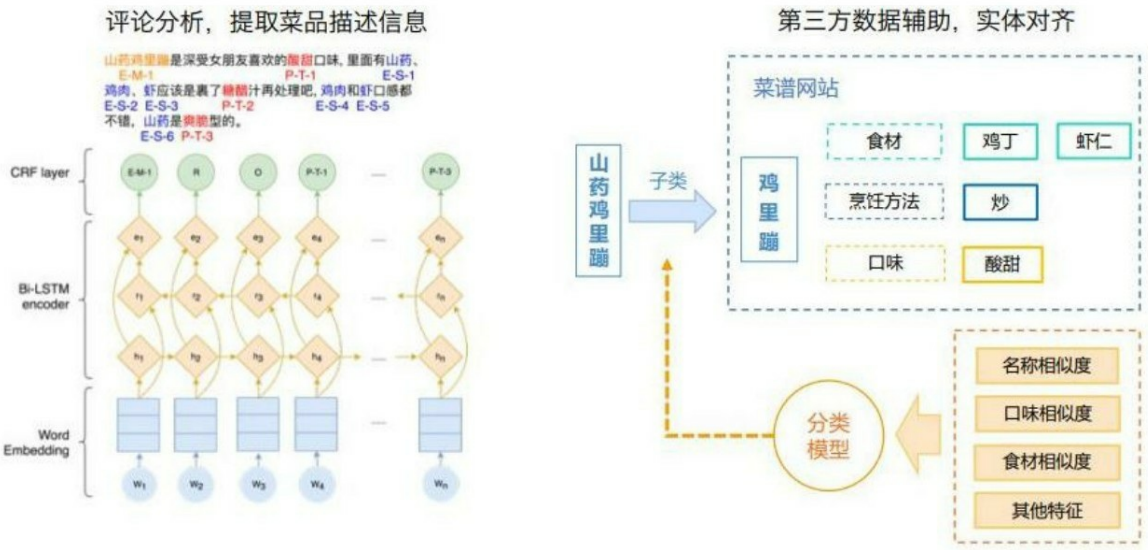

图9-10 菜品标签挖掘示意图

最后再对评论数据进行情感挖掘，主要采用CNN+LSTM 的模型，对每一个用户的评价进行分析，分析出用户的一些情感的倾向。同时，美团也正在做细粒度的情感分析，希望能够通过用户短短的评价，分析出用户在交通、环境、卫生、菜品、口味等不同维度方面的情感分析结果，如图9-11所示。

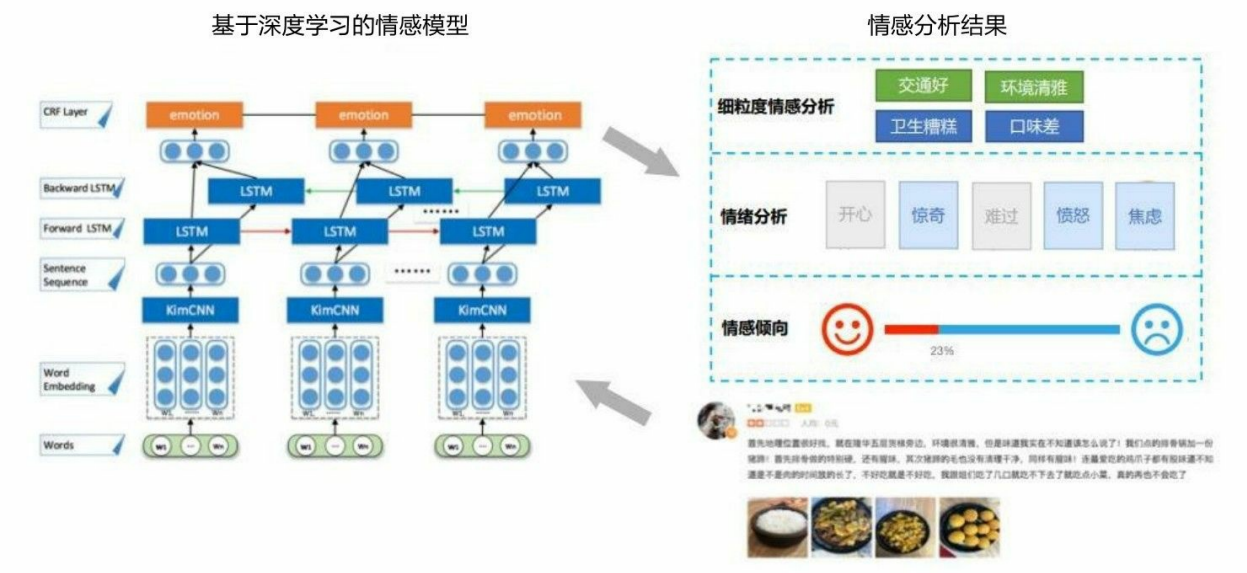

图9-11 情感挖掘示意图

2.美团**“**知识大脑**”**业务应用

依托深度学习模型，美团大脑充分挖掘、关联美团点评各个业务场景公开数据（如用户评价、菜品、标签等），构建餐饮娱乐“知识大脑”，并且已经开始在美团的不同业务中

落地，利用人工智能技术全面提升用户的生活体验。主要业务应用有智能搜索、ToB　商户赋能、金融风险管理和反欺诈。

（1）智能搜索：帮助用户做决策。知识图谱可以从多维度精准地刻画商家，已经在美食搜索和旅游搜索中应用，为用户搜索出更适合的店。基于知识图谱的搜索结果，不仅具有精准性，还具有多样性。如图9-12所示，当用户在美食类目下搜索关键词“鱼”时，未通过图谱搜索出来展示给用户的信息仅仅是包含关键词的“鱼”的相关结果；通过图谱可以认知到用户的搜索词是“鱼”这种“食材”。因此搜索的结果不仅有“糖醋鱼”“清蒸鱼”等精准的结果，还有“赛螃蟹”这样以鱼肉作为主食材的菜品，大大增加了搜索结果的多样性，提升用户的搜索体验。并且对于每一个推荐的商家，能够基于知识图谱找到用户最关心的因素，从而生成“千人千面”的推荐理由。例如，在浏览到大董烤鸭店的时候，偏好“无肉不欢”的用户 A 看到的推荐理由是“大董的烤鸭名不虚传”，而偏好“环境优雅”的用户 B 看到的推荐理由是“环境小资，有舞台表演”，不仅让搜索结果更具有解释性，同时也能吸引不同偏好的用户进入商家。

图9-12 知识图谱在点评搜索中的应用

对于场景化搜索，知识图谱也具有很强的优势。以七夕节为例，通过知识图谱中的七夕特色化标签，如约会圣地、环境私密、菜品新颖、音乐餐厅、别墅餐厅等，结合商家评论中的细粒度情感分析，为美团搜索提供了更多适合情侣过七夕节的商户数据，用于七夕场景化搜索的结果召回与展示，极大地提升了用户体验和用户点击转化。

（2）ToB 商户赋能：商业大脑指导店老板决策。美团大脑正应用在 SaaS 收银系统专业版中，通过机器智能阅读每个商家的每一条评论，可以充分理解每个用户对商家的感受。将大量的用户评价进行归纳总结，从而可以发现商家在市场上的竞争力、用户对于商家的总体印象趋势、菜品的受欢迎程度变化。进一步通过细粒度用户评论全方位分析，可以细致刻画商家服务现状，以及对商家提供前瞻性经营方向。通过美团 SaaS 收银系统专业版，这些智能经营建议将定期触达到各个商家，智能化指导商家精准优化经营模式。

在给店老板提供的传统商业分析服务中，主要聚焦于单店的现金流、客源分析。美团大脑充分挖掘了商户及顾客之间的关联关系，可以提供围绕商户到顾客，商户到所在商圈的更多维度的商业分析，在商户营业前、营业中以及经营方向，均可以提供细粒度的运营指导。

在商家服务能力分析上，通过图谱中关于商家评论所挖掘的主观、客观标签，例如“服务热情”“上菜快”“停车免费”等，同时结合用户在这些标签所在维度上的 Aspect 细粒度情感分析，告诉商家在哪些方面做得不错，是目前的竞争优势；在哪些方面做得还不够，需要尽快改进。因而可以更准确地指导商家进行经营活动。更加智能的是，美团大脑还可以推理出顾客对商家的认可程度，是高于还是低于其所在商圈的平均情感值，让店老板一目了然地了解自己的实际竞争力。

在消费用户群体分析上，美团大脑不仅能够告诉店老板顾客的年龄层、性别分布，还可以推理出顾客的消费水平，对于就餐环境的偏好，适合他们的推荐菜，让店老板有针对性地调整价格、更新菜品、优化就餐环境。

（3）金融风险管理和反欺诈：从用户行为建立征信体系。知识图谱的推理能力和可解释性在金融场景中具有天然的优势，美团 NLP 中心和美团金融共建的金融好用户扩散以及用户反欺诈，就是利用知识图谱中的社区发现、标签传播等方法来对用户进行风险管理，能够更准确地识别逾期客户以及用户的不良行为，从而大大提升信用风险管理能力。

在反欺诈场景中，知识图谱已经帮助美团金融团队在案件调查中发现并确认多起欺诈案件。由于团伙通常会存在较多关联及相似特性，关系图可以帮助识别出多层、多维度关联的欺诈团伙，能通过用户和用户、用户和设备、设备和设备之间的四度、五度甚至更深度的关联关系，发现共用设备、共用 Wi-Fi 来识别欺诈团伙，还可在已有的反欺诈规则上进行推理预测可疑设备、可疑用户来进行预警，从而成为案件调查的有力助手。

9.3.4 企业商业知识图谱的构建与应用[18]

丰富多维度的企业信息在基本面分析中十分重要，中国企业数量十分庞大，数据多源，需要构建统一的企业商业知识图谱。企业商业知识图谱包含企业、人物、专利等实体类型，以及任职、股权、专利所属权等关系类型，以完善企业及个人画像，助力企业潜在客户获取、客户背景调查、多层次研究报告、风险管控；辅助发现不良资产、企业风险、非法集资等。

典型的企业知识图谱，如量子魔镜[19]以全国全量企业的全景数据资源为研究基础，打造企业信用风险洞察平台；天眼查[20]、启信宝[21]则专注服务于个人与企业信息查询工具，为用户提供企业、工商、信用等相关信息的查询；企查查[22]立足于企业征信，通过深度学习、特征抽取以及知识图谱技术对相关信息进行整合，并向用户提供数据信息；中信建投将全国企业知识图谱整合进客户关系管理系统中，构建全面、清晰的客户视图，以实现高效客户关系管理。下面将企业商业知识图谱的构建方式进行梳理。

构建企业商业知识图谱，通常从相应网站中抽取企业信息、人物形象、诉讼信息以及信用信息，再添加上市公司、股票等概念和相应属性。企业招投标信息、上市公司的股票信息可从相关网站进行采集。企业的竞争关系、并购事件则从百科站点中进行抽取。这些信息存在于信息框、列表、表格等半结构化数据以及无结构的纯文本中。企业商业知识图谱如图9-13所示。

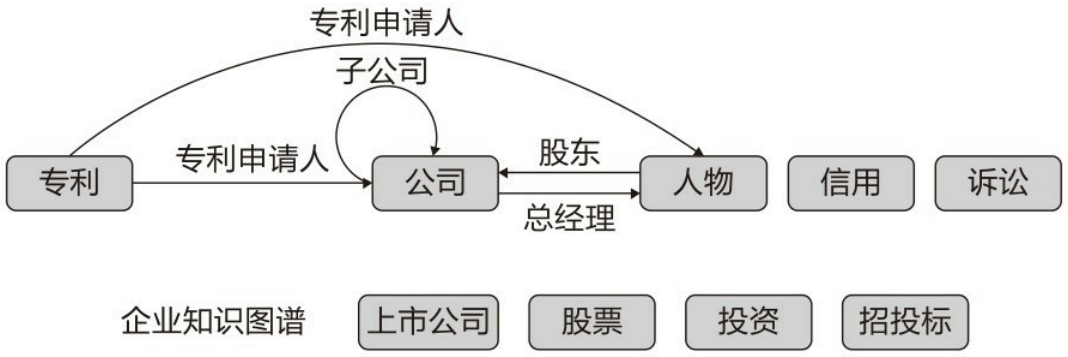

图9-13 企业商业知识图谱

企业商业知识图谱数据源主要包含两大类：

1）半结构化的网页数据，其中包括全国企业信用信息公示系统、中国裁判文书网、中国执行信息公开网、国家知识产权局、商标局、版权局等。

2）文本数据，如招投标信息公告、法律文书、新闻、企业年报等。通过 D2R 工具、

包装器、文本信息抽取等方式对以上数据分别进行抽取。由于数据来源多种多样，一方面涉及人物重名现象，另一方面，企业全称和简称产生的不一致问题也非常明显。因此，公司和人物两类实体是企业知识图谱融合的主要目标。公司的融合推荐基于公司名的全称进行链接，人物实例的融合推荐使用基于启发式规则进行集成。

全国企业商业知识图谱包含全国上千万家企业信息，10亿级别的三元组，形成知识图谱庞大而复杂，因此对存储方式提出了挑战，要求能够对海量的图数据进行存储，且具有良好的可伸缩性和灵活性。对此，推荐采用图数据库的方式进行存储，并可以扩展分布式存储方案以提高服务可用性与稳定性。

企业商业知识图谱的应用主要集中于金融反欺诈、辅助信贷审核的功能。例如，在金融反欺诈中，多个借款人联系方式的属性相同，但地址属性不同，可通过不一致性验证的方式来判断借款人是否有欺诈风险。

除此之外，通过异常关联挖掘、企业风险评估、关联探索、最终控制人和战略发展等方式，全国企业知识图谱为行业客户提供智能服务和风险管理。

异常关联挖掘是通过路径分析、关联探索等操作，挖掘目标企业谱系中的异常关联。基于企业商业知识图谱从多维度构建数据模型，进行全方位的企业风险评估，有效规避潜在的经营风险与资金风险，如图9-14所示。

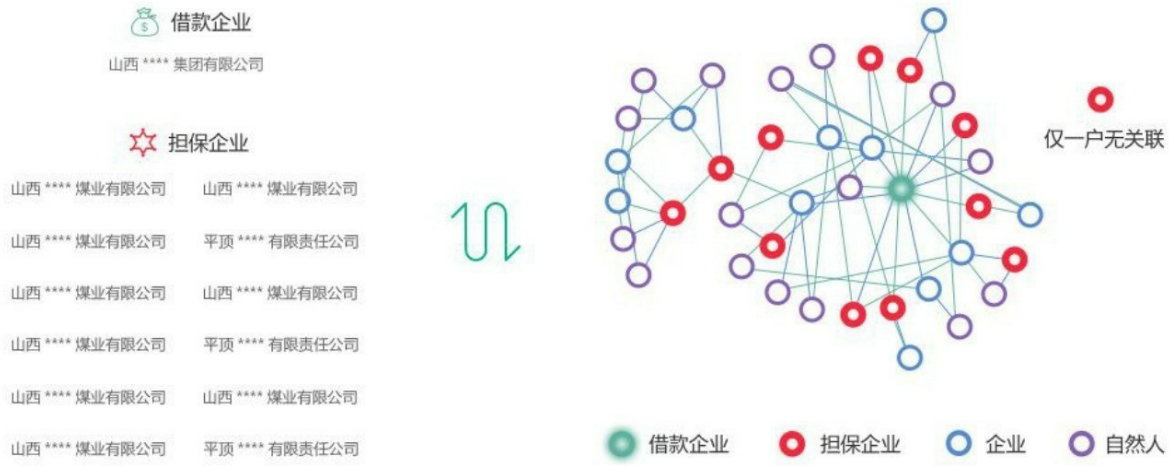

图9-14 异常关联挖掘

最终控制人分析是基于股权投资关系寻找持股比例最大的股东，最终追溯至自然人或国有资产管理部门，如图9-15所示。

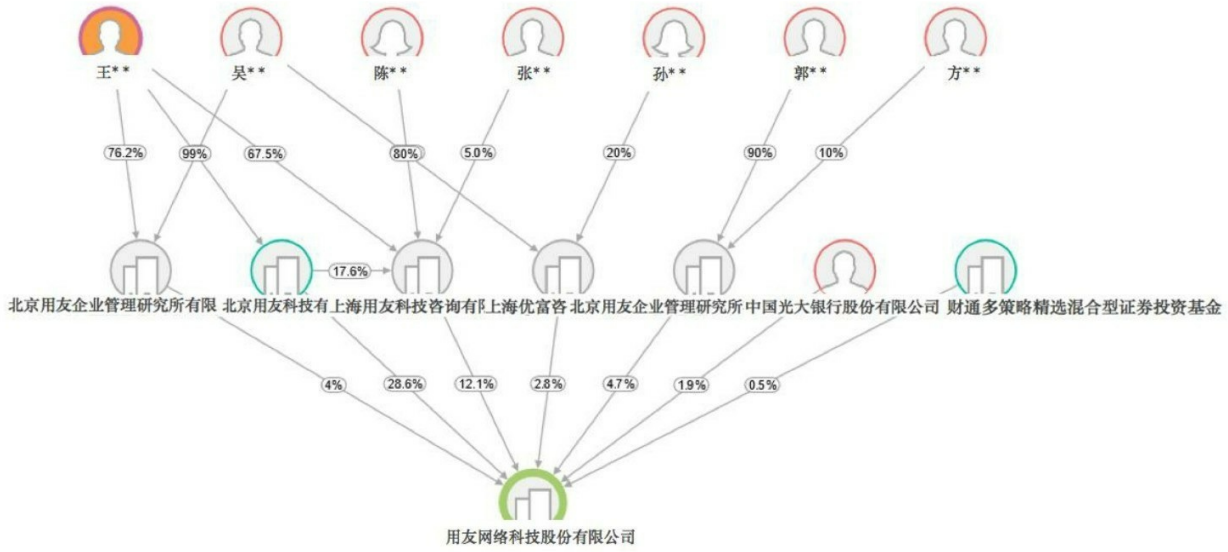

图9-15 最终控制人分析示例

战略发展则以“信任圈”的展现形式，将目标企业的对外投资企业从股权上加以区分，探寻其全资、控股、合营、参股的股权结构及发展战略，从而理解竞争对手和行业企业的真实战略，发现投资行业结构、区域结构、风险结构和年龄结构等，如图9-16所示。

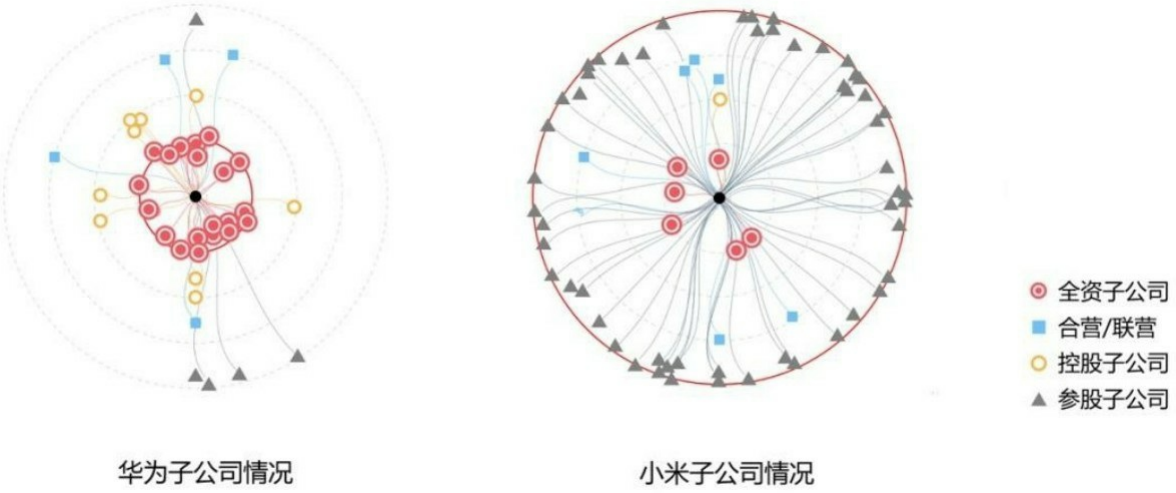

图9-16 企业社交图谱

9.3.5 创投知识图谱的构建与应用[23]

创业投资（以下简称“创投”）知识图谱聚焦于工商知识图谱的一部分数据内容，旨在展现企业、投融资事件、投资机构之间的关系。据IT桔子的不完全统计，截至2019年2

月，全国拥有初创公司超过12万家，投资机构超过7000家，有12万多名创业者，投资事件超过6万起。

作为公司发展过程中的重要阶段，创投领域的发展正得到越来越多数据与技术公司的关注。2007 年，在美国旧金山创立的 Crunchbase[24]，其核心业务是围绕初创公司及投资机构的生态为企业提供数据服务。国内企业中 TechNode 于 2017 年发布了数据棱镜平台[25]，构建创投知识图谱，为专业人员提供创业投资数据分析工具；因果树[26]是一家人工智能股权投融资服务平台，依托大数据和人工智能技术，提升一级市场效率，推动一级市场量化。

创投知识图谱的核心是投资，主要描述创业企业与投资机构之间以投资为主线的多种关系。因此，首先要理解创投领域的相关概念与关系。创投领域 Schema 中涉及的概念主要包括初创公司、投资机构、投资人、公司高管、行业以及投融资事件等。融资事件是创投领域的核心，不同于实体节点，融资事件描述的是一个事实，具有抽象性。典型的创投Schema如图9-17所示。

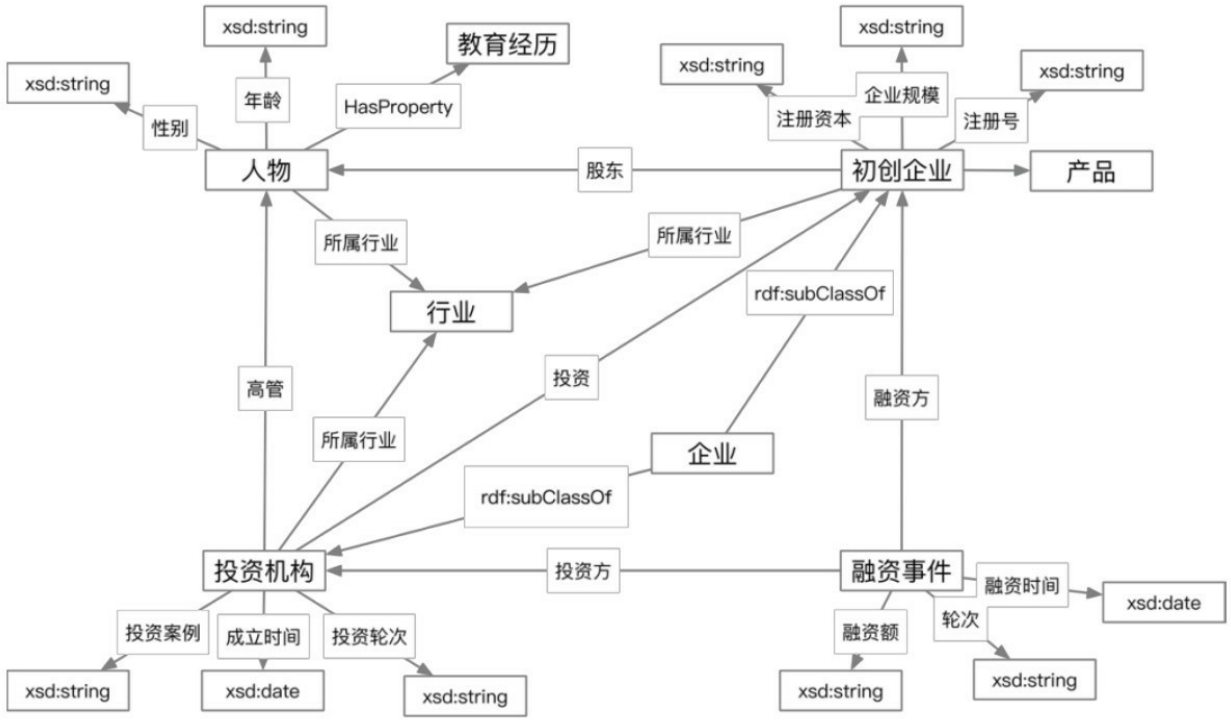

图9-17 典型的创投Schema

创投数据主要来源于虎嗅、IT 桔子、36Kr 等科技型媒体网站。IT 桔子是结构化的公司数据库和商业信息服务提供商，以融资事件为核心，关注 IT 互联网行业，其中包含了各类结构化的投资机构库和融资信息。虎嗅和 36Kr 则主要是以商业科技资讯为主的新闻

数据来源网站。构建创投知识图谱时，同样需要考虑数据融合的问题，典型问题包括：

1）数值属性表示不一致，例如金额的阿拉伯数字与中文写法的区别；

2）实体同义，例如企业的全称与简称；

3）不同数据源中的数据冲突。一般采用先实体对齐后属性对齐的方法来进行融合操作。

创投知识图谱的存储主要考虑融资事件的存储设计，通常采用两种方式对此类信息进行存储。第一种是在传统三元组的基础上加入其他描述字段，存储时间、轮次等信息；第二种方式是通过匿名节点存储事件，把时间、地点等相关信息作为事件节点的属性。对于融资事件来说，虽然它不是客观世界中一个具体的事物，但它包含了丰富的属性信息，如融资时间、融资轮次、融资额等。因此比较适合单独引入一类节点来进行存储和表示。

对于创投知识图谱的知识计算，主要通过使用社区发现、基于图的排序、最短路径等图算法，对合作分析、时序、相似公司等应用进行能力输出。例如，通过最短路径算法辅助合作分析，基于社区发现算法寻找行业研究热点，利用图排序算法进行权威分析等，通过分析展现公司的发展情况。

创投领域知识图谱主要的应用形态包括知识检索以及可视化决策支持。依托创投知识图谱，知识检索可以在原有知识全文搜索的基础上实现语义搜索与智能问答的应用形态。其中，语义搜索提供自然语言式的搜索方式，由机器完成用户搜索意图识别，如图9-18所示为语义搜索示例。而作为知识搜索的终极形态，智能问答允许用户通过对话的方式对领域内知识进行问答交互，同时通过配置问题模板实现复杂业务问题的回答，如图9-19所示为智能问答示例。


LengJing.io
北京小桔
问棱镜
关键字 北京小桔
企业 北京小桔科技有限公司
问棱镜 北京小桔科技有限公司怎么样
问棱镜 北京小桔科技有限公司新闻信息怎么样
问棱镜 北京小桔科技有限公司融资情况怎么样
摩拜单车
北京小桔科技有限公司
美团怎么样
蓝山投资怎么样
小米科技怎么样
北京小桔科技有限公司怎么样
最近的新闻
最近的xinw
北京小桔科技有限公司怎么样
北京小桔科技有限公司新闻信息怎...
清除搜索记录

图9-18 语义搜索示例

图9-19 智能问答示例

通过图谱可视化技术，决策支持可对创投图谱中的初创公司发展情况、投资机构投资偏好等进行解读。通过节点探索、路径发现、关联探寻等可视化分析技术，展示公司的全方位信息；通过知识地图、时序图谱等形态，对地理分布、发展趋势等进行解读，为投融资决策提供支持。如图9-20～图9-22所示分别为投融资知识图谱示例。


路径分析
图统计
图解读
搜索...
路径分析推荐
北京小桔科技有限公司 | 阿里巴巴（中国）有限公司
2017.04.15
路径分析推荐
北京小桔科技有限公司 | 晨兴资本
2017.03.08
新建分析
中信资本
投资企业
北京小桔科技有限公司
DST
中投公司
Temasek淡马锡
阿里巴巴（中国）有限公司

图9-20 路径分析

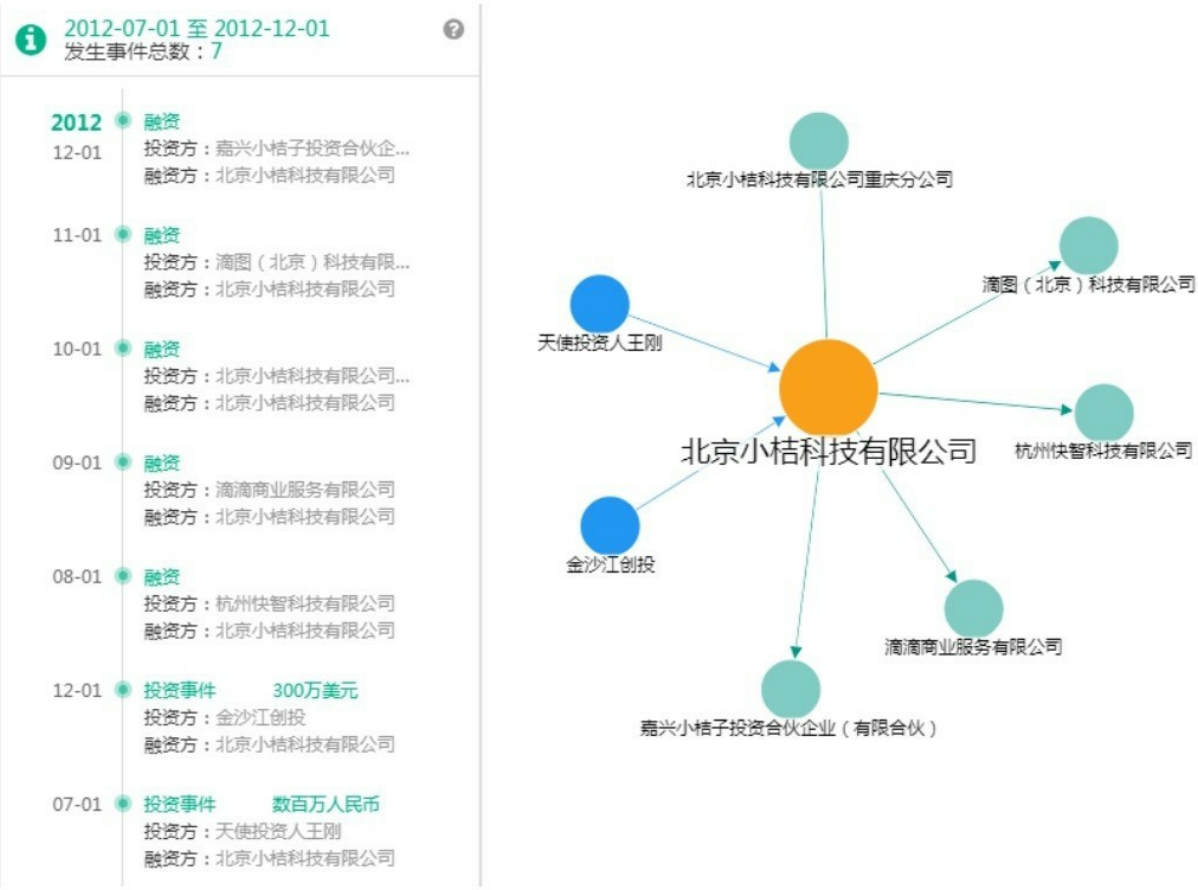
2012-07-01 至 2012-12-01
发生事件总数：7
2012 12-01 融资
投资方：嘉兴小桔子投资合伙企...
融资方：北京小桔科技有限公司
11-01 融资
投资方：滴图（北京）科技有限...
融资方：北京小桔科技有限公司
10-01 融资
投资方：北京小桔科技有限公司...
融资方：北京小桔科技有限公司
09-01 融资
投资方：滴滴商业服务有限公司
融资方：北京小桔科技有限公司
08-01 融资
投资方：杭州快智科技有限公司
融资方：北京小桔科技有限公司
12-01 投资事件 300万美元
投资方：金沙江创投
融资方：北京小桔科技有限公司
07-01 投资事件 数百万人民币
投资方：天使投资人王刚
融资方：北京小桔科技有限公司
北京小桔科技有限公司重庆分公司
滴图（北京）科技有限公司
天使投资人王刚
北京小桔科技有限公司
杭州快智科技有限公司
金沙江创投
滴滴商业服务有限公司
嘉兴小桔子投资合伙企业（有限合伙）

图9-21 时序分析

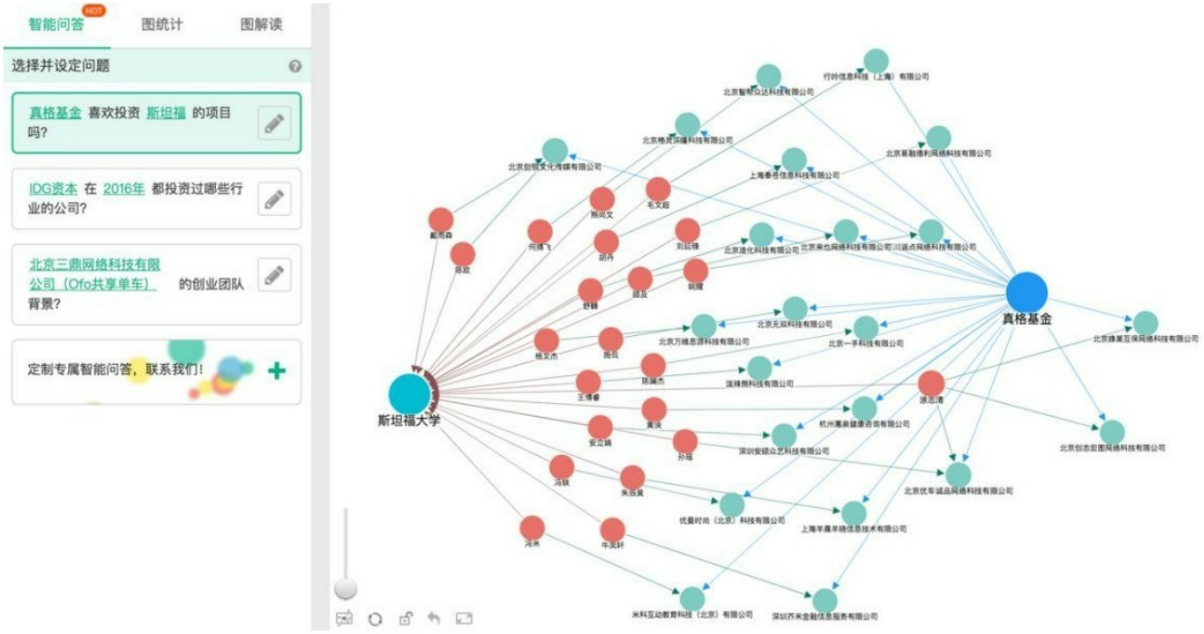

图9-22 自然语言BI

9.3.6 中医临床领域知识图谱的构建与应用[27]

中医药学是一门古老的医学，历代医家在数千年的实践中积累了丰富的临床经验，形成了完整的知识体系，产生了海量的临床文献。利用信息技术手段开展中医临床知识的管理和服务是一项开创性的探索，在临床上具有极大的应用价值。知识图谱有助于实现临床指南、中医医案以及方剂知识等各类知识的关联与整合，挖掘整理中医临证经验与学术思想，实现智能化、个性化的中医药知识服务，因此在中医临床领域具有广阔的应用前景。

中医临床领域有其自身的特点和需求，需要专门研究中医临床知识建模方法，以解决中医临床知识的获取、分类、表达、组织、存储等核心问题。只有采集加工高质量的中医临床知识，才能建立准确、实用、完整的中医临床知识图谱。中国中医科学院中医药信息研究所相关学者以“证、治、效”为中心，对中医临床领域庞大的知识内容进行系统梳理，初步建立了一个中医临床知识图谱系统。该系统以中医临床领域本体作为骨架，集成了名医经验、临床指南、中医医案、中医文献和方剂知识等多种知识资源，并实现了各类知识点之间的知识关联。知识图谱为中医临床知识体系的系统梳理和深度挖掘提供了新颖的方法，有助于实现中医临床知识的关联、整合与可视化，促进中医临床研究，辅助中医临床决策。

中医临床知识是解决中医临床实际过程中特定问题的结合，主要包括：看创指南、名

医经验、临床术语、古籍和期刊文献资源（包括RCT　　文献质量评价结果）、中药方剂等。这些信息分散于不同的组织机构和信息系统之中，形成一个个“知识孤岛”，尚未得到有效整合，严重影响了临床应用的效果。

但通过疾病、症状、方剂、中药等核心概念构成的中医临床知识图谱，可在这些“知识孤岛”之间建立联系，增强中医药知识资源的连通性，面向中医药工作者提供临床知识的完整视图，如图9-23所示为中医临床知识图谱示意图。

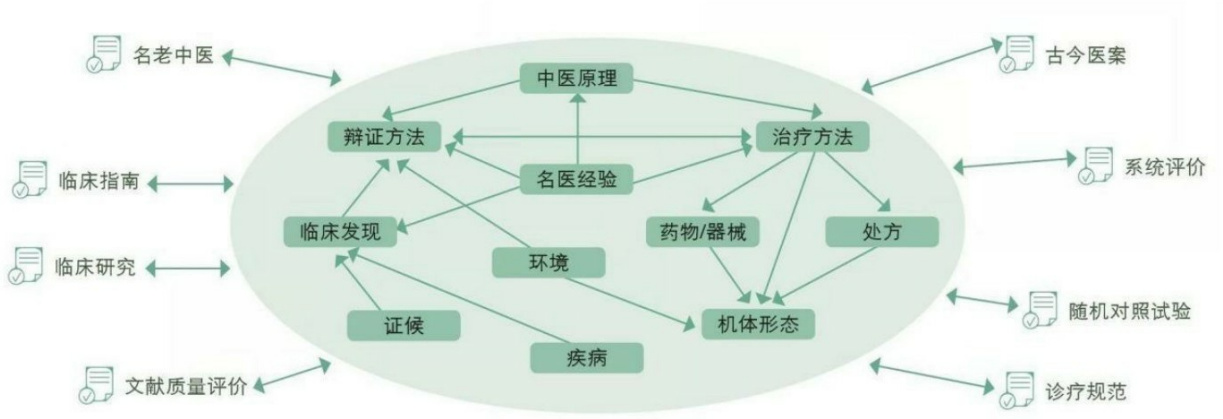

图9-23 中医临床知识图谱示意图

中医临床知识图谱的构建包括以下三个部分：

首先，基于领域专家设计中医临床领域的顶层本体，形成业界公认的技术规范。知识工程师们都可依据该规范进行知识图谱的加工，所产生的知识图谱互相兼容并能最终融合在一起。

其次，构建目标领域的语义网络，作为知识图谱的骨架。例如，中医临床术语系统（Traditional Chinese Medicine Clinical Terminology System,TCMCTS）就是一个专门面向中医临床的大型语义网络，共收录约11万个概念、27万个术语以及100多万条语义关系。[28]在建立语义网络之后，就可以进行领域知识的填充工作了。

最后，从术语系统、数据库和文本等知识源获取知识，对知识图谱内容进行填充。可将本领域中已有的术语系统和数据库的内容转换为知识图谱，从而避免知识资源的重复建设。针对自由文本，可采用自然语言处理和机器学习等方法，从古今中外的各类中医药文献中自动发现实体和语义关系，以自动或半自动的方式填充知识图谱。

在中医临床领域，构建知识图谱的一个核心的知识源是中医医案。中医医案是中医临床思维活动和辨证论治过程的记录，是中医理法方药综合应用的具体反映形式[29]。特别是名老中医的医案，对于中医理论和方法的传承具有重要意义。中医临床知识以医案形式分散于文献之中，这不利于知识检索以及临床数据的分析与挖掘。

从中医医案到知识图谱的知识转换是中医临床知识图谱构建中的核心任务。通过探索医案文本语义分析与知识获取的方法，中国中医科学院中医药信息研究所的学者们研发了中医医案语义分析与挖掘工具，实现医案文本预处理、分词、语义标注、语义检索、医案文本浏览等功能。通过这套工具，从中医古代医案中抽取结构化的中医临床知识，填入中医临床知识图谱。所产生的知识图谱主要包括：名医（如“施今墨”）的擅长疾病、经验方以及弟子等信息；方剂（如“竹叶石膏汤”）的作用、操作方法，以及相关疾病、症状等信息；疾病（如“肺胀”）的临床表现、治疗方法以及相关病症、养生方法、名医等信息；中药（如“杏仁”）所治疗的疾病以及相关方剂、名医等信息。

从知识学的角度分析，中医临床知识从低到高可分为“事实性知识”“概念性知识”“策略性知识”等多个层次。中医医案属于基础性的“事实型知识”，它直接记录中医临床活动中发生的事实。中医临床知识图谱则属于“概念性知识”，它用于梳理概念体系以及表示概念之间的关系。从医案知识向知识图谱的转换过程，实质上是一个知识抽象和归纳的过程。在这个过程中，一方面要完成知识抽取：对海量医案文本进行分析和标注，从中抽取中医知识；另一方面，要实现知识的结构化表示，也就是从医案文本到结构化知识的转换。在最高层则是问题求解和过程控制所需的“策略性知识”（通常用规则、过程等表示），它们是临床决策支持系统的基础。可见，知识图谱处于中间层，在多维度、多层次、多主题的知识点之间建立关联，在中医临床知识系统中起到重要的“粘合剂”作用。

知识图谱有助于对中医临床知识进行分类整理和规范化表达，促进中医临床知识的共享、传播与利用，在临床诊疗、临床研究、教育、培训等方面都具有应用价值。特别是可以将中医临床知识图谱集成到知识服务系统之中，用于改进知识检索、知识问答、决策支持和知识可视化等多种服务的效果，从而提升知识服务能力。如图9-24所示，知识图谱系统以图形化的方式呈现中医名家、疾病、特色疗法、方药、养生方法等概念之间的相互关系，实现中医临床知识体系可视化。系统提供检索框，用于检索知识图谱中的概念。

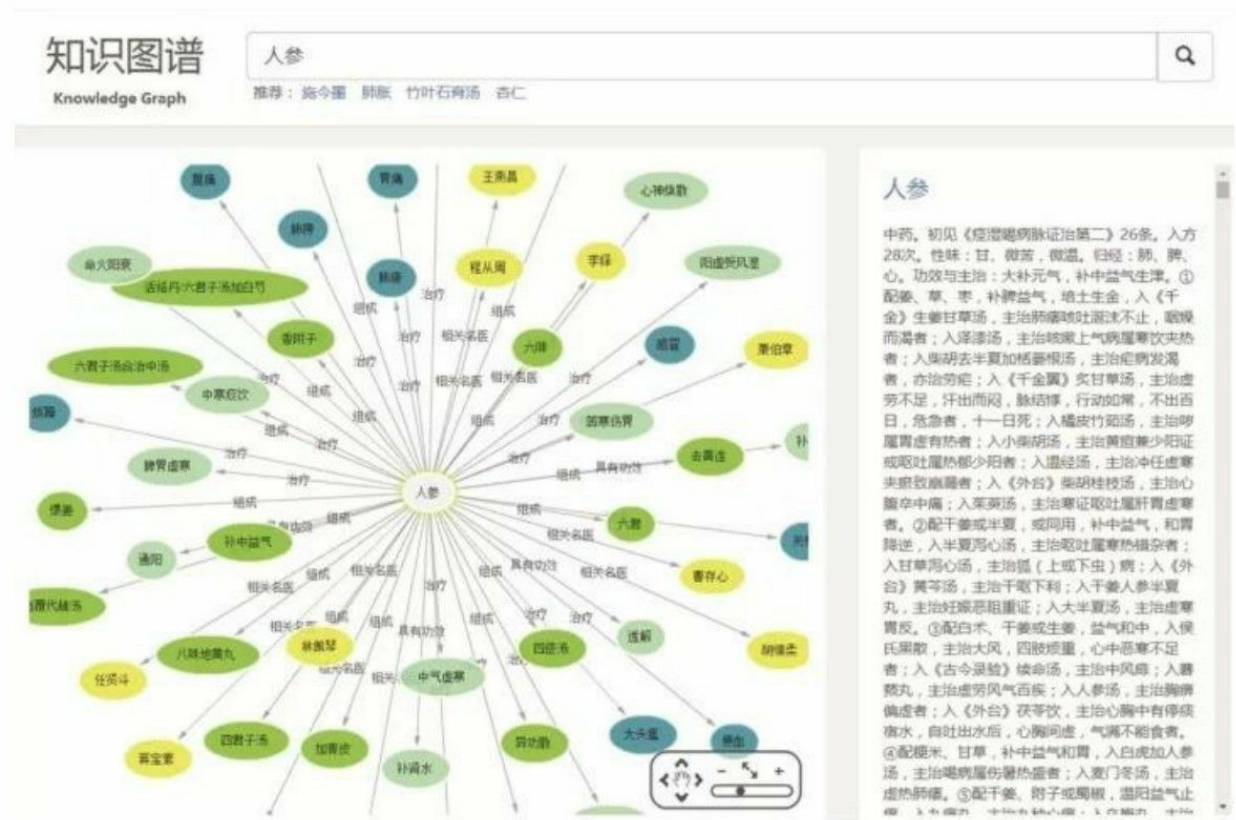

图9-24 中医临床知识图谱界面示意图

知识图谱系统以图形化的方式呈现中医名家、疾病、特色疗法、方药、养生方法等概念之间的相互关系，实现中医临床知识体系可视化。系统提供检索框，用于检索知识图谱中的概念。

使用知识图谱，用户可快速找到与当前研究主题相关的医案、指南和知识库内容，辅助用户进行决策。系统协助用户在概念层次上浏览中医临床知识，发现概念或知识点之间的潜在联系，从而更好地驾驭复杂的中医药知识体系。

中医临床知识图谱分析和揭示“证、治、效”之间的相关关系，提供了新颖的理念和方法。“证、治、效”是中医临床的灵魂，揭示三者之间的关联关系对于提高中医临床疗效具有重要意义。由于中医疗效的判断十分复杂，加入疗效这个因素后，使得三者关系的维度过高，目前的计算机模型很难处理，但可以选择验案作为研究方证对应关系的数据资源，因为验案本身都具有良好疗效。可在验案的基础上构建中医临床知识图谱，全面收集中医临床中与“证、治、效”相关的信息，从而再现中医验案中蕴涵的相关关系（如方剂与证候的相关关系、症状组合与证候的相关关系、药物组合与方剂的相关关系等），揭示症状组合规律、方剂配伍规律以及基于药物组合和症状组合的方证对应规律等。最后，可将这些相关关系和规律提供给临床医生，作为支持临床决策的参考性依据。

知识图谱是在“大数据”时代背景下出现的一项新颖的知识管理技术。在“大数据”时代，不再热衷于寻找因果关系，转而将注意力放在相关关系的发现和使用上。知识图谱从多个维度来描述中医药领域对象，反映中医药事物之间的相关关系，它将是中医药大数据方法学体系中的核心组成部分。大数据通过识别有用的关联关系来分析一个现象，而不是揭示其内部的运作机制。基于相关关系分析的预测是大数据的核心。中医的思想方法不是严格的逻辑推理，而是一种关联式的思考。这种理念上的相似性，使得中医药工作者更易接受并使用“大数据”的方法与技术。利用中医临床知识图谱，能够发现中医药概念之间的相关关系，揭示各种临床规律，从而不断完善中医临床知识体系，直接推动中医临床研究的快速发展。

9.3.7 金融证券行业知识图谱应用实践[30]

金融证券行业正面临着数据爆炸的问题。传统的金融数据服务商历时数十年，已收集整理了大量高质量的结构化数据，并分门别类地展示给用户。如何有效地使用这些数据，需要用户具备专业的金融经济知识，深刻理解某个数据的变动可能引发的关联、传导效应，从而帮助用户做出各种投资决策。金融行业的研究人员相当于在大脑里存储或训练了一个知识图谱，将相关的行业、产品、公司等因素联系在一起，当观察到某个数据变量发生变化时，可以分析推理出各种观点并进行预测。

然而，一个人的脑容量或记忆是有限的，一位专业的行业分析师通常只能对几个行业了如指掌。因此，对市场进行全行业的分析服务需要一支分析师团队。通过人与人之间的交流，以及研报与研报之间的关联和对接，来实现整个经济金融体系的传导与联系。近年来，非结构化数据的井喷式涌现给这种传统的运作方式带来了挑战。财经新闻、经济产业信息每时每秒都在更新；上市公司的数目众多，所涉及的定期报告、临时报告数量巨大；基于互联网平台的股吧、论坛、门户网站、微信、微博等每时每刻也在产生着大量的资讯，上述信息都将可能对证券市场产生各种各样的影响。这使得从海量资讯触发源上，以及分析数据所需的知识的广度、深度上，均对传统的资讯处理模式提出了极大的挑战。

现代信息技术人工智能的发展已经可以在很多方面提高信息分析和利用的效率。对结构化数据的分析挖掘已经取得了很多进展，很多成熟的分析预测算法还是针对结构化、关系数据的。然而，非结构化数据的分析挖掘和利用尚处于起步阶段。领域知识建模在方法论上的正确性，是决定人工智能应用成功与否的关键因素。当前，“知识图谱”作为领域知识建模的工具正在受到越来越多的重视。基于知识图谱的领域建模、基于规模化大数据的处理能力、针对半结构化标签型数据的分析预测算法三者的结合，是人工智能的优势所在。构建金融证券领域知识图谱作为金融证券文本语义理解和知识搜索的关键基础技术，

为未来金融证券领域文本分析、舆情监控、知识发现、模式挖掘、推理决策等提供了坚实支撑。

金融领域的知识图谱与其他专业领域图谱相比有着很大的不同。金融领域本就是连接各行各业、世间万物的，因此金融知识图谱涉及经济、投资、产业、公司等相关的知识，其实是覆盖全行业的。但金融领域知识图谱与通用或百科类知识图谱不同，其行业、产业链知识，经济金融重要指标等大多是以投资的视角来筛选和组织的。

金融知识图谱常见的实体包括：公司、产品、证券、人等。实体间的关系，如公司-人之间，主要有股权关系和任职关系；公司-公司间关系，有股权关系、供应商关系、竞争关系等；公司-产品间关系，有生产关系、采购关系等；产品-产品间关系，主要有上下游关系等。有些实体和关系可以自动抽取生成，如公司-公司间的股权关系、公司-人之间的股权关系和任职关系，均可来源于工商局注册登记公开信息，其结构化程度很高，实体、关系抽取难度不大。而对于产品-产品间上下游关系，则很难有系统性的半结构化数据源，其实体和关系呈碎片化分散在百科类网站、研究报告、专家资料等文本或图像中，这给抽取和甄别带来了很大挑战。如图9-25所示为金融知识图谱示例。

金融知识图谱的建立可以分为以下三个部分：从海量异构非结构化数据中辨别金融实体；定义并挖掘金融实体之间的各种关系，从而生成知识图谱；定义并表达业务逻辑，在知识图谱上实现各种具体任务，如推理等。

本书对构建过程中主要用的关键技术进行简单的梳理：

1.实体-关系抽取

从海量异构非结构化数据中辨别金融实体，主要采用实体-关系抽取技术，即从文本中抽取出特定的实体信息，如时间、人物、地点、公司、产品等；以及实体间的各种关系，如地理位置关系、雇佣关系、股权关系等。实体确定了知识图谱中的点，而关系则确定了点与点之间的边。

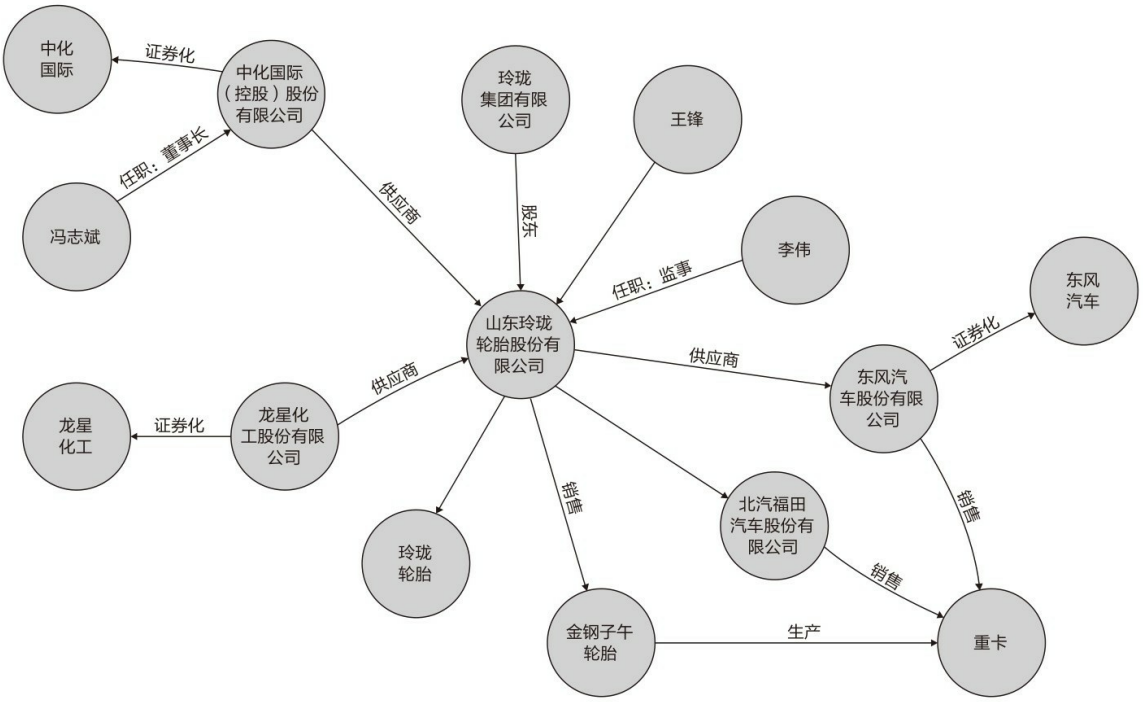

图9-25 金融知识图谱示例

常用的实体关系抽取方法有基于专家知识库的方法和基于机器学习的方法等。基于专家知识库的方法需要专家构筑大规模的领域知识库，需要大量的专家劳动。机器学习方法需要构造特征向量形式的训练数据，然后使用各种机器学习算法，如支持向量机等作为学习机构造分类器。这种方法被称作基于特征向量的学习算法。

通常来说，构造领域知识图谱会从大量特定类型的文本（尤其是高质量、模板化的专业资料）中提取实体关系。这类文本，或者是半结构化，或者是模块格式相对明确固定的，例如上市公司公告的 XBRL 格式数据。这类规范化数据源降低了信息提取的难度，大大提高了知识提取的准确度和效率。对于非结构化文本，实体识别和关系抽取需要基于NLP算法，以及深度学习算法（例如，用词向量的方式寻找近义词，提高实体模糊识别的准确度），是一个反复迭代、不断精进的过程。其中，关系抽取可以划分为确定类型的关系抽取和不确定类型的关系抽取。确定类型的关系抽取，例如“is-a”关系，可使用语法模式抽取固定模式，使用迭代方法扩展“is-a”关系，并对生成的“is-a”进行清洗。不确定类型的关系抽取常基于 NLP 将目标实体间的谓词提取出来作为候选关系，再进行下一步的筛选鉴别。

2.定义并挖掘金融实体间的各种关系，从而生成知识图谱

基于领域知识图谱的推理与业务场景息息相关。基于通用知识图谱的推理沿边的传递性并不强，例如精准搜索常常只用到一步到二步的推理，再往下传递时，其可信程度将会

大大降低。而金融知识图谱在与领域知识充分结合的前提下，是可以实现长链推理的。下面列举几个推理案例：

（1）关联关系推理。基于知识图谱中公司和人之间的股东、任职等关系，可以基于聚类算法发现利益相关团体。此时，当其中若干节点发生变动或大的事件时，可以通过沿知识图谱路径查询或子图发现等方法计算并绘制发生变动的实体间的关联情况，帮助监管层发掘潜在的关联或违规行为，大大提高关联发现的效率。如图9-26所示，该图为一个以某公司为核心的股权关系结构图，当该公司出现异常风险时，会影响到其核心关联节点。

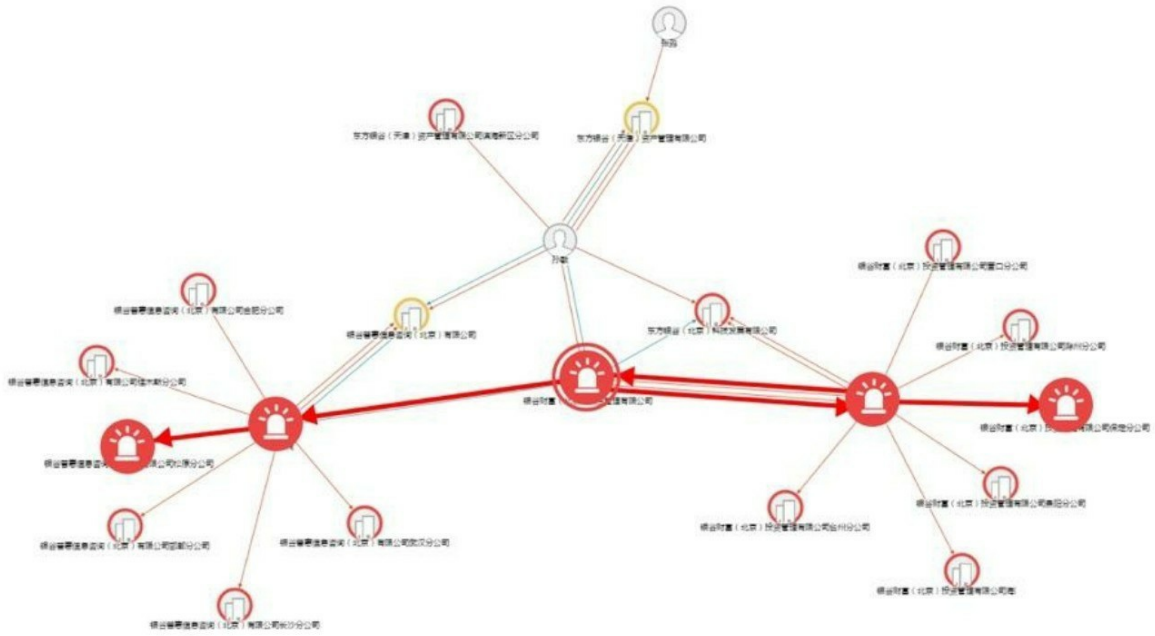

图9-26 关联关系推理示意图

（2）产业链关系推理。基于产业链知识图谱，可模拟经济学的涟漪效应：某产业链下游销量大涨，对整个产业链中游、上游的拉动是非常显著的，且可以沿图谱用量化的方式建模并形成自动化推理传导模型。同样的，上游原材料成本的上涨对于产业链中下游也可能形成链状的传导效应。这将帮助判断事件的重要程度，并即时给出事件的影响范围和程度，为各类投资决策做数据支持。如图9-27所示，某一稀有原材料上涨，其产业链中下游的产品可能因为成本上涨导致产品价格上涨等。

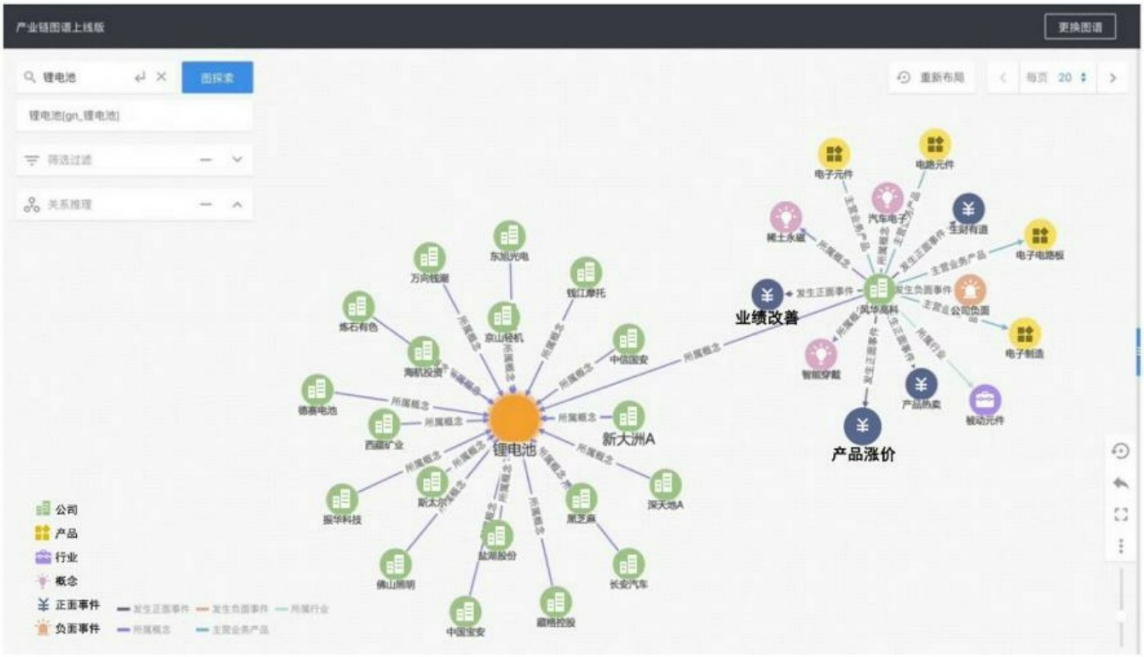

图9-27 产业链关系推理示意图

3.领域知识图谱数据库选型

构建领域知识图谱底层数据库有非常多的选择。从传统的关系数据库到 NoSQL，再到图数据库；不论是采用一种数据库还是多种数据库相结合的方式，都是研发领域知识图谱前需要反复斟酌和考虑的问题。数据库的选型需要充分考虑领域数据自身的特点（以结构化数据为主，还是非结构化数据为主），以及如何使用这些数据（例如，是否经常需要沿图谱进行推理，推理路径长短等）。通常来说，Neo4j 等图数据库擅长长链推理，但对单位基础数据的日常维护较弱；MongoDB、HBase 等 NoSQL数据库擅长处理文本类非结构化数据，对于传统数值型数据的很多处理则需要额外写代码维护；MySQL 等传统数据库擅长处理和维护结构化数据，在面对沿图谱进行推理等应用时则需要比图数据库更多的代码量。

从工程实现上来看，图数据库的使用频率和相关人才储备远低于关系数据库，如果选用图数据库作为主要的底层数据库，研发团队可能经常需要面临无人可招和遇到问题搜遍网络都无帖可解的窘境，即整个系统工期规划会难以预估。

构筑金融领域的知识图谱是一个既有着大量结构化数据，又需要整合非结构化文本数据信息，同时需要沿图谱进行推理的综合性项目。传统的金融数据供应商长期积累了大量结构化数据，例如价格、营收、利润、销量等数据，均为长时期时间序列格式。这与通用知识图谱相比，呈现出很大的不同。因此，在具体的数据库选型时，需要充分考虑未来的应用将以何种方式、何种频率使用数据，从而打造出因地制宜的高效底层数据库。

在知识图谱在金融证券行业应用方面，目前国内尚处于起步阶段。如果能基于知识图

谱技术框架，建立起一个全谱系的上市公司关联图，并将其直接关联、间接关联的各种实体、概念相联系，将极大地帮助证券行业监管层、投资者及其他各种参与者了解并把握市场的脉搏。而在具体业务应用方面，当监控到市场价格出现波动时，可以就股价出现异动的股票在知识图谱中追溯其异动产生的根源；挖掘学习实体之间的隐含关系，来发现潜在的关联与协同动作，以预防并打击违法、违规行为；自动学习并抽取公告摘要，快速传递并汇总全市场披露的动态信息，以减少信息不对称性并加强证券市场的透明度。

基于金融证券知识图谱可在多个智能金融应用场景中得到应用，这些应用场景包括：智能投研、智能投顾、智能风控、智能客服、智能监管、智能运营等。

智能投研专注于对基本面等信息的采集和分析。对智能投研技术的实用化来说，自然语言处理和产业链、作用链的知识图谱建模是最关键的技术。具体而言，通过构造上下游产业链知识图谱，基于经济基本面建立传导模型。当产业链中重要节点的状态发生变化时，将启动沿产业链传导推理引擎，自动给出影响范围、对象和程度，为事件引发的基本面分析做支持。不同于技术分析，基本面分析本身是一个非结构化的方式，无论是数据，还是市场逻辑。基于金融知识图谱和推理逻辑，把这些基础数据进行整合加工，从而找到未来趋势的变化或者解释已经发生过的事情。从局部来看，产业链知识图谱里面各种实体、属性、关系就像活细胞一样，相互关联、影响、作用着。这是“金融知识图谱+推理链”的共同作用结果。如图9-28所示为橡胶-轮胎-重卡产业链知识图谱局部示意图。当发生“重卡销量大增”事件时，可沿产业链向上游进行传导推理，并生成分析影响报告。

基于金融知识图谱，还可在风险评估与反欺诈方面展开应用。风险评估是大数据、互联网时代的传统应用场景，应用时间较早，应用行业广泛。它是通过大数据、机器学习技术分析用户行为数据后，进行用户画像，并进行信用评估和风险评估。

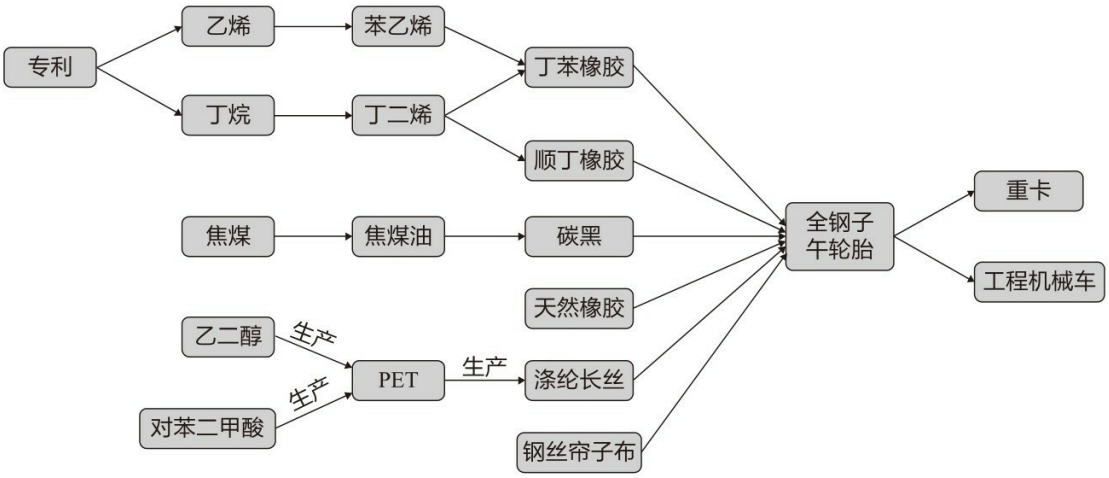

图9-28 橡胶-轮胎-重卡产业链知识图谱局部示例

NLP技术在风控场景中的作用是理解分析相关文本内容，为待评估对象打标签，为风控模型增加更多的评估因子。引入知识图谱技术后，可以通过人员关系图谱的分析，发现人员关系的不一致性或者短时间内变动较大，从而侦测欺诈行为。利用大数据风控技术，在事前能够预警，过滤掉带恶意欺诈目的人群；在事中进行监控，及时发现欺诈攻击；在事后进行分析，挖掘欺诈者的关联信息，降低以后的风险。

在金融行业中，风险评估与反欺诈的应用场景首先是智能风控。利用 NLP 和知识图谱技术改善风险模型以减少模型风险，提高欺诈监测能力。其次，还可以应用在智能监管领域，以加强监管者和各部门的信息交流，跟踪合规需求变化。通过对通信、邮件、会议记录、电话的文本进行分析，发现不一致和欺诈文本。例如，欺诈文本有些固定模式：如用负面情感词，减少第一人称使用等。通过有效的数据聚合分析，可大大减少风险报告和审计过程的资源成本。从事此类业务的金融科技公司很多，如 Palantir 最初从事的金融业务就是反欺诈。其他如Digital Reasoning、Rapid Miner、Lexalytics、Prattle等。

另一方面，金融知识图谱还可以用在客户洞察方面。客户关系管理（CRM）也是在互联网和大数据时代中发展起来，市场相对成熟，应用比较广泛，许多金融科技公司都以此为主要业务方向。现代交易越来越多是在线上而不是线下当面完成，因此如何掌握客户兴趣和客户情绪，越来越需要通过分析客户行为数据来完成。

NLP技术在客户关系管理中的应用，是通过把客户的文本类数据（客服反馈信息、社交媒体上的客户评价、客户调查反馈等）解析文本语义内涵，打上客户标签，建立用户画像。同时，结合知识图谱技术，通过建立客户关系图谱，以获得更好的客户洞察。这包括客户兴趣洞察（产品兴趣），以进行个性化产品推荐、精准营销等，以及客户态度洞察（对公司和服务满意度、改进意见等），以快速响应客户问题，改善客户体验，加强客户联系，提高客户忠诚度。客户洞察在金融行业的应用场景主要包括智能客服和智能运营。例如在智能客服中，通过客户洞察分析，可以改善客户服务质量，实现智能质检。在智能运营（智能 CRM）中，根据客户兴趣洞察，实现个性化精准营销。国外从事这个业务方向的金融科技公司有Inmoment、Medallia、NetBase等。[31]

总体来说，基于金融知识图谱的应用，有如下三大特点：

（1）广覆盖。广泛覆盖全量信息源，覆盖宏观、中观、微观各维度信息，覆盖上市公司及非上市公司，以方便后续算法拓展所有可能的深度关联关系。

（2）深加工。基于知识图谱与智能推理链，实现从数据到智慧的深加工。

（3）浅表达。以可视化的方式和自然语言与用户交互，一目了然，受众更广。

然而，领域知识图谱对专业知识的基础需求，远远大于通用知识图谱。在建设初期需要大量的专家工作。基于此，可以尝试从两个方面入手来构筑大型领域知识图谱。

一方面，开启知识众包时代，建立新的协作方式。构建用户友好的知识众包协作平

台，使得专家能很方便地利用碎片化时间在平台上贡献自己的知识，同时设计相应的知识回报模式。就平台自身而言，如何设计自动内容校验和精华内容提取算法，从大量专家碎片化知识中提取重要内容以添加到“主图谱”中，是一个需要长期不断探索的课题。

另一方面，通过知识自动抽取、自动生长构建“活”的知识图谱。这意味着需要有新的知识持续不断地输入知识图谱中；通过知识图谱定义的作用链进行自动推理；知识图谱自身可以备靠大数据，在“人工+自动”模式下自我生长。

通过这两方面的相辅相成、交叉验证，以真正将海量非结构化信息自动化利用起来，成为领域应用决策的坚实支撑。

9.4 本章小结

结合知识图谱研究发展态势，并结合当前知识图谱的构建与应用未来现状，对知识图谱未来技术发展及趋势发展做一个展望。

1.知识图谱构建

现阶段，基于本体工程的知识描述和表示仍是知识图谱建模的主流方法，而且仅仅用到一些RDFS及OWL中定义的基础元属性来完成知识图谱模式层构建，知识图谱所关注的重点也仍然是数据中的概念、实体、属性等。随着人们对知识的认知层次的提升，势必会对现有的知识表示方法进行扩展，逐步扩充对于时序知识、空间知识[22]、事件知识[23]等的表示。而知识图谱本身也会逐步将关注重点转移到时序、位置、事件等动态知识中去，来更有效地描述事物发展的变化，为预测类的应用形态提供支持。

其次，对于知识图谱构建任务来说，最困难、最无法标准化实现的一个环节就是对于文本数据的信息抽取。知识图谱面向开放领域的信息抽取普遍存在着召回率低、算法准确性低、限制条件多、拓展性差等问题。随着计算机计算能力的日益提高与深度学习技术不断研究发展，NLP 领域发生了翻天覆地的变化，CNN、RNN 等经典神经网络结构已经被应用于 NLP 中，尝试完成机器翻译、命名实体识别任务。未来，深度学习的思想和方法会越来越多地应用于文本信息抽取中，优化抽取方式，提高知识的覆盖率与准确率[24-25]。其他如跨语言知识融合[26]、知识嵌入[27]等方向也会在深度学习技术的加持下激起新的研究浪潮。

2.知识图谱应用

在知识图谱应用方面，未来将会出现更多的应用形态，如基于知识图谱的智能文本编制，通过知识图谱将行业中的业务知识与文档相结合，在文档编制过程中进行实时的智能提示、知识校验、知识生产等，辅助文档编制。又如基于知识图谱的自然语言理解与自然语言生成，通过知识图谱对知识的建模能力，结合深度学习对知识的学习与抽象能力，实现以自然语言形式进行输入和输出的下一代问答系统。随着知识表示技术和推理技术的发展，结合一些新型的可视化方法，还可以展望一些预测分析类的应用形态，如疾病预测、行情预测、政治意识形态检测[28]、城市人流动线分析[22]。除此之外，知识图谱在辅助多媒体数据处理方面也是一个有待深入研究的方向，如物体检测[29]、图像理解[30]等。

总之，知识图谱作为人工智能技术中的知识容器和孵化器，会对未来 AI 领域的发展起到关键性的作用。无论是通用知识图谱还是领域知识图谱，其构建技术的发展和对应用

场景的探索仍然会不断地持续下去。知识图谱技术不单指某一项具体的技术，而是知识表示、抽取、存储、计算、应用等一系列技术的集合。随着这些相关技术的发展，我们有理由相信，知识图谱构建技术会朝着越来越自动化方向前进，同时知识图谱也会在越来越多的领域找到能够真正落地的应用场景，在各行各业中解放生产力，助力业务转型。