知识图谱

发展报告（2022）

KNOWLEDGE GRAPH DEVELOPMENT REPORT

中国中文信息学会
语言与知识计算专委会

中国 北京

2022.08

目录

序言

当前人工智能正在经历从感知智能到认知智能的重要发展阶段。认知是人们获取和应用知识的过程，因此，作为人类对客观世界认知的一种表现形式，知识图谱是认知智能研究不可或缺的组成部分。知识图谱可以帮助机器积累人在解决问题中使用的知识，可以帮助组织互联网资源，进而用知识赋能行业智能应用，知识图谱及其知识引擎技术已经成为人工智能系统的基础设施。《知识图谱发展报告》（2022）是中国中文信息学会语言与知识计算专委会邀请知识图谱领域专家结合人工智能和知识图谱技术的最新发展，在《知识图谱发展报告》（2018）基础上对本方向前沿技术和应用的又一次系统总结，并对未来前沿趋势进行展望。

近年来，随着人工智能特别是大数据、深度学习和大规模预训练模型的快速发展，知识图谱的理论、方法和应用也有了很大进展。

在知识表示和建模中，知识图谱表示形式更加多样化，从单一语言和符号表示的知识图谱，到多语言和多模态的知识图谱；从结构化知识表示发展到与半结构化和非结构化数据融合的概念-实体-上下文一体化知识表示，从符号知识表示到融合符号和数值的知识表示。

在知识获取方面，低资源、真实场景下的知识获取技术也有了长足进步，由传统限定领域的知识抽取，到如今开放领域的多类别知识抽取；由基于知识库的关系获取，到以知识为指导的面向大规模预训练技术的关系获取；由粗粒度有监督学习到细粒度小样本学习，以及由单一模态的概念抽取到跨模态的联合学习。

在知识图谱应用方面，知识图谱领域应用越来越广泛，以多模态知识为驱动的虚拟数字人推动着人工智能走向更广阔的应用场景，“知识图谱+产业”的新范式凸显着以知识为中心的应用与现实业务的深度融合。“知识图谱+其他学科（如区块链、物联网）”的交叉研究也正在兴起和发展。

知识图谱未来发展趋势和面临的挑战在于，能否利用大规模预训练模型进一步促进知识表示、获取和推理技术的发展，能否基于认知推理实现具有认知能力的人工智能新架构，能否利用知识的可解释性释放更多产业潜能和应用。

本发展报告的定位是深度科普，旨在向政府、企业、媒体中对知识图谱感兴趣的社会各界人士简要介绍相关领域的基本概念、基本方法和应用方向，向高等院校、科研院所和高新技术企业中从事相关工作的专业人士介绍相关领域的前沿技术和发展趋势。

本报告共由 13 章组成，每一章按照 1）任务定义、目标和研究意义；2）研究内容和关键科学问题；3）技术方法和研究现状；4）技术展望与发展趋势等四部分的结构形成每一章

的内容。每一章我们邀请了本专业领域内的专家协同撰写完成。具体结构如下：

- 序言：李涓子（清华大学）、赵军（中国科学院自动化研究所）
- 知识表示与建模：张文，耿玉霞，许泽众，陈华钧（浙江大学）
- 知识表示学习：刘知远、汪华东（清华大学）
- 实体抽取：林鸿宇、韩先培（中国科学院软件研究所）
- 实体关系抽取：曾道建（湖南师范大学）、陈玉博、刘康（中国科学院自动化研究所）
- 事件知识获取：丁效（哈尔滨工业大学）
- 知识融合：胡伟（南京大学）、漆桂林（东南大学）
- 知识推理：张小旺（天津大学）、李炜卓（南京邮电大学）、张文（浙江大学）、漆桂林（东南大学）
- 知识图谱的存储和查询：彭鹏（湖南大学）
- 通用和领域知识资源：王昊奋（同济大学）、曹征晖（复旦大学）、林俊宇（中国科学院信息工程研究所）
- 知识图谱质量评估与管理：李直旭（复旦大学）、王萌（东南大学）、漆桂林（东南大学）、阮彤（华东理工大学）
- 基于知识的问答与对话：何世柱、张元哲、刘康（中国科学院自动化研究所）
- 基于知识的搜索与推荐：程龚（南京大学）
- 知识图谱交叉前沿：张文、毕祯，朱渝珊，李娟，陈卓，陈华钧（浙江大学）

发展报告最后由刘康（中国科学院自动化研究所）、程龚（南京大学）、侯磊（清华大学）、张元哲（中国科学院自动化研究所）、吴天星（东南大学）、陆垚杰（中国科学院软件研究所）等根据反馈意见对初稿进行校对并统一成文。

由于时间仓促，本报告难免有疏漏甚至错误的地方，仅供有志于语言与知识计算研究和开发的同仁参考，并激发更广泛的思考和讨论。期待在我们的共同努力下，知识图谱以及语义计算技术能够取得更辉煌的成果。

李涓子（清华大学）、赵军（中国科学院自动化研究所）

2022 年 8 月

第一章 知识表示与建模

张文[2]，耿玉霞[1]，许泽众[1]，陈华钧[1]

1. 浙江大学 计算机科学与技术学院，浙江省 杭州市 310007；

2. 浙江大学 软件学院，浙江省 宁波市 315048

一、任务定义、目标和研究意义

知识是人类通过观察、学习和思考有关客观世界的各种现象而获得和总结出的被广泛论证的正确的信息，知识具有三大特点：合理（Justified）、真实（True）和被相信（Believed）。在人类社会中，知识表示将人类的认知知识以特定的形式进行描述、表达和传承，人类表示知识的形式多种多样，包括声音、文字、绘画、音乐、数学语言、物理模型以及化学公式等，这些丰富的知识表示方法让人类更准确地表达自己的认知，有力地促进了社会文明进步。

对于机器而言，知识表示（Knowledge Representation, KR）将现实世界中的各类知识表达成计算机可存储和可计算的结构，使得计算机可以无障碍地理解所存储的知识。上世纪 90 年代，MIT AI 实验室的 R. Davis 定义了知识表示的五大特点：

- 客观事物的机器标识（A KR is a surrogate），即知识表示首先需要定义客观实体的机器指代或指称。
- 一组本体约定和概念模型（A KR is a Set of ontological commitments），即知识表示还需要定义用于描述客观事物的概念和类别体系。
- 支持推理的表示基础（A KR is a Theory of Intelligent Reasoning），即知识表示还需要提供机器推理的模型与方法。
- 用于高效计算的数据结构（A KR is a medium of efficient computation），即知识表示也是一种用于高效计算的数据结构。
- 人可理解的机器语言（A KR is a medium of human expression），即知识表示需要接近人的认知，是人可理解的机器语言。

自人工智能提出至今，知识表示已经探索过语义网络、专家系统、语义网、知识图谱等形态，形成了基于框架的语言、产生式规则、RDF 以及 OWL 等知识表示语言。近年来，人工智能依靠机器学习技术的进步，在数据感知方面取得了巨大的进步，可以精准地完成图像识别、语音识别等任务。但当前人工智能在语言理解、视觉场景分析、决策分析等方面依然面临巨大的挑战，其中一个关键挑战便是如何让机器掌握大量的知识，尤其是常识知识，这体现了知识表示的重要性。

二、研究内容与关键科学问题

根据知识呈现的形态和方式，我们可以将知识分为不同的类型，包括本体知识、规则知识以及事件知识等。其中本体知识表达实体和关系的语义层次，用于建模领域的概念模型；规则知识表达实体和关系之间存在的推理规律，是更抽象的知识；事件知识包含多种事件要素，是更多维更复杂的知识。本章主要针对本体知识和规则知识的表示与建模展开三个方面的介绍，包括当前建模语言、建模工具以及应用实践示例。事件知识相关内容参见第五章。

1. 本体知识

在万维网中，我们可能会用不同的术语来表达相同的含义，或者一个术语有多个含义。因此，消除术语差异是很有必要的。目前较受欢迎的解决方案就是，对某个领域建立一个公共的本体，鼓励大家在涉及该领域的时候都使用公共本体里的术语和规则。

本体最先是哲学领域提出的研究概念，其作用主要是为了更好地对客观事物进行系统性的描述，即总结、提炼描述对象的共性，从而将客观事物抽象为系统化、规范化的概念或专业术语。概括而言，哲学本体关心的是客观事物的抽象本质。应用至计算机领域，本体可以在语义层次上描述知识，因此可以用于建立某个领域知识的通用概念模型，即定义组成“主题领域”的词汇表的“基本术语”及其“关系”，以及结合这些术语和关系来定义词汇表外延的“规则” [Neches et al., 1991]。

具体来说，“领域”是指一个本体，描述的是一个特定的领域，如“大学”、“公司”等；“术语”指给定领域中的重要概念，例如大学这一领域中涉及的有教工、学生、课程等概念；“基本术语之间的关系”包括类的层次结构（类比 taxonomy）等关系，比如大学师生员工中包含了教工和学生，学生又可分为本科生和研究生，教工同理，而学生和教工是两个并列的概念，该层次关系如下图 1 所示；“词汇表外延的规则”，则类似数据库中的“约束”，包括概念的属性约束（如 X 教 Y）、值约束（如只有教职人员才能授课）、不相交描述（如教职人员和普通员工不相交）以及对象间的逻辑关系规定（如一个系至少要有 10 个教职员工）等。

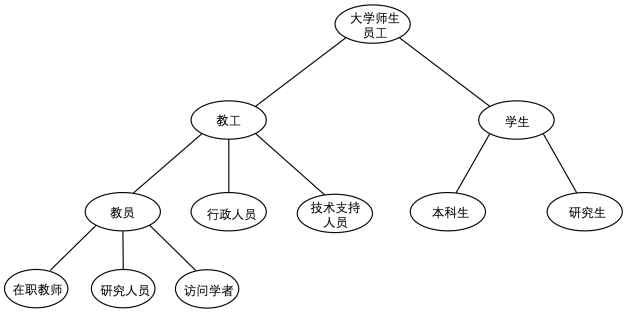


图 1 学校领域概念及概念间层次关系

通过对事物所具有的概念、概念的关系、概念的属性及概念的约束等明确、清晰地描述，本体体现了客观事物内在、外在的关系。从上述本体的定义中，我们可以看出本体四个重要的特点，即概念化、明确性、形式化和共享性。概念化是说本体表示的是各种客观存在的抽象模型，它并不描绘实体的具体形象而是表达出一个抽象的本质概念；明确性主要体现在描述客观事物时，利用自身概念化的表述优势和系统化的思想，准确地展示描述对象的特征；形式化则侧重使用特定的、严格规范化的、无歧义的语言对客观事物进行描述，以达到明确清晰的目的；共享性则是指本体所描述和表达的知识信息具有共享特性，希望能够被用户普遍认同并使用。

而本体与知识图谱之间又有着什么样的联系呢？从逻辑结构上看，知识图谱一般可分为两层，数据层存储知识图谱中的所有三元组信息，模式层（也称 schema 层或本体层）位于数据层之上，对数据层知识结构进行提炼，即通过在模式层上建立约束和规则，可规范图谱中的实体、关系、实体属性、属性值之间的联系，以及完成在知识图谱上的推理。基于知识图谱，本体既可以以模式层的形式出现，表达数据层的抽象知识，也可以以数据层的形式出现，表达资源之间的约束关系，尤其是层次约束关系。

2. 规则知识

传统知识推理历史悠久，相对完备，其理论支持也比较完备，其所基于前提和规则更容易被理解，具有较好的解释性。其中，规则是传统推理中一种重要的方式，一般而言，知识图谱中的规则被表示为以下形式：

$$head \leftarrow body$$

其中，body 表示规则的主体，head 表示规则的头部，一条规则被表示为由主体推导出头部。规则头由一个二元的原子构成，而规则的主体则由一个或者多个一元原子或者二元原子所构成。原子就是包含了变量的三元组，其本身也有肯定和否定之分。如果主体中仅仅包含肯定的原子，那么这样的规则也可以被称之为霍恩规则。

对于规则，其质量评价方法一般包括三种，分别为支持度（support），置信度（confidence），规则头覆盖度（head coverage）。支持度表示满足规则主体和规则头的实例的个数，即该规则在知识图谱中成立的实例数；置信度为满足规则主体的实例的个数和支持度的比值；规则头覆盖度即满足规则头部的实例数量和支持度的比值。基于以上指标，可以对规则的质量有一个比较直观的判断。

作为一种抽象知识，规则的典型应用是根据给定的一套规则，通过实际情况得出结论。这个结论可能是某种静态结果，也可能是需要执行的一组操作。应用规则的过程称为推理。如果一个程序处理推理过程，则该程序称为推理引擎。推理引擎是专家系统的核心模块。其

中，有一种推理引擎以规则知识为基础进行推理，其具有易于理解、易于获取、易于管理的特点，这样的推理引擎被称为“规则引擎”。

三、技术方法和研究现状

1. 本体知识建模

1） 本体知识建模语言

本体构建之前，需要选择合适的本体描述语言。本体描述语言是本体构建环节中的重要工具，客观的信息资源只有经过本体语言的描述转化后才能够在计算机、网络上实现输入、导出、分类、语义关联、逻辑推理等一系列的功能。按复杂程度递进，目前比较有代表性的本体描述语言有：XML、RDF、RDFS 和 OWL。

XML（Extensible Markup Language），可扩展标记语言，是 W3C 组织创建的一种定义标记的通用元语言，它向用户提供统一的框架，以便在不同应用之间交换数据和元数据。它能自定义和为其他语言提供语法支持，XML 数据表示形式简单，无任何语义约束，能够轻易的读写。XML 可用于数据存储、编码、交换和数据分析、处理等方面，可用于 Web 服务、语义网构建等，可支持基于 XML 语言的开发，可用于通信协议、办公软件开发。

RDF（Resource Description Framework），资源描述框架，是 W3C 组织制定的第一个用于对任意资源进行语义信息描述的语言。RDF 即描述对象（“资源”）和对象间关系的数据模型，并为这种数据模型提供一个简单的语义。它由一系列陈述（statement）——即“对象-属性-值”三元组（object-attribute-value triple)组成，我们熟悉的知识图谱常使用这种三元组表示形式。

RDFS（RDF Schema）主要用于描述 RDF 词表，刻画 RDF 资源的属性和类的词汇描述语言。RDFS 定义 RDF 数据模型所使用的词汇，规定什么属性可以作用于什么类型的对象，属性可以取什么值，也可以描述对象之间的关系。从语义网的观点来看，RDFS 使机器可以解读语义信息。

OWL（Web Ontology Language），网络本体语言，是 W3C 组织推出的新的本体语言标准，相比于 XML、RDF、RDFS 增加了更多描述属性和类的词汇，支持基于描述逻辑的推理过程。OWL 提供了 3 个表达能力不同、计算效率各异的子语言：OWL Lite、OWL DL 和 OWL Full。OWL Lite 主要面向需要构建分类层次和约束简单的本体用户；OWL DL 主要提供给需要构建最强表达能力且保持计算的完整性和可判性的用户；OWL Full 主要提供给追求最强表达能力和完全自由的 RDF 语法的使用者。

2）本体知识建模工具

本体建模工具有 Protégé、Apollo、OntoStudio、TopBraid Composer、Semantic Turkey、Knoodl、Chimaera、OliEd、WebODE、Kmgen 和 DOME。其中，Protégé 是大众最熟悉最常用的一个工具，最早开发于 1987 年，主要使用 OWL 语言对知识进行表示，其最初目的是通过减少知识工程师的手动操作来消除知识建模的瓶颈，经过若干次的版本迭代，逐渐演化成了现在的基于框架的本体编辑建模工具。Protégé 可以用于概念建模、实体编辑、模型处理以及模型交换等。Protégé 支持用户界面和 OWL 语言两种方式进行本体建模工作。

在实际应用中，本体建模工具的选取建议关注以下特性：

- 该工具是否拥有可视化用户界面。可视化界面往往会使本体构建简单很多，用户通过对可视化界面进行操作，而无须掌握复杂的程序语言即可构建本体。
- 该工具是否支持分布式构建和存储。考虑到一些承载海量领域知识的大型本体的建模需求，本体的构建会消耗大量的存储空间。这时分布式的存储能力显得尤为重要。同时，对于需要异地协同工作来创建本体的场景，支持分布式构建尤为重要。
- 该工具是否支持推理。本体构建完毕后，应该如何验证本体的正确性？特别是面对包含庞大数据规则的本体，逐一对数据进行验证是不现实的。而基于本体进行推理则是一种常用的方法，即按照指定的规则在本体中进行推理，若推理可以得到正确的结果，则相当于对本体中结构和数据的正确性进行了验证。
- 该工具是否被持续维护。在使用本体建模工具构建本体时，可能会遇到各种各样的问题，工具的开发者无法确保仅使用一个版本就满足用户的所有需求且不会出现任何漏洞。同时，系统环境和需求的不断变化也对本体建模工具的兼容性和功能不断提出新的要求。选择一个有开发者持续维护甚至有完整成熟社区的本体建模工具尤为重要，这将决定使用者是否可以专注于本体的构建而不是不断对工具进行调整，进而提升本体建模的效率。

3）本体知识建模的应用实践示例

本体可用于网站的组织和导航。如网站页面左边中往往会列出在概念层次结构中最高层的术语，用户可以点击其中之一来浏览相关子目录。

本体可用于提高网络搜索的精确度及支持特殊查询。搜索引擎可以根据本体中的搜索关键词来查找相关的“概念”，而不是硬匹配“关键词”，同时可以根据语义层面的相关性消除术语差别。同时，本体可支持网络搜索中不同粒度级别的查询，如果查询失败，搜索引擎可以向用户推荐更一般（即粒度更粗）的查询（或搜索引擎主动执行这样的查询）；反之，如果查询结果过多，搜索引擎可以建议用户使用更特殊（即粒度更细）的查询。

基于本体推理的中医药诊疗实践[陈华钧 et al., 2011]。传统的中医学以阴阳五行作为理

论基础，将人体看成是气、形、神的统一体，通过望、闻、问、切，四诊合参的方法，探求病因、病性、病位、分析病机及人体内五脏六腑、经络关节、气血津液的变化、判断邪正消长，进而得到病名，归纳出证型，以辨证论治原则。基于本体推理的中医药五行诊疗系统，将抽象晦涩的中医五行理论构建成结构化的语义本体，有助于促进中医药理论知识的形式化表达；以语义规则的形式表现中医五行理论中的生克乘侮关系及病机推理相关的逻辑关系，是一种新的表达方法，能有效、直观地表达中医五行的内部机制，同时，结合 flex 技术展示中医五行的诊疗过程，有助于帮助普通用户理解中医理论的科学性。

基于本体建模对化工生产过程进行控制[荣冈 et al., 2015]。现代工业中，由于生产过程通常都比较复杂，在生产进行之前一般会进行建模仿真，以指导生产过程顺利进行。仿真技术，如离散时间仿真范式（DEVS），是非常重要的一类建模与仿真方法，多年来在大量人造系统中得到了广泛应用。而基于本体构建 DEVS 模型对化工生产过程进行控制，则利用本体强大的表达能力建立 DEVS 本体模型，并利用本体强大的推理能力对 DEVS 本体模型进行校验，避免将错误代入仿真过程，保证化工生产过程的安全进行。使用本体描述 DEVS 模型的基本组成元素以及基本组成元素之间的关系，得到 DEVS 本体模型；依据化工生产过程构建生产模型，并利用该生产模型进行 DEVS 本体模型实例化，得到初始 DEVS 实例；对初始 DEVS 实例进行推理校验，然后对获得的冲突进行修正，得到 DEVS 实例；将 DEVS 实例映射为 DEVS 模型代码；将 DEVS 模型代码应用于化工生产过程仿真，并依据仿真结果控制化工生产过程。

2. 规则知识建模

1）规则知识的建模语言

在规则引擎中，通常会使用某种表述性的语言来描述规则。所以规则建模语言也是规则引擎的一个重要组成部分。目前的规则建模语言，并没有一个通用的标准获得规则引擎厂商的广泛支持，大部分规则建模语言都是厂商私有的。大体来说，规则建模语言可以分为结构化的（Structured)和基于标记的（Markup，通常为 XML）两类。下面介绍几种常用的逻辑规则编程语言。

Prolog 的命名来自“逻辑编程”（programming of Logic），广泛应用在人工智能的研究中。它创建在逻辑学的理论基础之上，最初被运用于自然语言等研究领域。它可以用来建造专家系统、自然语言理解系统、智能知识库等。只要给出事实和规则，它会自动分析其中的逻辑关系，然后允许用户通过查询，完成复杂的逻辑运算。

Datalog 是一种数据查询语言，语法与 Prolog 相似。Datalog 不是某一种具体的语言，而是一个规范，bddbddb、DES、OverLog、Deals 等都按照 Datalog 的语法实现了自己的语

言。Datalog 的语法是 Prolog 的子集，但是 Datalog 的语义与 Prolog 不同，Prolog 程序里的事实和规则的出现顺序决定了执行结果，Datalog 程序对事实和规则的出现顺序不做要求。

Rule Markup Language (RuleML)是一系列 Web 文档和数据语言的统一系统，通过模式语言进行句法指定，最初为 XML 开发并转换为其他格式，如 JSON。RuleML 允许部分受约束的语义简档和完全指定的语义。作为一种基于研究的语言系统，RuleML 可以作为 Prolog 和 N3，F-logic 和 TPTP，RIF 和 Common Logic 等语言的连接器。RuleML 已经在其他规则语言之间（如 SWRL 和 SWSL）提供了适应，扩展的互操作桥梁。

Semantic Web Rule Language（SWRL）的规则部分概念是由 RuleML 所演变而来，并结合了 OWL 本体论的部分概念。其以语义的方式呈现规则。SWRL 已经是 W3C 规范中的一员。通过两者的组合可以使得在撰写规则时，直接使用本体论中所描绘的关系和词汇，而本来这些类别之间的关系可能还需要额外的描述，但在 SWRL 中可以直接使用本体论描述。

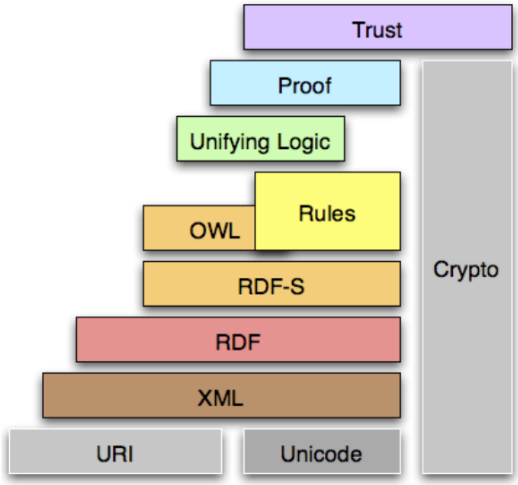


图 2 规则在语义网技术栈中的定位

2）规则知识的建模工具

Cyc 推理引擎。Cyc 是一个人工智能项目，致力于将各个领域的本体和常识知识进行整合，并在此基础上实现知识推理。它的目标是使人工智能的应用能够以类似于人类推理的方式工作。部分项目以 OpenCyc 的形式发布，OpenCyc 项目以开源许可的形式向开发者和用户提供 API，并可以下载数据集。

KAON2 是用于管理 OWL-DL、SWRL 和 F-Logic 本体论的基础架构。它能够操纵 OWL-DL 本体论；可以使用 SPARQL 完成查询。KAON2 是 KAON 项目（通常称为 KAON1）的继任者。与 KAON1 的主要区别在于支持的本体语言：KAON1 使用 RDFS 的专有扩展，而 KAON2 基于 OWL-DL 和 F-Logic，并且与 KAON1 不兼容。

Drools 是一个业务规则管理系统，具有基于前向链接和后向链接推理的规则引擎，可以

快速，可靠地评估业务规则和进行复杂的事件处理，其基于 Rete 算法的增强算法实现。Drools 作为一个易于访问企业策略、易于调整以及易于管理的开源业务规则引擎，符合业内标准，速度快、效率高。业务分析师人员或审核人员可以利用它轻松查看业务规则，从而检验已编码的规则是否满足所需的业务要求。

Flora-2 是一个开源的基于语义规则的规则引擎。系统的语言来源于 F-logic，HiLog 和 Transaction logic。基于 F-Logic 和 HiLog 意味着面向对象的语法和高阶表示是 Flora-2 系统的主要特征，其还支持一种可废止的推理形式，称为具有默认值和论证理论的逻辑编程(LPDA)。

Prova 是一个基于规则的脚本系统，用于中间件。该语言通过使用允许调用 Java 函数的 prolog 语法结合了命令式和声明式编程。Prova 通过提供适当的语言语法与 Java 的本机语法集成、代理消息传递和反应规则来扩展 Mandarax。

3）规则知识建模的应用实践示例

规则引擎在应用中作为一个嵌入在应用程序中的组件，其核心思想是将复杂多变的规则从业务流程中解放出来，以规则脚本的形式存储在文件或数据库中。使得业务规则的变化不需要修改代码重启机器就可以在线完成。

（1）1970 年代，斯坦福大学利用 LISP 语言开发了世界上第一个基于规则的系统——MYCIN 系统，主要用于血液疾病的诊断，并给出了相应的治疗方法。在该应用中，知识与控制分离，即知识抽象出相应的规则，与评价和执行的控制逻辑程序分离。这是业务规则引擎技术的启蒙时期。

（2）规则引擎也适用于政务服务。例如，在税收制度的发展中，需要明确业务规则，如公司名称的长度、合伙纳税人的比例、外商投资企业的比例等。税制的改革和完善必然会改变税收业务规则，而如果将相关的规则直接写入源代码，在方案的进行修改的时候必然会导致项目的重新部署，给维护工作带来极大的不便。使用规则引擎可以很好地解决以上问题。

（3）在电商场景下，我们可以利用规则在电商场景下进行同款商品挖掘，并且在达到目的的基础上可以得到符号化的知识表示，将其作为一种选择策略辅助人工进行判断。同款商品规则建模发现旨在将不同平台、不同商家销售的同一款商品挂载到统一的产品实体上。这有助于打通不同平台商品实体之间的联通，提供跨域的实体对齐，从而构建一个更完备的商品知识图谱。规则的作用是帮助业务人员对齐新的商品对，并且规则具备可解释性，业务人员能够很清楚地知道两个商品是基于哪些重要的属性和属性值来判断对齐与否，因此规则在实际的电商场景上有很大的应用。但目前规则库中的规则主要是由业务专家构造，为了提升规则建模效率降低规则建模成本，部分规则也采取了自动化的规则挖掘方法。

四、技术展望与发展趋势

近年来，本体知识建模利用自然语言处理、机器学习等技术从多源异构数据中进行自动化的构建取得了长足的进展。自动构建的过程中，如果数据是结构化的(例如图表数据)，已知属性名称、属性间的层次结构等，构建本体相对较为容易。如果缺乏以上信息，则只能通过文本信息等非结构化数据提炼知识构建本体，技术上将面临很多挑战。整体来看，呈现以下趋势：

- 多模态数据及数据的结构化工作。随着数据资源的丰富，越来越多的本体构建工作需要处理多种模态的数据，进行模态融合和语义的对齐，例如，从图片或文本中提取出结构化的知识，进行语义对齐。
- 低资源场景下的本体构建。由于标记数据的缺乏或相关领域数据保密的要求，本体的自动化构建面临小样本甚至是零样本的挑战，近年来的很多工作围绕小样本和零样本利用集成学习、多任务学习、预训练模型、元学习等技术结合深度学习模型进行探索。
- 大规模本体构建。随着算力的不断提升，现有实用系统可以有效处理更大规模的本体数据，数据量大、种类多样、结构不同都为本体构建带来巨大的挑战。

规则引擎的发展也遇到了很多问题，需要在未来进一步研究和解决。主要问题如下：

- 规则可视化配置。需要设计更高级的方案，让业务人员通过界面引导配置各种规则，而不是让技术人员从后台手动配置，彻底解放技术人员。
- 规则执行的效率。在规则数量不断增加，业务数据被索引并不断增长的情况下，如何快速选择规则，做出快速准确的决策，不会使规则数量成为系统的瓶颈。
- 规则的维护。当规则数量增加时，如何维护这些规则？更改规则时如何保证与之前发布的规则不冲突？规则能够更好的维护将使这些规则不会相互干扰，相互独立。

应用方面，随着企业智能化进程的加快，知识工程与产业互联的结合更加紧密，除了在数据治理、搜索与推荐、问答等通用领域有所突破之外，在智能生产、智慧城市、智能管理、智能运维等众多领域，以及工业、金融、司法、公安、医疗、教育等众多行业也都有进一步的场景化落地的突破。但落地的热潮在应对不同领域的知识建模需求时，需要行业专家与 AI 技术人才进行深度的磨合和协作。

参考文献

[Neches et al., 1991] Neches R, Fikes R E, Finin T, et al. Enabling technology for knowledge sharing[J]. AI magazine, 1991, 12(3): 36-36.

[陈华钧 et al., 2011] 陈华钧. 基于本体推理的中医药五行诊疗系统：中国，CN102156801A [P]. 2011-08-17.

[荣冈 et al., 2015] 荣冈. 一种基于本体构建模型的化工生产过程控制方法：中国，CN104678780A [P]. 2015-06-03.

第二章 知识表示学习

刘知远，汪华东

清华大学 计算机科学与技术系，北京 100084

一、任务定义、目标和研究意义

知识表示是知识获取与应用的基础，因此知识表示学习问题，是贯穿知识图谱的构建与应用全过程的关键问题。人们通常以网络的形式组织知识图谱中的知识，网络中每个节点代表实体（人名、地名、机构名、概念等），而每条连边则代表实体间的关系。然而，直接应用符号表示的知识图谱存在计算效率低、数据稀疏等诸多挑战性难题。近年来，以深度学习为代表的表示学习[Bengio et al., 2013]技术得到了广泛研究，在自然语言处理、图像分析和语音识别领域取得了巨大成功。表示学习旨在将研究对象的语义信息表示为稠密低维实值向量。在该低维向量空间中，两个对象距离越近，则说明其语义相似度越高。知识表示学习，则是面向知识图谱中的实体和关系进行表示学习。

知识表示学习实现了对实体和关系的分布式表示，它具有以下主要优点：

（1）显著提升计算效率。知识图谱的三元组表示实际就是基于独热表示的。如前所分析的，在这种表示方式下，需要设计专门的图算法计算实体间的语义和推理关系，计算复杂度高，可扩展性差。而表示学习得到的分布式表示，则能够高效地实现语义相似度计算等操作，显著提升计算效率。

（2）有效缓解数据稀疏。由于表示学习将对象投影到统一的低维空间中，使每个对象均对应一个稠密向量，从而有效缓解数据稀疏问题，这主要体现在两个方面。一方面，每个对象的向量均为稠密有值的，因此可以度量任意对象之间的语义相似度。另一方面，将大量对象投影到统一空间的过程，也能够将高频对象的语义信息用于帮助低频对象的语义表示，提高低频对象的语义表示的精确性。

（3）实现异质信息融合。不同来源的异质信息需要融合为整体，才能得到有效应用。例如，人们构造了大量知识图谱，这些知识图谱的构建规范和信息来源均有不同。大量实体和关系在不同知识图谱中的名称不同。如何实现多知识图谱的有机融合，对知识图谱应用具有重要意义。通过设计合理的表示学习模型，将不同来源的对象投影到同一个语义空间中，就能够建立统一的表示空间，实现多知识图谱的信息融合。

综上，由于知识表示学习能够显著提升计算效率，有效缓解数据稀疏，实现异质信息融合，因此对于知识图谱的构建、推理和应用具有重要意义，值得广受关注、深入研究。

二、研究内容和关键科学问题

知识表示学习是面向知识图谱中实体和关系的表示学习。通过将实体或关系投影到低维向量空间，我们能够实现对实体和关系的语义信息的表示，可以高效地计算实体、关系及其之间的复杂语义关联。这对知识图谱的构建、推理与应用均有重要意义。目前，已经在知识图谱补全、关系抽取等任务中取得了瞩目成果。但是，知识表示学习仍然面临很多挑战。

1. 复杂关系建模

现有知识表示学习方法无法有效地处理知识图谱中的复杂关系。这里的复杂关系定义如下。按照知识图谱中关系两端连接实体的数目，可以将关系划分为 1-1、1-N、N-1 和 N-N 四种类型。例如 N-1 类型关系指的是，该类型关系中的一个尾实体会平均对应多个头实体，即我们将 1-N、N-1 和 N-N 称为复杂关系。研究发现，各种知识获取算法在处理四种类型关系时的性能差异较大，在处理复杂关系时性能显著降低。如何实现表示学习对复杂关系的建模成为知识表示学习的一个难点。

2. 多源信息融合

知识表示学习面临的另外一个重要挑战如何实现多源信息融合。现有的知识表示学习模型仅利用知识图谱的三元组结构信息进行表示学习，尚有大量与知识有关的其他信息没有得到有效利用，例如：（1）知识图谱中的其他信息，如实体和关系的描述信息、类别信息等；（2）知识图谱外的海量信息，如互联网文本蕴含了大量与知识图谱实体和关系有关的信息。如何充分融合这些多源异质信息，实现知识表示学习，具有重要意义，可以改善数据稀疏问题，提高知识表示的区分能力。

3. 关系路径建模

在知识图谱中，多步的关系路径也能够反映实体之间的语义关系。Lao 等人曾提出 Path-Constraint Random Walk [Lao & Cohen et al., 2010]、Path Ranking Algorithm [Lao et al., 2011] 等算法，利用两实体间的关系路径信息，预测它们的关系，取得显著效果，说明关系路径蕴含着丰富信息。如何突破知识表示学习孤立学习每个三元组的局限性，充分考虑关系路径信息是知识表示学习的关键问题。

4. 时序信息建模

当前的知识图谱的研究主要集中在事实不随时间变化的静态知识图谱上，而对于知识图谱的时序动态性则很少被研究。实际上，知识图谱的大量知识具有时效性，随着时间发展是动态变化的，如：美国总统在 2010 年是“贝拉克·奥巴马”，在 2020 年是“乔·拜登”。因此，

对知识图谱中的时序信息建模是十分重要的。充分建模知识图谱富含的时序信息，利用时序分析和图神经网络等技术，对于分析图谱结构随时间的变化规律和趋势，以及知识推理都具有重要意义。

5. 模型知识增强

语言模型是自然语言理解的核心能力，以预训练语言模型 BERT、GPT 为代表的最先进的深度学习方法，仍然面临鲁棒性差、可扩展性差和可解释性差等问题。此外，语义的深度理解离不开多类型知识推理，因此建立面向预训练语言模型的模型知识增强机制，是知识融合的关键科学问题。知识表示学习是构建结构化符号知识到深度语言模型的桥梁，如何低成本植入结构化知识到预训练语言模型增强模型的语义理解能力，是目前知识表示学习的热点方向。

三、技术方法和研究现状

知识表示学习是近年来的研究热点，研究者提出了多种模型，学习知识图谱中的实体和关系的分布式表示。本节将围绕知识表示学习中关键科学问题对于相关技术研究进展进行介绍。

1. 复杂关系建模

TransE [Bordes et al., 2013]将知识图谱中的关系看作实体间的某种平移向量，由于模型简单，在大规模知识图谱上效果明显。但是也由于过于简单，导致 TransE 不能处理知识图谱中的复杂关系。为了解决 TransE 模型在处理 1-N、N-1、N-N 复杂关系时的局限性，许多 TransX 系列模型被提出。TransH 模型[Wang et al., 2014b]提出让一个实体在不同的关系下拥有不同的表示。TransR [Lin et al., 2015b]进一步认为不同的关系拥有不同的语义空间，对每个三元组将实体利用矩阵投影到对应的关系空间中，再建立从头实体到尾实体的平移关系。Qian 等人[Qian et al., 2018]提出了 TransAt 模型,通过引入注意力机制来进行表示学习。TransMS [Yang et al., 2019]则使用非线性函数来传播多向语义。

部分工作则从表示空间着手，ManifoldE [Xiao et al., 2016a]将传统的基于“点”的表示扩展为流形表示，并设计了 Sphere 和 Hyperplane 两种流形的设置。ComplEx [Trouillon et al., 2016]从复数空间上建模实体和关系嵌入，以更好地捕获对称和非对称的关系。RotatE [Yang et al., 2019]在复数空间上将关系看做是头实体到尾实体的旋转。HAKE [Zhang et al., 2020c]则是将实体映射到极坐标系，通过在链接预测任务上的实验表明 HAKE 能有效地在知识图中建立语义层次模型。目前的研究大多集中在静态知识图谱上，但时序知识图谱也同样重要。

[Sadeghian et al., 2021]提出了一种用于学习实体、关系和时间表示的模型 ChronoR ，可以通过使用高维旋转作为变换算子，捕捉到时间和多关系特征之间的丰富信息，并在时序知识图谱链接预测任务取得优异效果。

可以看到，在 TransE 之后，在如何处理复杂关系建模的挑战问题上，提出了多种模型，从不同角度尝试解决复杂关系建模问题，可谓百花齐放。在相关数据集合上的实验表明，这些方法均较 TransE 有显著的性能提升，验证了这些方法的有效性。

2. 多源信息融合

现有知识表示学习模型如 TransE 等，仅利用知识图谱的三元组结构信息进行表示学习，尚有大量与知识有关的其他信息没有得到有效利用。因此需要对现有知识表示学习模型进行多源信息融合，融合包括文本描述、类别、属性以及图片等多源异质信息。在融合上述信息方面，许多研究工作被提出[Ji et al., 2021]。

文本描述。多数知识图谱中含有大量对实体进行描述的文本信息，这些文本数据包含着丰富的语义信息。[Wang et al., 2014a]提出融合文本信息的知识表示学习方法，其利用 Word2Vec 学习维基百科正文中的词表示，利用 TransE 学习知识图谱中的知识表示。然后利用维基百科正文中的链接信息（锚文本与实体的对应关系），让文本中实体对应的词表示与知识图谱中的实体表示尽可能接近。此外，DKRL[Xie et al., 2016a]考虑知识图谱中提供的实体描述文本信息，给出了两种融合本文描述信息的模型：一种是 CBOW，将文本中的词向量简单相加作为文本表示；一种是 GCN，能够考虑文本中的词序信息。最近几年，预训练语言模型在各种 NLP 任务中表现出优越性能，其通过海量数据训练实现了对文本的丰富语义模式和语言信息的编码。[Wang et al., 2021]提出了预训练语言表示和知识表示联合学习的统一模型 KEPLER，如图 1 所示，其通过联合学习不仅能够将事实知识信息更好的嵌入到预训练语言模型中，同时通过基于文本训练的预训练语言模型可以得到文本语义增强的知识表示。

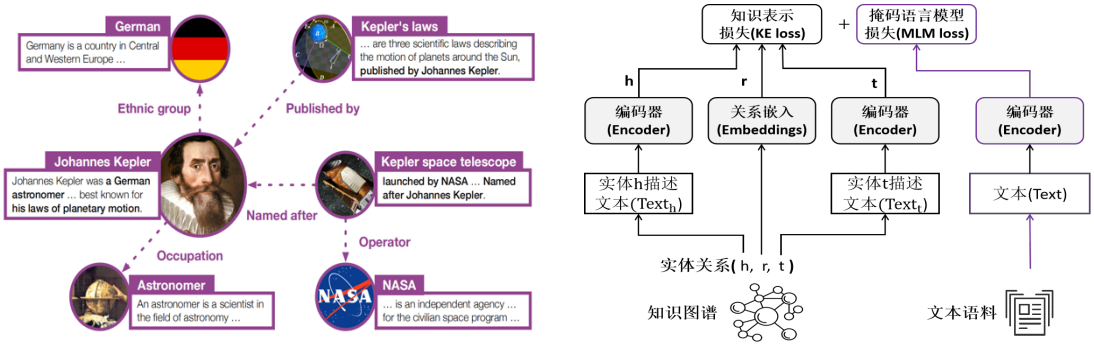


图 1 多源信息融合：含文本描述的知识图谱例子（左图），KEPLER 框架（右图）

实体类别。实体由层次类或类型和语义类别的关系来表示。融合实体相关类别信息有助于增强实体的语义表示。[Guo et al., 2015]提出 SSE(Semantically Smooth Embedding)模型，尝试引入实体的语义类别信息，使得来自同一类别的实体在嵌入空间更为接近。[Xie et al., 2016c]提出融合类别的知识表示学习模型 TKRL，该模型是第一个借助层次结构信息将实体类别信息编码到知识表示的方法。其他融合类别信息到知识表示中的工作可以参考 [Zhang et al., 2018; Niu et al., 2020a]。

视觉信息。知识图谱中实体通常包含着丰富的视觉信息，如人物照片、动物图片、公司 Logo 等。IKRL[Xie et al., 2016b]提出了一种将图像信息融入到知识图谱中来进行知识表示的学习方法，该方法在知识补全和三元组分类任务中均取得了不错的性能，也说明了跨模态的图像信息对于图谱是一个有效的补充。在 IKRL 的基础上，[Mousselly-Sergieh et al., 2018]提出了一种同时融入基于语言学和图像信息的多模态知识表示方法，并构建了一个大规模的多模态知识表示数据集。其他工作可以参考[Wang et al., 2019; Zhang et al., 2020a]。

逻辑规则。另外一种可以被利用的信息是逻辑规则。[Guo et al., 2016]提出 KALE 是将逻辑规则和知识图谱进行共同表示的典型工作。KALE 在一个统一的框架中表示三元组和给定的逻辑规则，并获得实体和关系的向量表示。 具体而言，其将三元组看成原子公式，并利用转移模型进行建模。规则被形式化为复杂公式，并利用 t 阶模糊逻辑建模，并 将复合公式的真值定义为其成分真值的组合。[Guo et al., 2018]又进一步提出了基于软规则的改进方法 RUGE。

多语言信息。多语言知识图谱(如 DBpedia)一般都包含几种不同语言实体中的机构性知识，并且它们对于跨语言应用都是有用的资源。因此多语言知识图谱的表示方法也是值得关注的一个研究领域。Chen 等人提出了 MTransE [Chen et al., 2017], 是第一个将知识表示推广到多语言场景的工作。MTransE 分别在独立空间中对实体和关系进行编码，并可以对任意实体或关系向量进行跨语言转换，且多语言图谱的嵌入模型保留了单语嵌入时的优良特性。IPTransE [Zhou et al., 2017]则将不同 KG 的实体和关系联合编码到一个统一的低维语义空间中,并提出了一种迭代和参数共享的方法来提高跨语言对齐性能。[Sun et al., 2018]则提出了一种基于实体对齐的知识图谱嵌入方法。

不确定信息。一些具有不确定性信息的知识图谱（如 NELL）给每个三元组添加一个置信度来描述三元组的不确定性。那么不确定知识图谱表示学习任务，需要实体与关系的表示向量同时嵌入图谱的结构信息与置信度信息。UKGE [Chen et al., 2019]首先关注了不确定信息的，通过引入规则作为先验知识，并利用概率软逻辑方式进行置信度推断。而[Zhang et al., 2021]的工作则关注不确定知识图谱中长尾关系的少样本问题，提出了基于高斯分布的度量

学习方法，利用 Gaussian Embedding 方式建模实体及关系的语义不确定性。[Boutouhami et al.,2019]则考虑知识图谱中存在不确定本体信息问题，提出不确定本体感知知识图谱嵌入模型 UOKGE，根据置信度分数学习不确定本体感知知识图上的实体、类和属性的嵌入。

已有工作表明，多源信息融合能够有效提升知识表示的性能，特别是可以有效处理新实体的表示问题。从目前来看，多源信息融合的知识表示学习处于快速发展的阶段，尽管如此，考虑的信息源非常有限，有大量的信息（如音频、视频等）未被考虑，具有广阔的研究前景。

3. 关系路径建模

关系路径是指两个实体之间的多步关系，而不仅限于两个实体之间直接相连的关系。目前许多研究方法主要基于三元组（头实体，关系，尾实体）方式学习图嵌入表示，这类 Triple-level 学习方法仅从一个局部的视图（即一跳关系邻居）中学习实体嵌入，忽略了图谱的关系路径信息。实际上，在知识图谱中，多步关系包含了两个实体之间丰富的语义关系，有助于多步推理，如图 2 所示。为此，[Lin et al.,2015a]提出考虑关系路径的表示学习方法，以 TransE 作为扩展基础，提出 Path-based TransE（PTransE）模型，该模型将关系路径建模成一组关系的组合，并给出了相加、相乘和循环网络等多种关系组合形式。为了在知识图谱中结合更多的信息，[Guo et al.,2019]提出了循环跳跃网络模型 RSN，沿着关系路径对实体和关系进行联合学习，该模型将递归神经网络与残差连接相结合，以捕获知识图谱中长期依赖关系。以上方法利用路径上关系或实体表示的数值计算结果作为关系路径的表示，存在误差传播和可解释性差的问题，[Niu et al.,2020b]为此提出一种联合路径和规则的知识表示学习模型 RPJE。受到神经架构搜索（NAS）的启发，[Zhang et al., 2020b]提出将 Interstellar 作为一种处理关系路径中信息的循环架构搜索问题，以获取路径中的短期和长期信息。此外，基于图神经网络 GNN 被广泛用于知识图谱的图嵌入学习建模，其通过多层聚合方式可以实现对图中多跳依赖信息的建模，代表性方法有 R-GCN [Schlichtkrull et al, 2018]、基于注意力的特征嵌入模型 [Nathani et al., 2019]等。最近受 Transformer 强大的语义编码能力启发，研究者提出利用 Transformer 和预训练语言模型的关系路径编码方法，如 CoKE [Wang et al., 2019b]给出了上下文（边和路径）知识图嵌入方法，使用 Transformer 编码器获得上下文信息。

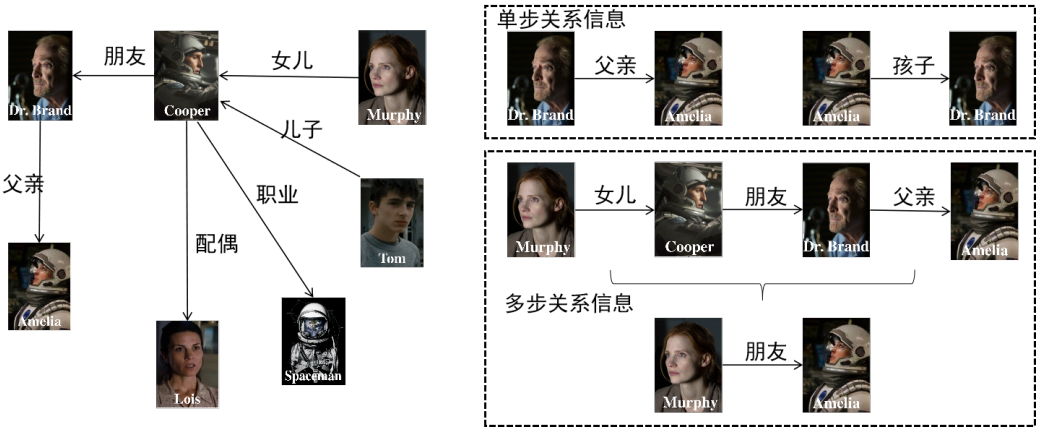


图 2 知识图谱中单步和多步关系信息示例 (原图来自[Zhang et al., 2020b])

以上关系路径建模的相关研究实验表明，考虑关系路径能够极大提升知识表示学习的区分性，提高在知识图谱补全等任务上的性能。关系路径建模目前许多相关工作还比较初步，特别是在关系路径的可靠性计算、关系路径的语义组合操作、与复杂推理联合建模等方面，还有很多细致的考察工作需要完成。

4. 时序信息建模

目前知识图谱表示学习的研究主要集中在静态知识图谱上。但许多事实在时间序列中是不断变化发展，所以时序知识图谱也同样重要。围绕着时序知识图谱，许多研究开始将时间信息纳入知识图谱表示学习和相关任务中，对时间序列中的知识进行表示学习。这些工作可以分为两类：外推任务(Extrapolation task)和插值任务(Interpolation task)[Liao et al., 2021]。

外推任务，旨在对未来的事实进行预测。为了解决外推任务，[Trivedi et al., 2017]提出了一种知识进化算法，该算法通过时间点过程根据时间 t−1 的状态来估计一个事实在时间 t 时是否成立。[Jin et al., 2019]使用一个邻域聚合器来考虑并发事件，并利用递归神经网络(RNNs)来捕获时间序列的时间条件联合概率分布。[Xu et al., 2020]提出了 ATiSE, 则是考虑了知识图谱在时间演化过程中的不确定性因素，采用多维高斯分布函数来对图谱进行表示学习。[Liao et al., 2021]构建了动态贝叶斯知识图嵌入模型(DBKGE)，在联合度量空间中动态地跟踪实体的语义表示，并对未来做出预测。

插值任务, 建立在一个插值任务公式上，目的是预测一个事实在给定的时间点是否有效，也称为时序知识图补全。目前对时序知识图谱表示学习的研究大多是基于插值任务的。[Leblay & Chekol, 2018]在扩展现有的关系嵌入模型的基础上，提出了各种考虑时间信息的方法。[Garcia-Duran et al., 2018]将谓词序列和时间戳序列进行拼接构成一个关系序列，然后输入到 LSTM 中进行编码，用以进行时间信息感知的表示学习。[Dasgupta et al., 2018]提出了一种基于超平面的学习知识图谱表示的方法，该将时间戳转化为一种关系依赖的超平面，

并将实体和关系进行映射，从而有效计算评价分数。

从目前研究来看，围绕时序知识图谱的表示学习已经成为当前图谱表示学习领域的研究热点，相关研究进展较为显著，但相对于静态知识图谱的研究，其相关理论和技术体系还很不完善，存在许多挑战问题有待研究，如何建模时序关系的依存、时序逻辑推理等。

5. 模型知识增强

目前预训练语言模型(PLM)主要采用互联网获取的海量通用文本语料训练得到，实现了对文本丰富语义模式的编码，但由于没有自觉运用结构化知识，依然严重缺乏知识运用和推理能力，缺乏可解释性和鲁棒性。为此，许多学者研究了融合结构化知识的 PLM 及其学习框架[Yang et al., 2021]，融合方法大致分为以下几种[Han et al., 2021]，如图 3 所示：

知识增广，从输入端增强模型，有两种主流的方法：一种方式是直接把知识加到输入，另一方法是设计特定模块来融合原输入和相关的知识化的输入表示。目前，基于知识增广的方法已经在不同任务上取得良好效果，比如信息检索[Guu et al., 2020]、问答系统[Xiong et al., 2019]等。

知识支撑，关注于对带有知识的模型本身的处理流程进行优化。一种方式是在模型的底部引入知识指导层来处理特征，以便能得到更丰富的特征信息。例如，使用专门的知识记忆模块来从 PLM 底部注入丰富的记忆特征[Ding et al., 2020]。另一方面，知识也可以作为专家在模型顶层构建后处理模块，以计算得到更准确和有效的输出。例如，利用知识图谱来改进语言生成质量[Gu, et al., 2018]。

知识约束，利用知识构建额外的预测目标和约束函数，来增强模型的原始目标函数。例如，远程监督学习利用知识图谱启发式标注语料作为新的目标，并广泛用于系列 NLP 任务，如实体识别[Xin, et al., 2018]、关系抽取[Han, et al., 2018b]等。或者利用知识构建额外的预测目标，比如 ERNIE [Zhang et al., 2019c], CoLAKE [Sun, et al., 2020a]和 KEPLER [Wang et al., 2021]等工作，都是在原始的语言建模之外构建了相应额外的预训练目标。

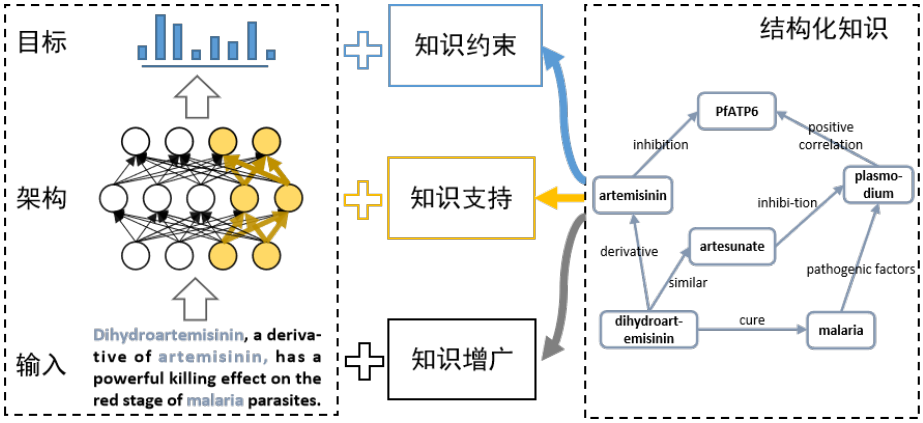


图 3 融合结构化知识到预训练语言模型的三种途径

6. 知识表示学习开源工具

目前围绕知识图谱表示，已经有大量的模型被提出，这些模型在基准数据集上取得了很好的性能。但是这些模型算法实现在一定程度上是分散的且不系统的。为了进一步促进这些模型的研究和开发，许多相关开源工具被提出，如 Han 等人发布了 OpenKE 工具包[Han et al., 2018a]，其是高效训练知识图谱表示的早期工具包之一，提供了包括 TransE、TransR、CompIEx 等 8 种常见模型的实现。 Facebook 研发了 Pytorch-BigGraph 工具包[Lerer et al., 2019]，其开发重点是大模型图谱的尺度化以及机器集群上的分布式训练，但该软件包不支持分布式训练。[Zhu et al., 2019]发布了 GraphVite 工具包，聚焦在知识表示的多 GPU 训练，不仅支持一般图谱的表示学习同时支持知识图谱的表示学习。该工具方法以增加嵌入的持久性为代价，减少了 CPU 和 GPU 之间的数据移动。为了适应不断增长的知识图谱，[Zheng et al., 2020]提出了一种可有效计算知识图谱表示的开源软件包 DGL-KE，其引入了各种新的优化方法，利用多处理、多 GPU 和分布式并行性，实现对具有数百万个节点和数十亿条变的知识图谱的训练的高效加速。表 1 给列出了目前代表性开源工具包, 包括其已支持的最大知识图谱情况。尽管目前开源工具包在图谱高效训练上取得了巨大进展，但在处理实体超过千万级以上超大规模知识图谱时在模型性能、训练时间、内存消耗等方面还存在巨大挑战。

表 1 知识表示学习的开源工具

研发机构	名称	语言	GPU	最大 KG		时间
				关系	实体	
清华大学	Fast-TransX	C	×	100M	40M	2017
	OpenKE	Pytorch&C	√	21M	5M	2018
Amazon	DGL-KE	Pytorch	√	338M	86M	2019
蒙特利尔大学	GraphVite	Python&C++	√	21M	5M	2019
Facebook	BigGraph	Pytorch	√	**2.7B**	**121M**	2019
Accenture	AmpliGraph	TensorFlow	√	100M	1M	2019
加州大学	Pykg2vec	Pytorch	√	87K	41K	2019
曼海姆大学	LibKGE	Pytorch	√	1M	123K	2020
MIT	Scikit-KGE	Python	×	21M	5M	2021
波恩大学	PyKEEN	Python	√	21M	5M	2021

7. 测试基准数据集

为了评测知识表示学习算法的性能，目前已经有许多测试数据集被提出。这些数据集主

要从现有公开知识图谱基础上抽取子集构造，如：以语言知识图谱 WordNet 构造的数据集包括 WN18、WN11、WN18RR 等，以世界知识图谱 Freebase 构造的数据集如 FB40K、FB5M、FB86M 等，以链接知识库 Wikidata 构造的的数据集有 Wikidata5M、Wikidata68M、WikiKG90Mv2 等。此外，也有部分数据集通过其他类型知识图谱构造，如多语言知识图谱 YAGO 和跨语言知识图谱 XLORE。表 2 列举了目前代表性的测试基准数据集及其统计情况。可以看出，这些测试基准数据集覆盖了几万实体到近亿实体的不同尺度规模，可以充分满足目前的知识表示学习领域对模型算法预测准确率或训练速度的测试需要。另外，在融合异构信息的知识表示学习方面，Wikidata5M、WikiKG90Mv2 等数据集也提供了实体在维基百科的描述文本信息。另外，也有一些专门针对时序知识图谱的基准数据被提出，如 ICEWS14、ICEWS05-15、 GDELT 等。

表 2　知识图谱表示学习测试基准

数据集	关系	实体	训练集	验证集	测试集	来源
WN18	18	40,943	141,442	5,000	5,000	[Bordes et al.,2013]
WN11	11	38,696	112,581	2,609	10,544	[Socher et al.,2013]
WN18RR	11	40,943	86,835	3,034	3,134	[Dettmers et al.,2018]
FB13	13	75,043	316,232	5,908	23,733	[Socher et al.,2013]
FB15K	1,345	14,951	483,142	50,000	59,071	[Bordes et al.,2013]
FB15K-237	237	14,541	272,115	17,535	20,466	[Schlichtkrull,2018]
FB40K	1,336	39,528	370,648	67,946	96,678	[Lin et al., 2015b]
FB5M	1,192	5,385,322	19,193,556	50,000	59,071	[Wang et al., 2014b]
FB86M	14824	86,054,151	338,586,277	-	-	openke.thunlp.org
YAGO3-10	37	123,182	1,079,040	5,000	5,000	[Ali et al.,2021]
Wikidata5M	822	4,594,485	20,614,279	5,163	5,163	[Wang et al.,2021]
Wikidata68M	595	20,982,734	68,902,802	-	-	openke.thunlp.org
WikiKG90Mv2	1,387	91,230,610	601,062,811	15,000	15,000	KDD CUP 2021
XLORE	138581	10,572,210	35,954,250	100,000	100,000	openke.thunlp.org

四、技术展望与发展趋势

近些年，虽然面向知识图谱的知识表示学习领域发展迅速，相关基础理论及其应用技术趋于完备，但是仍然存在许多挑战问题有待进一步研究，本节将对知识表示学习的未来方向

进行展望。

面向不同知识类型的知识表示学习。已有工作将知识图谱的关系划分为 1-1、1-N、N-1 和 N-N 四类，这种关系类型划分略显粗糙，无法直观地解释知识的本质类型特点。根据认知科学研究[Kemp & Tenenbaum, 2009; Tenenbaum et al., 2011]，人类知识包括以下几种结构：（1）树状关系，表示实体间的层次分类关系；（2）二维网格关系，表示现实世界的空间信息;（3）单维顺序关系，表示实体间的偏序关系；（4）有向网络关系，表示实体间的关联或因果关系。认知科学对人类知识类型的总结，有助于对知识图谱中知识类型的划分和处理。未来有必要结合人工智能和认知科学的最新研究成果，有针对性地设计知识类型划分标准，开展面向不同复杂关系类型的知识表示学习研究。

面向多源信息融合的知识表示学习。在多源信息融合的知识表示学习方面，相关工作还比较有限，主要是考虑实体描述的知识表示学习模型，以及文本与知识图谱融合的知识表示学习，这些模型无论是信息来源，还是融合手段都非常有限。我们认为在多源信息融合的知识表示学习方面，我们还可以对下列方面进行探索:（1）融合知识图谱中实体和关系的其他信息，知识图谱中拥有关于实体和关系的丰富信息，如描述文本、层次类型等。有机融合这些信息，将显著提升知识表示学习的表示能力;（2）融合互联网文本、图像、音频、视频信息，互联网海量文本、音频、视频数据是知识图谱的重要知识来源，有效地利用这些信息进行知识表示可以极大地提升现有知识表示方法的表示能力;（3）融合多知识图谱信息，人们利用不同的信息源构建了不同的知识图谱。如何对多知识图谱信息进行融合表示，对于建立统一的大规模知识图谱意义重大。

考虑复杂推理模式的知识表示学习。考虑关系路径的知识表示学习，实际上是充分利用了两实体间的关系和关系路径之间的推理模式，来为表示学习模型提供更精确的约束信息。例如，根据三元组（康熙, 父亲，雍正）和（雍正, 父亲，乾隆）构成的“康熙”和“乾隆”之间“父亲+父亲”的关系路径，再结合三元组（康熙, 祖父，乾隆），通过构建“父亲+父亲=祖父”的推理模式，提升知识表示的精确性。 此外，知识图谱中还有其他形式的推理模式，例如三元组（美国，总统，奥巴马）和（奥巴马，是，美国人）之间就存在着推理关系，但是两者的头、尾实体并不完全一致。如果能将这些复杂推理模式考虑到知识表示学习中，将能更进一步提升知识表示的性能。 在该问题中，如何总结和表示这些复杂推理模式，是关键难题。目前来看，一阶逻辑是对复杂推理模式的较佳表示方案。

超大规模知识图谱的知识表示学习。虽然已经出现了 GraphVite、OpenKE、DGL-KE、BigGraph 等知识表示学习开源工具，但这些工具还主要针对百万级实体规模以内的知识图谱，处理的最大知识知识图谱规模，这限制了大规模知识图谱的应用潜力。目前知识图谱的

规模越来越大，如 Wikidata 已经含有了超过 9 千万实体、14.7 亿的关系，而且这种规模仍然呈现快速增长趋势。如何将现有知识表示学习方法适配到千万级以上实体规模的图谱上仍然是一个挑战，需要解决在优化过程中大规模知识图谱的高质量负采样、模型多维并行训练机制（如数据并行、模型并行、流水并行等）以及并行训练中高效内存和通信管理等关键问题。

大规模知识图谱的在线知识表示学习。在实际中，知识图谱的规模不断扩大的，且知识信息也随着时间不短更新，如 DBpedia 每天提取维基百科的更新流，以保持其知识图谱包含最新信息，阿里的产品知识图谱需要相当频繁地更新，由于其电商平台每天都有大量的新产品上线。但是，现有的知识表示学习方法主要是聚焦在静态的知识图谱忽略了知识图谱的动态性。此外，大规模知识图谱稀疏性很强，初步实验表明，已有表示学习模型在大规模知识图谱上性能堪忧，特别是对低频实体和关系的表示效果较差，根据知识图谱动态更新实体和关系表示突破该问题的重要途径。因此，我们需要设计高效的在线学习方案。除了充分融合多源信息降低稀疏性之外，我们还可以探索如何优化表示学习的方式，借鉴课程学习和迁移学习等算法思想，进一步改善知识表示的效果。

参考文献

[CIPS2018] 中国中文信息学会语言与知识计算专委会. 知识图谱发展报告(2018) 第二章 知识表示学习[C]. 2018:22-30.

[Ali et al.,2021] Ali M, Berrendorf M, Hoyt C T, et al. PyKEEN 1.0: a Python library for training and evaluating knowledge graph embeddings[J]. Journal of Machine Learning Research, 2021, 22(82): 1-6.

[Bengio et al., 2013] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on PAMI, 2013, 35(8): 1798-1828.

[Bordes et al., 2013] Bordes A, Usunier N, et al. Translating embeddings for modeling multi-relational data[C], in Proceedings of NIPS 2013, 2787-2795.

[Boutouhami et al., 2019] Boutouhami K, Zhang J, Qi G, et al. Uncertain ontology-aware knowledge graph embeddings[C]. Joint International Semantic Technology Conference. Springer, Singapore, 2019: 129-136.

[Chen et al., 2017] Chen M, Tian Y, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment[C], in Proceedings of IJCAI 2017: 1511-1517.

[Chen et al., 2019] Chen X, Chen M, et al. Embedding uncertain knowledge graphs[C], in

Proceedings of the AAAI.2019, 33(01): 3363-3370.

[Dasgupta et al., 2018] Dasgupta S S, et al. Hyte: Hyperplane-based temporally aware knowledge graph embedding[C], in Proceedings of EMNLP, 2018: 2001-2011.

[Dettmers et al., 2018] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2d knowledge graph embeddings[C], in Proceedings of AAAI. 2018, 32(1): 1811-1818.

[Ding et al., 2020] Ding M, Zhou C, Yang H, et al. Cogltx: Applying bert to long texts[J], in Proceedings of NeurIPS, 2020, 33: 12792-12804.

[García-Durán et al., 2018] García-Durán A, Dumancic S, Niepert M. Learning Sequence Encoders for Temporal Knowledge Graph Completion[C], in Proceedings of EMNLP. 2018.

[Gu et al., 2018] Gu Y, Yan J, Zhu H, et al. Language Modeling with Sparse Product of Sememe Experts[C], in Proceedings of EMNLP 2018: 4642-4651.

[Gu et al., 2015] Guo S, Wang Q, Wang B, et al. Semantically smooth knowledge graph embedding[C], in Proceedings of ACL-IJNLP 2015: 84-94.

[Guo et al., 2016] Guo S, Wang Q, Wang L, et al. Jointly embedding knowledge graphs and logical rules[C], in Proceedings of EMNLP 2016: 192-202.

[Guo et al., 2018] Guo S, Wang Q, Wang L, et al. Knowledge graph embedding with iterative guidance from soft rules[C], in Proceedings of AAAI, 2018, 32(1): 4816-4823.

[Guo et al., 2019] Guo L, Sun Z, Hu W. Learning to exploit long-term relational dependencies in knowledge graphs[C], in Proceedings of ICML 2019: 2505-2514.

[Guu et al., 2020] Guu K, Lee K, Tung Z, et al. Retrieval augmented language model pre-training[C], in Proceedings of ICML 2020: 3929-3938.

[Han et al., 2018a] Han X, Cao S, Lv X, et al. Openke: An open toolkit for knowledge embedding[C], in Proceedings of EMNLP: system demonstrations. 2018: 139-144.

[Han et al., 2018b] Han X, Liu Z, Sun M. Neural knowledge acquisition via mutual attention between knowledge graph and text[C], in Proceedings of AAAI 2018.

[Han et al., 2021] Han X, Zhang Z, Liu Z. Knowledgeable machine learning for natural language processing[J]. Communications of the ACM, 2021, 64(11): 50-51.

[Ji et al., 2021] Ji S, Pan S, et al. A survey on knowledge graphs: Representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.

[Jin et al., 2019] Jin W, Qu M, Jin X, et al. Recurrent event network: Autoregressive structure inference over temporal knowledge graphs[J]. arXiv preprint arXiv: 1904.05530, 2019.

[Kemp & Tenenbaum, 2009] Kemp C, Tenenbaum J B. Structured statistical models of inductive reasoning[J]. Psychological review, 2009, 116(1): 20.

[Lao&Cohen et al.,2010] Lao N, Cohen W W. Relational retrieval using a combination of path-constrained random walks[J]. Machine learning, 2010, 81(1): 53-67.

[Lao et al., 2011] Lao N, Mitchell T, Cohen W. Random walk inference and learning in a large scale knowledge base[C], in Proceedings of EMNLP 2011: 529-539.

[Lerer et al., 2019] Lerer A, Wu L, Shen J, et al. Pytorch-biggraph: A large-scale graph embedding system[J]. arXiv preprint arXiv:1903.12287, 2019.

[Leblay & Chekol, 2018] Leblay J, Chekol M W. Deriving validity time in knowledge graph[C]//Companion Proceedings of the The Web Conference. 2018: 1771-1776.

[Lin et al., 2015a] Lin Y, Liu Z, et al. Modeling Relation Paths for Representation Learning of Knowledge Bases[C], in Proceedings of EMNLP 2015: 705-714.

[Lin et al., 2015b] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C], in Proceedings of AAAI 2015.

[Liao et al., 2021] Liao S, Liang S, Meng Z, et al. Learning dynamic embeddings for temporal knowledge graphs[C], in Proceedings of WSDM. 2021: 535-543.

[Mousselly-Sergieh et al., 2018] Mousselly-Sergieh H, Botschen T, Gurevych I, et al. A multimodal translation-based approach for knowledge graph representation learning[C], in Proceedings of SEM. 2018: 225-234.

[Niu et al., 2020a] Niu G, et al. AutoETER: Automated Entity Type Representation for Knowledge Graph Embedding[J]. arXiv preprint arXiv: 2009. 12030, 2020.

[Niu et al., 2020b] Niu G, Zhang Y, Li B, et al. Rule-guided compositional representation learning on knowledge graphs[C], in Proceedings of AAAI 2020, 34(03): 2950-2958.

[Nathani et al., 2019] Nathani D, Chauhan J, Sharma C, et al. Learning Attention- based Embeddings for Relation Prediction in Knowledge Graphs[C], in Proceedings of ACL 2019: 4710-4723.

[Qian et al., 2018] Qian W, Fu C, Zhu Y, et al. Translating embeddings for knowledge graph completion with relation attention mechanism[C], in Proceedings of IJCAI 2018: 4286-4292.

[Sadeghian et al., 2021] Sadeghian A, Armandpour M, Colas A, et al. ChronoR: Rotation Based Temporal Knowledge Graph Embedding[C], in Proceedings of AAAI 2021, 35(7): 6471-6479.

[Sun et al., 2018] Sun Z, Hu W, Zhang Q, et al. Bootstrapping entity alignment with knowledge graph embedding[C], in Proceedings of IJCAI 2018: 4396-4402.

[Schlichtkrull et al, 2018] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[C]. European semantic web conference. Springer, 2018: 593-607.

[Sun et al., 2020a] Sun T, Shao Y, Qiu X, et al. CoLAKE: Contextualized Language and Knowledge Embedding[C], in Proceedings of ACL 2020: 3660-3670.

[Tenenbaum et al., 2011] Tenenbaum J B, Kemp C, Griffiths T L, et al. How to grow a mind: Statistics, structure, and abstraction[J]. science, 2011, 331(6022): 1279-1285.

[Trouillon et al., 2016] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[C], in Proceedings of ICML 2016: 2071-2080.

[Trivedi et al., 2017] Trivedi R, Dai H, Wang Y, et al. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs[C], in Proceedings of ICML. PMLR, 2017: 3462-3471.

[Wang et al., 2014a] Wang Z, Zhang J, Feng J, et al. Knowledge graph and text jointly embedding[C], in Proceedings of EMNLP. 2014: 1591-1601.

[Wang et al., 2014b] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C], in Proceedings of AAAI 2014, 28(1).

[Wang et al., 2019] Wang Z, Li L, Li Q, et al. Multimodal data enhanced representation learning for knowledge graphs[C], in Proceedings of IJCNN 2019: 1-8.

[Wang et al., 2019b] Wang Q, Huang P, Wang H, et al. CoKE: Contextualized knowledge graph embedding[J]. arXiv preprint arXiv:1911.02168, 2019.

[Wang et al., 2021] Wang X, Gao T,et al., KEPLER: A unified model for knowledge embedding and pre-trained language representation[J]. TACL, 2021, 9: 176-194.

[Xiao et al., 2016a] Xiao H, et al. From one point to a manifold: knowledge graph embedding for precise link prediction[C], in Proceedings of IJCAI 2016: 1315-1321.

[Xie et al., 2016a] Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions[C], in Proceedings of AAAI 2016, 30(1).

[Xie et al., 2016b] Xie R, Liu Z, Luan H, et al. Image-embodied Knowledge Representation Learning[C], arXiv preprint arXiv:1609.07028, 2016.

[Xie et al, 2016c] Xie R, Liu Z, Sun M. Representation learning of knowledge graphs with hierarchical types[C], in Proceedings of IJCAI. 2016: 2965-2971.

[Xin et al., 2018] Xin J, Lin Y, Liu Z, et al. Improving neural fine-grained entity typing with knowledge attention[C], in Proceedings of AAAI 2018.

[Xiong et al., 2019] Xiong W, Du J et al., Pretrained Encyclopedia: Weakly Supervised Knowledge-

Pretrained Language Model[C], in Proceedings of ICLR 2019.

[Xu et al., 2020] Xu C, Nayyeri M, Alkhoury F, et al. Temporal Knowledge Graph completion based on time series Gaussian embedding[C], International Semantic Web Conference. Springer, Cham, 2020: 654-671.

[Yang et al., 2019] Yang S, Tian J, Zhang H, et al. TransMS: Knowledge Graph Embedding for Complex Relations by Multidirectional Semantics[C], in Proceedings of IJCAI. 2019: 1935-1942.

[Yang et al., 2021] Yang J, Xiao G, Shen Y, et al. A Survey of Knowledge Enhanced Pre-trained Models[J]. arXiv preprint arXiv:2110.00269, 2021.

[Zhang et al., 2018] Zhang Z, Zhuang F, Qu M, et al. Knowledge graph embedding with hierarchical relation structure[C], in Proceedings of EMNLP 2018: 3198-3207.

[Zhang et al., 2019c] Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced Language Representation with Informative Entities[C], in Proceedings of ACL 2019: 1441-1451.

[Zhang et al., 2020a] Zhang Y, Fang Q, Qian S, et al. Multi-modal Multi-relational Feature Aggregation Network for Medical Knowledge Representation Learning[C], in Proceedings of ACM Multimedia. 2020: 3956-3965.

[Zhang et al., 2020b] Zhang Y, Yao Q, Chen L. Interstellar: Searching Recurrent Architecture for Knowledge Graph Embedding[J] , in Proceedings of NeurIPS 2020, 33: 10030-10040.

[Zhang et al., 2020c] Zhang Z, Cai J, et al. Learning hierarchy-aware knowledge graph embeddings for link prediction[C], in Proceedings of AAAI 2020.

[Zhang et al., 2021] Zhang J, Wu T, Qi G. Gaussian Metric Learning for Few-Shot Uncertain Knowledge Graph Completion[C], DASFAA. Springer, 2021: 256-271.

[Zheng et al., 2020] Zheng D, Song X, Ma C, et al. DGL-KE: Training knowledge graph embeddings at scale[C], in Proceedings of SIGIR. 2020: 739-748.

[Zhou et al., 2017] Zhou X, Zhu Q, Liu P, et al. Learning knowledge embeddings by combining limit-based scoring loss[C], in Proceedings of CIKM 2017: 1009-1018.

[Zhu et al., 2019] Zhu Z, Xu S, Tang J, et al. Graphvite: A high-performance cpu-gpu hybrid system for node embedding[C], in Proceedings of WWW 2019: 2494-2504.

第三章 实体抽取

林鸿宇，韩先培

中国科学院软件研究所 中文信息处理实验室，100190

一、任务定义、目标与研究意义

实体是世界构成的基本单元，文本中的实体是承载着文本信息的重要基本单位。一段文本中所蕴含的信息通常可以被表述为其所包含的实体以及实体之间的相互联系。例如，“美国白宫首席副新闻秘书卡琳·让·皮埃尔新冠检测结果呈阳性”中所包含的信息可以由句子中所提及的实体“美国”、“白宫”、“卡琳·让·皮埃尔”、“新冠”以及它们之间的相互关联所构成。而一个知识图谱通常是一个以实体为节点的巨大知识网络，包括实体、实体属性以及实体之间的关系。因此，上述文本中所蕴含的实体以及实体间关联信息是知识图谱中知识的最重要来源。实体抽取的主要目标是识别文本当中的实体提及，并将其划分到指定的给定类别[Chinchor & Robinson, 1998]。常用实体类别包括人名、地名、机构名、日期等。例如，给定“美国白宫首席副新闻秘书卡琳·让·皮埃尔新冠检测结果呈阳性”以及待抽取的实体类别“人名”与“地名”，实体抽取模型需要识别出“卡琳·让·皮埃尔”是一个人名实体提及，“美国”是一个地名实体提及等。实体抽取是海量文本分析和知识图谱构建的核心技术，也是文本语义理解的基础，为解决信息过载提供了有效手段。互联网海量的文本数据中蕴含大量有价值的信息，针对性地挖掘并剔除无关与冗余的信息，可以帮助人类高效获取信息。通过以实体为核心建立海量信息的表示、关联和结构，实体抽取可以为互联网信息的挖掘提供高效手段，为用户信息需求的精准满足提供基础支撑。实体抽取技术通过将文本结构化为以实体为中心的语义表示，为分析非结构化文本提供了核心技术手段，是实现大数据资源化、知识化和普适化的核心技术，已被广泛应用于舆情监控、网络搜索、智能问答等多个重要领域。

二、研究内容与挑战

实体抽取的主要研究对象是如何从文本中识别指定类别的实体。一个实体抽取系统通常包含两个部分：实体边界识别和实体分类，其中实体边界识别判断一个字符串是否组成一个完整实体，而实体分类将识别出的实体划分到预先给定的不同类别中去。实体抽取是一项自然语言处理的基础技术，目前中英文上通用的特定领域（人名、地名、机构名）实体抽取性能 F1 值都能达到 90% 以上。然而，对于实体抽取而言，当下最核心的挑战是如何能够将限定领域上的优良表现迁移至开放领域，从而在众多不同的领域与类别上均实现较好的性

能。相比于传统的限定领域实体抽取，开放领域实体抽取面临以下几个核心挑战：

1）类别开放：限定领域实体抽取通常只关注于非常稀少的特定实体类别。然而，开放域实体抽取需要处理为数众多、粒度不一的各种实体类别。为满足各领域的实体识别需求，命名实体的类别范围不断扩大。例如，语言数据协会根据《华尔街日报》的文章构建了包括了 64 个实体类别的 BBN 数据集，随后又构建了包括 87 个实体类别的 OntoNotes 数据集。Ling 等人从 Freebase 中选取了 112 个实体类别作为识别目标并构建了一个细粒度实体抽取数据集 [Ling & Weld, 2012]。Choi 等提出的极细粒度实体抽取，利用 WordNet 将实体扩充到了 10331 个粒度不一的开放类别 [Choi et al., 2018]。除了通用领域的命名实体识别，还有许多研究者针对特定领域进行了实体类别扩展。针对计算机科学领域，Jain 等提出了 SciREX 数据集，其中涵盖数据集、评价指标、任务和方法等四大类实体 [Jain et al., 2020]。针对生物领域，Li 等人根据 PubMed 的文章构造了 BC5CDR 数据集，重点关注疾病和化学药物实体 [Li et al., 2016]。众多的开放类别不仅数量不固定、粒度参差不齐，类别之间还具有上下位和共现关系等复杂关联。因此，传统的限定域实体抽取中孤立地考虑每个实体类别的方法是低效且不切实际的。

2）实体结构复杂：传统的限定领域实体抽取通常关注于平实体抽取 (Flat NER)，即不考虑实体提及中存在的嵌套、重叠以及不连续的情况。然而，对于开放领域的实体抽取而言，实体提及间存在复杂结构是非常常见的现象。例如“中华人民共和国教育部”中包含有“中华人民共和国教育部”和“中华人民共和国”两个不同的实体提及，而“心、肺功能异常”中则包含着“心功能异常”与“肺功能异常”两个不同的症状实体。这类复杂结构在开放领域的实体抽取问题中分布非常广泛，而以 CRF 为代表的传统的实体抽取模型表达能力不足，很难建模上述复杂结构。

3）标注资源缺乏：由于类别开放以及实体结构复杂的特点，我们很难为所有待抽取的实体类别构建足够数量的标注资源。因此，在开放实体抽取中有大量的实体类别仅有极少量的标注数据，只能提供极少量的信息。同时，虽然现在存在一定数量的外部标注/半标注实体抽取资源，但是这些资源通常标注质量较差，带有噪声，并且外部资源与任务目标可能存在知识不匹配的问题。因此，如何利用极少量的标注资源获得一个有效的实体抽取模型是开放领域实体抽取的又一大重要挑战。

上述三个挑战是限制了实体抽取从限定域迈向开放域的核心因素。近年来，实体抽取领域的研究工作大多围绕着解决上述三个挑战来进行的，而深度神经网络以及预训练语言模型的兴起为解决上述挑战带来了新的机遇。下面我们将简要介绍目前实体抽取领域的研究现状与发展趋势。

三、研究现状与发展趋势

基于深度神经网络的实体抽取方法当前已经居于统治地位。相比传统统计方法，深度学习方法的主要优点是其训练是一个端到端的过程，无需人工定义相关的特征。此外，深度学习方法还可以学习任务特定的表示，并建立不同模态、不同类型、不同语言之间信息的关联，从而取得更好的实体分析性能。近五年来，预训练深度语言模型的飞速发展更是为实体抽取领域带来了深刻的变革。对于实体抽取而言，深度学习与预训练语言模型不仅仅带来了一个更好的语言学编码器，还提供了一种有效的知识融合手段，打通了实体类别、语言、模态以及各种可用的资源之间的鸿沟，有效地提升了小样本、低资源、细粒度实体抽取的能力，为解决前述的三大挑战提供了重要的技术基础。下面本文将分别从模型架构、学习算法以及模态融合三个层间介绍实体抽取领域的研究现状与发展趋势。

1. 模型架构：从序列标注到生成模型

传统方法通常将实体抽取建模为一个序列标注问题，通过对输入中的每一个字符进行标记并整合相关标记来完成实体抽取。其中最常用的方法是基于条件随机场（Conditional Random Field，CRF）的序列标注模型 [Lafferty et al., 2001]。然而，CRF 由于其自身的语义表达能力有限，使其难以面对开放领域的嵌套、重叠以及不连续实体等复杂结构。

为了解决上述问题，有许多相关聚焦于设计面向复杂结构实体抽取的特定抽取结构。针对嵌套与重叠实体结构，Finkel 和 Manning 首次提出将依存树上的节点视为候选实体 [Finkel & Manning, 2009]。Wang 等人设计了一种超图结构，通过将嵌套以及重叠实体建模为超图中不同节点连接的方式来进行实体识别 [Wang & Lu, 2018]。Wang 等提出了一个新的基于转移的模型，该模型通过一系列特别设计的转移动作来构建嵌套实体提及 [Wang et al., 2018]。Lin 等人设计了一种基于锚点-指针网络的框架，通过将实体抽取转化为锚点与边界两步骤抽取的问题，来识别不同锚点对应的嵌套与重叠实体 [Lin et al., 2019]。针对于非连续实体的抽取，近期的工作主要聚焦于扩展常用的 BIO 标记的表达能力，并引入超图 [Dai et al., 2020]、团 [Yu et al., 2021] 等特殊结构，使得模型能够处理非连续实体抽取的问题。虽然这些方法在特定的实体抽取数据集与特定的结构上已经取得了很好的效果，但是其标记结构的适用范围较窄，且结构的设计必须能够防止产生歧义以及不一致性。然而表达能力更强、更无歧义的标注模式将不可避免地导致训练和解码过程中更高的时间复杂度。近年来，大规模预训练语言模型的兴起为解决复杂结构的实体抽取带来的新的思路。许多学者开始聚焦于将实体抽取任务与自然语言处理领域中的其它常见的任务范式—例如区块 (Span) 抽取范式与生成范式—进行对接。这类基础的范式通常较为灵活，因此可以很好地表达复杂

的实体结构。同时，通过使用这些范式建模实体抽取任务可以非常有效地利用现有的其它任务的资源，使得开放域实体抽取模型可以在仅有少量相关训练数据的情况下取得较好的性能。为此，Li 等人提出了基于阅读理解模型架构的实体抽取模型，通过将实体抽取转化为一个基于阅读理解的区块抽取任务，来统一建模各类实体抽取任务 [Li et al., 2020]。近期，Yan 与 Lu 等人则提出通过生成模型，来将实体抽取任务直接转化为生成目标实体位置或实体区块的生成任务，从而更直接地完成实体抽取[Yan et al., 2021, Lu et al., 2022]。这类模型目前已经表现出了非常好的性能，同时对数据的依赖度较低且可复用性与可迁移性较好。但是，由于这些模型通常具有较高的模型复杂度，其解码过程相比于传统的序列标注模型代价更高，因此如何设计更好的生成架构，降低解码复杂度，提升解码效率将是未来这类模型迈向实际应用过程中的重要挑战。

2. 学习算法：从粗粒度有监督学习到细粒度小样本学习

绝大部分传统的实体抽取研究集中在构建更精准的模型和方法，这些方法通常面向预先定义好的粗粒度实体类别，使用大规模标注语料训练模型参数。然而，在构建开放领域实体抽取系统时，这些有监督方法往往依赖于大规模的训练语料来提升模型性能，因此无法被用于开放类别、资源缺乏的实体抽取任务当中。近年来，有许多工作重点关注于解决实体抽取中类别开放的细粒度实体抽取与资源缺乏的小样本实体抽取两大挑战。对于开放类别的细粒度实体抽取，当前的工作主要聚焦于两条技术路线。一是利用额外的数据来学习更好的开放类别实体表示，这类工作包括有基于远距离监督的方法 [Choi et al., 2018, Onoe et al., 2021] 与基于数据增强 [Xin et al., 2018, Dai et al., 2019] 的方法，已经在许多数据集上取得了一定的效果。这些方法的最主要优势在于其主要是对数据层面进行处理，因此不需要在模型层面进行改动，使得其可以直接接入下游各种不同的实体抽取模型。但是由于额外的数据通常是通过弱监督方式构造得来的，因此必然面临着数据质量差、存在大量噪声数据的问题。此外，由于开放域实体类别众多，因此即便利用众多的外部数据也无法保证覆盖所有的实体类别。因此，如何解决数据质量与数据覆盖度的问题是这类方法所面临的核心挑战。第二类技术路线则是从模型层面下手，通过充分利用标签间的关联关系与贡献关系来完成开放类别的实体抽取。Ren 等提出利用预定的标签结构来学习更好的类别表示 [Ren et al., 2016, Xu & Barbosa, 2018, Abhishek et al., 2017]。Liu 等人则提出了一种全新的标签推理网络，通过一个生成式的框架来自动地捕捉标签间所蕴含的隐式关系 [Liu et al., 2021b]。这类方法充分地利用了类别标签间的关联信息作为辅助，有效地提升了在资源稀缺的实体类别的抽取性能。但是，由于这类方法均依赖于类别间的关联信息，而这类关联信息通常难以获取，并且需要引入一定的类别间先验假设，因此如何在更自由的类别体系当中更好地捕捉类别间的关联是这一路线

的核心挑战。学习算法层面另一个研究重点是在稀缺资源条件下的实体抽取问题，并由此衍生出了小样本实体抽取这一研究方向。小样本实体抽取通常将实体抽取的过程分为三个阶段：1）预学习，即在现有的数据以及现有的实体类别上学习得到一个较具通用性的抽取模型；2）微调，即利用新类别上的小样本数据微调上述的通用性模型，以得到一个新类别的实体识别模型；3）预测，即利用微调后的模型来进行实体抽取。这一方向上的基线方法是直接使用新类别上的小样本数据直接对模型进行训练。然而，由于训练样本数量较少，这一方法通常不能取得满意的性能。为此， 现有的小样本实体抽取方法通常可以被归纳为以下三大类别 [Huang et al., 2021]：1）基于原型学习的方法 [Ravi & Larochelle, 2017, Tian et al., 2020, Geng et al., 2019, Snell et al., 2017]， 即利用少量样本获取特定类别的原型，并利用这一原型进行实体抽取；2）基于弱监督学习的方法 [Ghaddar & Langlais, 2017]，即利用少量样本，从大规模语料库中获取更多的样本扩充训练数据，从而进行有监督学习；3）基于自学习的方法 [Xie et al., 2020]，即通过小样本学习得到一个模型，然后通过模型-数据之间的相互迭代，使得模型能够在少量标注数据和大规模无标注数据上进行自我学习。这三种方法实质上是分别从学习层面、数据层面以及模型层面来提升小样本实体抽取的性能，因此相互之间可以互补，并在近期在预训练语言模型上衍生出了基于 Prompt 的微调 [Ding et al., 2021] 等相关工作。然而，现有的小样本学习方法在实体抽取上与有监督学习方法之间仍存在着较大的差距，因此如何利用少量样本获取更好的实体抽取模型仍然是一个尚待解决的重要问题。

3. 模态融合：从单语单模到多语多模

深度学习和预训练模型为实体抽取领域带来了另一大进展是打通了不同语言与不同模态间的信息。特别是随着近年来多语言多模态预训练模型的快速兴起，不同语言不同模态的数据可以被映射到同一个语义空间中，使得跨语言跨模态之间的语义可以进行交互计算，这为多语多模实体抽取提供了坚实的基础。不同语言与模态之间的信息存在天然的互补关系，富资源的语言可以为低资源语言的实体抽取提供知识的迁移，而图像、音频等模态的信息则可以为本文的实体抽取提供额外的依据，因此，多语言多模态的实体抽取日渐成为了当下的一大研究热点。在多语言实体抽取方面，绝大多数工作遵循的核心思路是“单语标注，多语使用”，即通过充分利用富标注信息语言的标注预料，并通过多语之间的语义对齐来提升资源缺乏语言的实体抽取性能。这方面的工作主要包括数据对齐、表示对齐以及基于知识蒸馏的方法。在数据对齐方面，Tedeschi 等人提出了利用 Wikipedia 中的多语对齐信息来自动构建多语言实体抽取对齐数据的方法 [Tedeschi et al., 2021]。Liu 等人提出了一种多语言数据增强的方法来完成零样本条件下的跨语言实体抽取能力迁移 [Liu et al., 2021a]。在表示层学

习方面，大多数工作通过利用多语言神经网络或者预训练模型，将多语言的表示映射到统一空间，并在这一空间上进行实体抽取，从而使模型具有仅利用部分语言的训练数据实现多语实体抽取的能力 [Shaffer, 2021, Fan et al., 2021, Rahimi et al., 2019]。而基于多语言微调和蒸馏的方法通常首先学习一个基础的多语言模型，然后在少量样本上分别微调某个特定语言的解码参数，从而完成跨语言之间的知识迁移 [Dhamecha et al., 2021, Wang et al., 2020]。这些方法在多语言，特别是资源匮乏的语言的实体抽取任务上已经展现出了良好的效果。在多模态实体抽取方面，目前的工作主要通过引入语音或者是图像中的额外信息，从而辅助完成文本中的实体抽取。多模态融合的主要技术手段包括有表示层的融合 [Zhang et al., 2021] 以及跨模态多任务学习 [Sui et al., 2021] 等。这些工作在短文本以及不规范文本实体抽取等文本单模态存在歧义的场景中取得了明显的提升。

四、产业发展现状

实体抽取是自然语言处理最基本的技术之一。近年来，国内外众多著名人工智能厂商纷纷构建自己的人工智能开放平台和相应的开源工具，为其他行业提供人工智能服务。绝大多数相关的人工智能开放平台均对外提供有实体抽取的服务与接口，国内具有代表性的平台包括有百度 AI 开放平台、阿里灵杰、华为 AI 开放平台和腾讯 AI 开放平台等。除了在常见类别的限定域实体上，这些人工智能开放平台还根据不同的业务场景，提供了法律、金融、医疗等诸多领域的领域特定实体抽取服务，帮助相关领域企业实现数字化、智能化转型。在开源工具方面，哈工大研发的 LTP 语言技术平台、复旦大学研发的 FudanNLP、斯坦福大学研发的 CoreNLP 和 Stanza、清华大学研发的 THULAC、HanNLP 以及 spaCy 工具包等在学术界和产业界均有着较大的影响力，这些开源工具一般都内置命名实体识别模型，但是通常仅支持对基本实体类型的识别，例如人名、地名、机构名等。因此，这些开源工具对于开放领域的实体抽取的支撑仍略显不足。由于实体抽取的基础性和重要性，实体抽取技术在众多领域都有着广泛的应用。在新闻媒体领域，实体抽取技术能够帮助新闻采编工作更加高效。在法律服务领域，实体抽取技术通过识别法律文书中的法律术语等相关信息，构建法律领域知识图谱，对类案文书、法律规则、相应法条进行自动推荐，从而帮助法官从繁重的文书工作中解脱出来。在电商领域，实体抽取技术能够提取快递单据中的文本信息，并输出包含姓名、电话、地址等的结构化信息，帮助快递或电商企业提高单据处理效率。在医疗领域，实体抽取技术能识别电子病案中的医学实体，进而辅助医生及时发现病历书写中的缺陷，全面提升病历质量，帮助医院优化诊疗流程、提高诊疗效率和全面提高医疗质量。此外，实体抽取技术也是构建行业知识图谱的关键技术。例如，在汽车领域，针对汽车这种属性较多的

实体领域，汽车知识图谱可以将不同品牌、不同型号的汽车产品信息整合，为消费者提供全面的导购服务；在政务领域，利用实体抽取技术构建知识图谱可以聚合政策信息，提供统一的数据访问视图，支撑高效政务搜索和问答，提升政务处理效率；在油气勘探领域，基于勘探知识图谱可以提供丰富的油气应用，例如语义搜索、油气知识推荐等，支撑油气勘探开发、降本增效等；在各行业的客服领域，构建基于知识图谱的多轮对话系统，可以分析用户对话中的实体和关系，根据实体和关系进行知识图谱的查询和推理，从而选择相应的对话策略，减少人工成本，提高工作效率等。

五、总结与展望

实体抽取是自然语言处理与知识图谱领域的基础性技术。近年来，实体抽取领域逐渐从限定领域迈向开放领域，由此面临着类别开放、实体结构复杂、标注资源缺乏等的重要挑战。近年来，深度学习和大规模预训练语言模型的兴起已经为实体抽取领域带来了范式级别的改变。这种深刻的改变体现在模型架构、学习算法与模态融合等多个层面，并显著地改变了实体抽取领域的技术发展路线。然而在当下，开放领域的实体抽取仍然面临着诸多挑战，大模型对于实体抽取领域的影响还远远没有完全展现出来。在未来，如何设计出更通用、有效、高速的模型架构，如何更充分地利用现有资源，使得实体抽取模型具有更快速的跨类别泛化能力，如何更好的实现多模态多语言的打通融合，都将是实体抽取领域在大模型时代所面临的重要挑战。

参考文献

[Abhishek et al., 2017] Abhishek Abhishek, Ashish Anand, Amit Awekar. Fine-grained entity type classification by jointly learning representations and label embeddings. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 797–807, Valencia, Spain, 2017.

[Chinchor & Robinson, 1998] N. Chinchor, P. Robinson. MUC-7 named entity task definition (version 3.5). In Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998, 1998.

[Choi et al., 2018] Eunsol Choi, Omer Levy, Yejin Choi, Luke Zettlemoyer. Ultra-fine entity typing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 87–96, Melbourne, Australia, 2018.

[Dai et al., 2019] Hongliang Dai, Donghong Du, Xin Li, Yangqiu Song. Improving fine-grained

entity typing with entity linking. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6210–6215, Hong Kong, China, 2019.

[Dai et al., 2020] Xiang Dai, Sarvnaz Karimi, Ben Hachey, Cecile Paris. An effective transition-based model for discontinuous NER. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5860–5870, Online, 2020.

[Dhamecha et al., 2021] Tejas Dhamecha, Rudra Murthy, Samarth Bharadwaj, Karthik Sankaranarayanan, Pushpak Bhattacharyya. Role of Language Relatedness in Multilingual Fine-tuning of Language Models: A Case Study in Indo-Aryan Languages. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8584–8595, Online and Punta Cana, Dominican Republic, 2021.

[Ding et al., 2021] Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Haitao Zheng, Zhiyuan Liu, Juan-Zi Li, Hong-Gee Kim. Prompt-learning for fine-grained entity typing. ArXiv, abs/2108.10604, 2021.

[Fan et al., 2021] Yimin Fan, Yaobo Liang, Alexandre Muzio, Hany Hassan, Houqiang Li, Ming Zhou, Nan Duan. Discovering representation sprachbund for multilingual pre-training. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 881–894, Punta Cana, Dominican Republic, 2021.

[Finkel & Manning, 2009] Jenny Rose Finkel, Christopher D. Manning. Nested named entity recognition. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 141–150, Singapore, 2009.

[Geng et al., 2019] Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, Jian Sun. Induction networks for few-shot text classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3904–3913, Hong Kong, China, 2019.

[Ghaddar & Langlais, 2017] Abbas Ghaddar, Phillippe Langlais. WiNER: A Wikipedia annotated corpus for named entity recognition. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 413–422, Taipei, Taiwan, 2017.

[Huang et al., 2021] Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, Jiawei Han. Few-shot namedentity recognition: An empirical baseline study. In Proceedings of the 2021 Conference on Empirical

Methods in Natural Language Processing, pages 10408–10423, Online and Punta Cana, Dominican Republic, 2021.

[Jain et al., 2020] Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, Iz Beltagy. SciREX: A challenge dataset for document-level information extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7506– 7516, Online, 2020.

[Lafferty et al., 2001] John D. Lafferty, Andrew McCallum, Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, page 282–289, San Francisco, CA, USA, 2001.

[Li et al., 2016] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. Database: The Journal of Biological Databases and Curation, 2016, 2016.

[Li et al., 2020] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, Jiwei Li. A unified MRC framework for named entity recognition. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5849–5859, Online, 2020.

[Lin et al., 2019] Hongyu Lin, Yaojie Lu, Xianpei Han, Le Sun. Sequence-to-nuggets: Nested entity mention detection via anchor-region networks. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5182–5192, Florence, Italy, 2019.

[Ling & Weld, 2012] Xiao Ling, Daniel S. Weld. Fine-grained entity recognition. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, AAAI'12, page 94– 100, 2012.

[Liu et al., 2021a] Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, Chunyan Miao. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5834–5846, Online, 2021a.

[Liu et al., 2021b] Qing Liu, Hongyu Lin, Xinyan Xiao, Xianpei Han, Le Sun, Hua Wu. Fine-grained entity typing via label reasoning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4611–4622, Online and Punta Cana, Dominican Republic, 2021b.

[Lu et al., 2022] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun,

Hua Wu. Unified structure generation for universal information extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5755–5772, Dublin, Ireland, 2022.

[Onoe et al., 2021] Yasumasa Onoe, Michael Boratko, Andrew McCallum, Greg Durrett. Modeling fine-grained entity types with box embeddings. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2051–2064, Online, 2021.

[Rahimi et al., 2019] Afshin Rahimi, Yuan Li, Trevor Cohn. Multilingual ner transfer for low-resource languages. CoRR, abs/1902.00193, 2019.

[Ravi & Larochelle, 2017] Sachin Ravi, H. Larochelle. Optimization as a model for few-shot learning. In International Conference on Learning Representations, 2017.

[Ren et al., 2016] Xiang Ren, Wenqi He, Meng Qu, Lifu Huang, Heng Ji, Jiawei Han. AFET: Automatic fine-grained entity typing by hierarchical partial-label embedding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1369– 1378, Austin, Texas, 2016.

[Shaffer, 2021] Kyle Shaffer. Language clustering for multilingual named entity recognition. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 40–45, Punta Cana, Dominican Republic, 2021.

[Snell et al., 2017] Jake Snell, Kevin Swersky, Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30, 2017.

[Sui et al., 2021] Dianbo Sui, Zhengkun Tian, Yubo Chen, Kang Liu, Jun Zhao. A large-scale Chinese multimodal NER dataset with speech clues. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2807–2818, Online, 2021.

[Tedeschi et al., 2021] Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, Roberto Navigli. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2521–2533, Punta Cana, Dominican Republic, 2021.

[Tian et al., 2020] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In Andrea Vedaldi,

Horst Bischof, Thomas Brox, Jan-Michael Frahm, editors, Computer Vision – ECCV 2020, pages 266–282, Cham, 2020.

[Wang & Lu, 2018] Bailin Wang, Wei Lu. Neural segmental hypergraphs for overlapping mention recognition. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 204–214, Brussels, Belgium, 2018.

[Wang et al., 2018] Bailin Wang, Wei Lu, Yu Wang, Hongxia Jin. A neural transition-based model for nested mention recognition. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1011–1017, Brussels, Belgium, 2018.

[Wang et al., 2020] Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, Kewei Tu. Structure-level knowledge distillation for multilingual sequence labeling. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3317–3330, Online, 2020.

[Xie et al., 2020] Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, Quoc V. Le. Self-training with noisy student improves imagenet classification. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10684–10695, 2020.

[Xin et al., 2018] Ji Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun. Improving neural fine-grained entity typing with knowledge attention. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, 2018.

[Xu & Barbosa, 2018] Peng Xu, Denilson Barbosa. Neural fine-grained entity type classification with hierarchy-aware loss. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 16–25, New Orleans, Louisiana, 2018.

[Yan et al., 2021] Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, Xipeng Qiu. A unified generative framework for various NER subtasks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5808–5822, Online, 2021.

[Yu et al., 2021] Bowen Yu, Zhenyu Zhang, Jiawei Sheng, Tingwen Liu, Yubin Wang, Yucheng Wang, Bin Wang. Semi-open information extraction. In Proceedings of the Web Conference 2021, WWW '21, page 1661–1672, New York, NY, USA, 2021.

[Zhang et al., 2021] Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu,

Guodong Zhou. Multi-modal graph fusion for named entity recognition with targeted visual guidance. Proceedings of the AAAI Conference on Artificial Intelligence, 35(16):14347– 14355, 2021.

第四章 实体关系抽取

曾道建[1] 陈玉博[2] 刘康[2]

1. 湖南师范大学 信息科学与工程学院，长沙 410081

2. 中国科学院自动化研究所 模式识别国家重点实验室，北京 100190

一、任务定义、目标和研究意义

关系定义为两个或多实体之间的某种联系，实体关系抽取是检测和识别出实体之间具有的某种语义关系，并将结果以结构化的形式存储 [Martinez-Rodriguez et al., 2020]。例如：给定文本“华扬联众数字技术股份有限公司于 2017 年 8 月 2 日在上海证券交易所上市。”，通过实体关系抽取可以得到三元组 < 华扬联众数字技术股份有限公司，上市时间，2017 年 8 月 2 日 >，< 华扬联众数字技术股份有限公司，上市地点，上海证券交易所上市 >。实体关系抽取是信息抽取中的一个关键环节，具有重要的理论意义和广阔的应用前景。在理论方面，实体关系抽取涉及到自然语言处理、机器学习、逻辑推理、数据挖掘等多个学科的理论和方法，实体关系抽取不仅能得到结构化知识，而且对相关学科理论的完善和发展也将产生积极的促进作用；在应用方面，实体关系抽取可以为大规模知识图谱的构建提供核心技术，是实现文本从语法分析到语义分析的关键环节，同时也是智能信息服务的关键支撑，将促进以知识为核心的信息检索、智能问答、人机交互和海量数据管理等多个研究方向的快速发展，进而推进互联网相关产业的进一步发展。

二、研究内容与挑战

实体关系抽取是信息抽取领域的一个经典任务，根据抽取数据来源，主要可以分为结构化、半结构化、非结构实体关系抽取三类。目前研究工作主要针对抽取难度较大的非结构化文本展开，研究内容主要涉及语义关系表征、抽取数据处理、复杂关系建模等三方面，以下分别介绍具体研究内容：

语义关系表征：主要研究如何用特征来表示实体之间的语义关系，具体研究内容包括基于规则的方法和统计机器学习方法，统计方法又可分为特征向量、核函数、深度学习自动特征学习等。

抽取数据处理：主要研究如何处理不同类型的关系抽取数据，具体研究内容包括远程监督标注数据噪声处理、小样本关系抽取、数据隐私保护以及如何从预训练语言模型中抽取知识等。

复杂关系建模：主要研究如何处理实际应用场景中的复杂关系，具体研究内容包括文档、对话、多模态等复杂场景下的关系抽取、多元关系抽取、自动发现实体间的新型关系等。

实体关系抽取目前主要面临如下三个挑战：

1）自然语言表达的多样性：关系抽取的核心是将自然语言表达的关系知识映射到关系三元组上。然而，自然语言表达具有多样性和隐含性，导致关系抽取任务极具挑战性。自然语言表达的多样性指的是同一种关系可以有多种表达方式，例如“总部位置”这个语义关系可以用“X 的总部位于 Y”，“X 总部坐落于 Y”，“作为 X 的总部所在地，Y…”等等不同的文本表达方式。自然语言表达的多样性是关系抽取的一大挑战。

2）关系表达的隐含性：关系表达的隐含性是指关系有时候在文本中找不到任何明确的标识，关系隐含在文本中。例如：蒂姆·库克与中国移动董事长奚国华会面商谈“合作事宜”，透露出了他将带领苹果公司进一步开拓中国市场的讯号。在这一段文本中，并没有直接给出蒂姆·库克和苹果公司的关系，但是从“带领苹果公司”的表达，我们可以推断出蒂姆·库克是苹果公司的首席执行官 (CEO)。关系表达的隐含性是关系抽取的一大挑战。

3）实体关系的复杂性：关系抽取的目标是抽取实体之间的语义关系，然而，真实世界中同一对实体之间可能有多个关系，而且有的关系可以同时存在，而有的关系是具有时间特性的。比如：中国和北京的关系有多个，北京坐落于中国，北京是中国的首都，北京是中国的政治中心，北京是中国的文化中心。这些关系是可以同时存在的。但是如果两个人本来是夫妻关系，后来离婚了，他们就不是夫妻关系了，是前妻或者前夫的关系，这个类关系具有时空性，不能单独存在，实体关系的复杂性是关系抽取的又一挑战。

三、技术方法和研究现状

实体关系抽取在过去的 20 多年里都有持续研究，主要以 MUC、ACE、SemEval、KBP 等评测会议提出的任务展开，其技术方法也由人工标注语料、基于“特征工程”的机器学习方法发展到利用远程监督自动标注语料、机器自动学习特征的深度学习方法 [Zeng et al., 2014]，深度神经网络特别是 BERT[Devlin et al., 2019]、GPT[Radford et al., 2018] 等预训练语言模型为实体语义关系抽取带来了新的突破，与传统的非神经网络方法相比性能显著提升，为自动构建大规模知识图谱奠定带来了曙光，受到学术界和工业界的广泛关注，近年来在语义关系表征、抽取数据处理和复杂关系建模等研究方向上涌现出一大批新的工作，以下分别介绍具体：

1. 语义关系表征

目前，利用神经网络自动学习表征实体语义关系的特征是一种非常有效的方法，已得到

研究者共识。早期工作主要采用流水线的方法，即先进行实体识别后语义关系分类，Zeng 等[Zeng et al., 2014] 尝试使用卷积神经网络自动学习语义关系分类特征，之后研究人员陆续将关系表示涉及的句法结构等信息引入进来，进一步提升了语义关系抽取的性能。流水线方式忽视了实体识别和关系分类两个任务之间的关联性，并且不可避免地存在实体识别模块错误传递。针对此问题，Li 等 [Li & Ji, 2014] 最早提出使用联合模型捕获语义关系之间错综复杂的关联，并通过实验验证了联合抽取的可行性。

最近几年实体语义关系表征方向上的研究热点是实体关系联合抽取，基本出发点是利用实体识别任务帮助学习更好的语义关系特征。联合抽取又分为序列标注、表填充和序列生成等三种方法。序列标注方法通常在循环神经网络或者预训练语言模型基础上接一个命名实体识别序列标注网络，然后再接一个关系分类的网络。例如，Miwa 等 [Miwa & Bansal, 2016] 等首先使用长短记忆网络编码输入的句子，然后通过序列标注进行实体识别，最后考虑实体在依存句法树上的路径对检测到的实体进行关系分类，模型训练时利用实体标签和关系标签联合更新网络参数。Katiyar 等 [Katiyar & Cardie, 2017] 针对 Miwa 等所提方法依赖依存句法分析的问题，使用注意力机制帮助捕获实体对的语义关系特征，取得了更好的效果。Zheng 等 [Zheng et al., 2017] 提出了一种新的标注策略，将实体识别和关系分类任务融入标注策略，达到联合的目的，但是该标注策略无法处理三元组重叠的问题。Takanobu 等[Takanobu et al., 2019] 使用分层的强化学习标注框架来增强实体和关系之间的交互性，整个抽取的过程被分解为高层和低层并分别用于关系判定和实体抽取。Fu 等 [Fu et al., 2019] 提出两阶段图的方法，第一阶段使用多任务的方式找到实体和所有可能的关系得分，第二阶段构建实体关系图建模实体和关系之间的交互，实验结果显示对重叠关系的预测比以前的序列方法有较大的改进。Wei 等 [Wei et al., 2020] 提出了层级二值标注框架，首先通过序列标注的方式得到头实体边界，然后每种关系使用一个二值序列标注器找到头实体在此关系中对应的尾实体。上述方法共同特点是实体识别和关系抽取任务共享同一个网络编码，Zhong 等[Zhong & Chen, 2021] 认为命名实体识别和关系抽取表示特征应该不一样，底层共享一个表示层会限制模型的表达能力，提出了两个编码器组成的模型。

表填充方法最早由 Miwa 等 [Miwa & Sasaki, 2014] 提出，他们将句子中的词看作矩阵的横纵坐标，实体识别转换为填充表格的对角线元素，关系分类任务是填充上三角或者下三角矩阵，然后使用分类器填充表格元素。Gupta 等 [Gupta et al., 2016] 进一步使用循环神经网络依次填充表格，建模表格之间的依赖关系，从而捕获三元组之间的交互。Zhang 等[Zhang et al., 2017] 利用句法信息全局优化表格填充帮助更好地进行关系抽取。Adel 等[Adel & Schütze, 2017] 根据实体的位置将句子分为三段，同时预测关系和实体的类型，利用条件随

机场模型建模实体类型与关系之间的依赖关系。上述表格填充方法要求每个元素只能填充一个元素，无法处理三元组重叠的问题，受序列标注方法的启发，Bekoulis 等 [Bekoulis et al., 2018] 提出基于多头选择的方法，该方法单独使用序列标注层检测实体，然后允许每个词在表格中选择多个词构成三元组，从而解决了重叠三元组抽取的问题。上述方法在表填充时使用多任务学习的思路，将填充过程分成了两个阶段，存在暴露偏置问题。Wang 等 [Wang et al., 2020] 使用单阶段解码，将抽取框架统一为字符对链接问题，同时解决重叠关系和暴露偏置问题。

序列生成最早由 Zeng 等 [Zeng et al., 2018b] 提出，将联合抽取问题看作是一个序列到序列生成问题，使用带拷贝机制的编码器-解码器模型 CopyNet 解决此问题，解码时通过从原句子中拷贝实体和预测关系得到三元组。CopyNet 存在着无法处理由多个词构成的实体的问题，之后一系列改进序列到序列的模型相继被提出，Zeng 等 [Zeng et al., 2020a] 在其编码器端增加序列标注模块识别实体，Nayak 等 [Nayak & Ng, 2020] 提出新的解码策略从而避免无法处理词构成的实体。Ye 等 [Ye et al., 2020] 使用生成式 Transformer 并利用对比学习的方式训练模型，进一步提升语义关系特征的有效性。上述基于生成的方法使用自回归的方法解码，无法避免模型存在的暴露偏置问题。为此，Zhang 等 [Zhang et al., 2020a] 提出一种树状解码的策略，使得解码长度不依赖于三元组的个数，有效减轻了暴露偏置的影响。Sui 等 [Sui et al., 2021] 将联合抽取进一步看作是序列到集合问题，使用非自回归方法解码，彻底消除暴露偏置的存在。

2. 抽取数据处理

目前性能占据主导地位的神经网络实体关系抽取是典型的“数据饥渴”模型，需要大量高质量的标注数据，而人工标注数据费时费力、一致性差。为此，研究人员提出远程监督关系抽取。从该任务诞生开始就被广泛关注的数据噪声问题最近几年仍然是研究重点，Zeng 等 [Zeng et al., 2015] 先利用分段卷积神经网络学习每个句子的表示，然后使用多示例学习避免噪声的干扰。Lin 等 [Lin et al., 2016] 提出只选取每个包中一个句子作为包的表示会丢失信息，提出使用注意力机制对包中的示例进行加权得到包的表示向量。Jiang 等 [Jiang et al., 2016] 通过对包内所有的句子做最大池化操作，提取出示例之间的隐藏关联，并且针对实体对之间可能存在多种关系的问题，设计了一种多标签损失函数，使用 Sigmoid 计算每一个类别的概率，然后判断该包是否可能包含该类别。Zeng 等 [Zeng et al., 2018a] 利用强化学习抽取包中每个句子的关系，然后使用句子中的关系帮助包中关系的确定。Ma 等 [Ma et al., 2021] 采用负样本学习的方法，直接找出并过滤噪声样本。基于多示例学习的方法可以减轻数据的噪声，但是包中句子中很多其他有益的信息未被关注到，Chen 等 [Chen et al., 2021]

以句子为单位，使用示例对比学习的方法挖掘其中丰富的语义信息。上述方法主要针对错误正样本展开，由于知识库的不完备性，远程监督还面临着错误负样本类噪声，研究者也从正样本-未标注样本学习角度展开了关系抽取的研究 [Yang et al., 2019, Xie et al., 2021]。

远程监督为高效收集训练数据开启了新的纪元，但是真实场景中长尾知识而言，仍难以通过远程监督机制来得到训练实例。针对上述问题，Han 等 [Han et al., 2018] 首次将小样本学习引入到关系抽取，构建了小样本关系抽取数据集 FewRel，之后基于混合注意力机制的原型网络 [Gao et al., 2019a]、多级匹配和整合策略 [Ye & Ling, 2019]、预训练语言模型 [Baldini Soares et al., 2019]、基于贝叶斯的元学习 [Qu et al., 2020] 等方法相继被提出来完成该任务。Gao 等 [Gao et al., 2019b] 在 FewRel 基础上增加领域迁移和“以上都不是”检测任务，提出了 FewRel 2.0 数据集。另外，很多领域的数据隐私性要求极高（例如：金融、医疗、安全、军事等领域），无法直接获取数据。同时，针对真实应用场景中数据管理与隐私保护的要求日益严格，而现有方法的训练过程需要暴露大量数据。为了解决这一矛盾，Sui 等 [Sui et al., 2020a] 提出了联邦远程监督关系抽取任务，利用懒惰多示例学习算法通过跨平台之间的协作缓解联邦远程监督关系抽取中的数据噪声问题，并利用基于集成蒸馏的联邦训练框架降低联邦学习中的通信开销，增加了基于大规模预训练语言模型的关系抽取方法在联邦设定下的实用性[Sui et al., 2020b]。传统的实体关系抽取研究主要面向非结构化文本，近年来，随着大规模预训练语言模型的快速发展，研究者认为预训练的语言模型 (如 BERT 等) 中，除包含的语言学知识外还包含了事实性的知识，因此可以将预训练语言模型当作一个现成的、开放的知识库。

Petroni 等 [Petroni et al., 2019] 对语言模型记忆知识的能力进行了探测，针对该问题提出了语言模型分析（LAMA）任务，并基于多个知识源手工创建了单个词语的完形填空数据集。由于 LAMA 的查询生成过程非常简单，Jiang 等 [Jiang et al., 2021] 认为 LAMA 只是测量了语言模型所知道的下限，并提出了更高级的方法来生成更高效的查询，进一步挖掘模型提取知识的能力。Roberts 等 [Roberts et al., 2020] 使用了一种更具有难度的闭卷问答任务，让模型先在相关数据集上微调，在微调过程中模型需要学习如何挖掘之前预训练获得的知识并加以利用，实验表明预训练语言模型不仅存储了大量的知识，并且可以将这些知识迁移到下游任务中。Verga 等 [Verga et al., 2021] 在 BERT 架构基础上加入了一个实体记忆模块和事实记忆模块，通过加入对实体、关系和三元组事实知识的编码信息来增强文本表示，在一定程度上模块化地将模型中存储的事实知识分离出来。

3. 复杂关系建模

传统的关系抽取主要处理的是简单关系，复杂关系抽取试图提取涉及多个实体或在特定

约束下的更复杂关系，该方向的研究目前呈现百花齐放状态，包括文档级、多元关系、跨文档、增量式、多模态抽取等多个研究点。为了推动文档级关系抽取的研究，Yao 等 [Yao et al., 2019]提出了一个人工标注的大规模文档级语义关系抽取数据集 DocRED。 Christopoulou 等[Christopoulou et al., 2011]利用以边为中心的图神经网络建模跨句之间的实体交互。Nan 等 [Nan et al., 2020] 使用图神经网络学习文档地潜在结构，逐步汇总多跳信息进行语义关系推理。Zeng 等[Zeng et al., 2020b]使用两个图网络结构来实现语义关系抽取，一个图用于特征传播，另外一个用于关系推理。除了使用图网络外，研究者也开始尝试直接使用大规模语言模型建模文档，Zhou 等 [Zhou et al., 2021] 提出自适应阈值代替用于多标签分类的全局阈值，并直接利用预训练模型的自注意力得分找到有助于确定关系的相关上下文特征。上述方法主要关注文档中的二元关系，近年来也有工作探索多元关系抽取，Song 等 [Song et al., 2018] 提出基于图 LSTM 的关系抽取网络抽取多个句子中存在的多元关系，Jia 等 [Jia et al., 2019] 提出多尺度神经结构进行多元关系抽取，所用方法同时考虑了不同尺度的文本跨度和不同子关系的学习表示。

除文档抽取外，实际应用中，还面临从多个文档、多轮对话等场景抽取关系。为此，Yao 等 [Yao et al., 2021] 提出了一个新的跨文档抽取任务，并发布了数据集 CodRED。Zhang 等 [Zhang et al., 2020b] 探索了如何在实际的医疗对话中抽取出症状、检查、手术、一般信息及其相应的状态。另外，现有关系抽取任务设定一般假设有预先定义好的封闭关系集合，实体间的新型关系无法被有效获取。为解决关系抽取中的超出预定义关系的问题，Cui 等[Cui et al., 2021] 提出基于关系原型表示的持续关系抽取方法。Zhao 等 [Zhao et al., 2021] 在预定义关系数据集上预训练，然后通过最小化标记数据和未标记数据上的联合目标完成未标记数据聚类，最后进行增量式学习。

随着面向文本的关系抽取技术成熟，研究者也开始探索多模态关系抽取方法。Wan 等 [Wan et al., 2021] 提出基于小样本学习的方法，同时利用文本和面部图像进行社会关系抽取，并发布了由四部经典名著和相应的电视剧组成的多模态数据。Zheng 等 [Wan et al., 2021] 构造了一个多模态的关系分类数据集，给定图像和单句及两个实体进行关系分类，并验证了可以通过视觉信息帮助纯文本的关系分类。

四、发展趋势

实体关系抽取技术研究蓬勃发展，已经成为了信息抽取和自然语言处理的重要分支。这一方面得益于系列国际权威评测和会议的推动，如消息理解系列会议（MUC，Message Understanding Conference），自动内容抽取评测（ACE，Automatic Content Extraction）和文本

分析会议系列评测（TAC，Text Analysis Conference）。另一方面也是因为实体关系抽取技术的重要性和实用性，使其同时得到了研究界和工业界的广泛关注。实体关系抽取技术自身的发展也大幅度推进了中文信息处理研究的发展，迫使研究人员面向实际应用需求，开始重视之前未被发现的研究难点和重点。纵观实体关系抽取研究发展的态势和技术现状，本文认为实体关系抽取的发展方向如下：

1. 新类别/开放类别上的小样本学习能力

目前的小样本学习设定需要用一个巨大的训练集训练的，测试时只给出 N-way K-shot，在这 N*K 个样本上学习并预测，真实场景下的小样本学习不存在巨大的训练集。此外，真实应用中还需要考虑如何自动发现新类别，迫切需要利用小样本实现模型在新类别关系上的快速训练模型。从 GPT3 开始，预训练-提示 (Prompt) 学习范式受到研究者的关注，该范式将下游任务也建模成语言模型任务，在只给出几条或几十条样本作为训练集，借助与大规模预训练语言模型中蕴含的大量知识，取得了不错的小样本学习效果。此外，相对于传统的 Pretrain+Finetune 范式，Prompt 可以摆脱指数级的预训练参数量对巨大计算资源的需求，高效的利用预训练模型。基于上述分析，本文认为实体关系抽取发展方向之一是利用预训练—提示学习范式进行高效的新类别/开放类别上的小样本学习。具体包括：1）开放类别语义标签自动生成与新类别的挂载；2）提示学习中关系抽取任务模板的设计与自动学习；3）预训练-提示学习范式进行实体关系抽取的理论分析。

2. 数据隐私保护下的关系可信抽取

目前性能较好的实体关系抽取模型主要是基于有监督学习或者远程监督学习的。此类模型需要将大规模的标注数据集中暴露给模型。但是在金融、医疗、安全、军事等应用场景中，数据管理与隐私保护的要求日益严格，因此如何实现数据隐私保护下的实体关系抽取模型高效训练是目前技术在真实应用场景中落地的主要挑战之一。基于上述分析，本文认为实体关系抽取的发展方向之一是数据隐私保护下的关系可信抽取。具体包括：1）数据隐私保护下的大规模实体关系抽取数据自动生成；2）含噪数据下的实体关系抽取模型鲁棒性训练；3）数据隐私保护下的实体关系抽取模型高效训练。

3. 多模态关系抽取

目前实体关系抽取主要针对的是纯文本数据，而常见的文档具有多样的布局且包含丰富的信息，以富文本文档的形式呈现包含大量的多模态信息。从认知科学的角度来说，人脑的感知和认知过程是跨越多种感官信息的融合处理，如人可以同时利用视觉和听觉信息理解说话人的情感、可以通过视觉信息补全文本中的缺失信息等，实体关系抽取技术的进一步发展

也应该是针对多模态的富文档。基于上述分析，本文认为实体关系抽取的发展方向之一是多模态信息的融合。具体包括：1）面向关系的多模态预训练模型的设计；2）多模态信息抽取框架中跨模态对齐任务设计；3）多模态信息的提取和表示。

4. 数据驱动和知识驱动融合

现有的神经网络实体关系抽取方法依靠深度学习以数据驱动的方式得到各种语义关系的统计模式，其优势在于能从大量的原始数据中学习相关特征，比较容易利用证据和事实，但是忽略了怎样融合专家知识。单纯依靠神经网络进行实体关系抽取，到一定准确率之后，就很难再改进。从人类进行知识获取来看，很多决策的时候同时要使用先验知识以及证据。数据驱动和知识驱动结合是模拟人脑进行信息抽取的关键挑战。基于上述分析，本文认为信息抽取的发展方向之一是构建数据驱动和知识驱动融合抽取技术。具体包括：1）神经符号学习信息抽取框架的构建；2）学习神经网络到逻辑符号的对应关系；3）神经网络对于符号计算过程进行模拟。

参考文献

[Adel & Schütze, 2017] Heike Adel, Hinrich Schütze. Global normalization of convolutional neural networks for joint entity and relation classification. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1723–1729, Copenhagen, Denmark, 2017.

[Baldini Soares et al., 2019] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2895–2905, Florence, Italy, 2019.

[Bekoulis et al., 2018] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, Chris Develder. Joint entity recognition and relation extraction as a multi-head selection problem. Expert Systems with Applications, 114, 2018.

[Chen et al., 2021] Tao Chen, Haizhou Shi, Siliang Tang, Zhigang Chen, Fei Wu, Yueting Zhuang. CIL: Contrastive instance learning framework for distantly supervised relation extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6191–6200, Online, 2021.

[Christopoulou et al., 2019] Fenia Christopoulou, Makoto Miwa, Sophia Ananiadou. Connecting

the dots: Document-level neural relation extraction with edge-oriented graphs. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP- IJCNLP), pages 4925–4936, Hong Kong, China, 2019.

[Cui et al., 2021] Li Cui, Deqing Yang, Jiaxin Yu, Chengwei Hu, Jiayang Cheng, Jingjie Yi, Yanghua Xiao. Refining sample embeddings with relation prototypes to enhance continual relation extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 232–243, Online, 2021.

[Devlin et al., 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171–4186, 2019.

[Fu et al., 2019] Tsu-Jui Fu, Peng-Hsuan Li, Wei-Yun Ma. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1409–1418, Florence, Italy, 2019.

[Gao et al., 2019a] Tianyu Gao, Xu Han, Zhiyuan Liu, Maosong Sun. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19, 2019a.

[Gao et al., 2019b] Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, Jie Zhou. FewRel 2.0: Towards more challenging few-shot relation classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6250–6255, Hong Kong, China, 2019b.

[Gupta et al., 2016] Pankaj Gupta, Hinrich Schütze, Bernt Andrassy. Table filling multi-task recurrent neural network for joint entity and relation extraction. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2537–2547, Osaka, Japan, 2016.

[Han et al., 2018] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, Maosong Sun. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art

evaluation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4803–4809, Brussels, Belgium, 2018.

[Jia et al., 2019] Robin Jia, Cliff Wong, Hoifung Poon. Document-level n-ary relation extraction with multiscale representation learning. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3693–3704, Minneapolis, Minnesota, 2019.

[Jiang et al., 2016] Xiaotian Jiang, Quan Wang, Peng Li, Bin Wang. Relation extraction with multi-instance multi-label convolutional neural networks. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1471–1480, Osaka, Japan, 2016.

[Jiang et al., 2021] Zhengbao Jiang, Jun Araki, Haibo Ding, Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. Transactions of the Association for Computational Linguistics, 9:962–977, 2021.

[Katiyar & Cardie, 2017] Arzoo Katiyar, Claire Cardie. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 917–928, Vancouver, Canada, 2017.

[Li & Ji, 2014] Qi Li, Heng Ji. Incremental joint extraction of entity mentions and relations. In ACL (1), pages 402–412, 2014.

[Lin et al., 2016] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, Maosong Sun. Neural relation extraction with selective attention over instances. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2124–2133, Berlin, Germany, 2016.

[Ma et al., 2021] Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Xuanjing Huang, Yaqian Zhou. SENT: Sentence-level distant relation extraction via negative training. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6201–6213, Online, 2021.

[Martinez-Rodriguez et al., 2020] Jose L Martinez-Rodriguez, Aidan Hogan, Ivan LopezArevalo. Information extraction meets the semantic web: a survey. Semantic Web, 11(2):255–335, 2020.

[Miwa & Bansal, 2016] Makoto Miwa, Mohit Bansal. End-to-end relation extraction using LSTMs

on sequences and tree structures. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1105–1116, Berlin, Germany, 2016.

[Miwa & Sasaki, 2014] Makoto Miwa, Yutaka Sasaki. Modeling joint entity and relation extraction with table representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1858–1869, Doha, Qatar, 2014.

[Nan et al., 2020] Guoshun Nan, Zhijiang Guo, Ivan Sekulic, Wei Lu. Reasoning with latent structure refinement for document-level relation extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1546–1557, Online, 2020.

[Nayak & Ng, 2020] Tapas Nayak, Hwee Tou Ng. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. Proceedings of the AAAI Conference on Artificial Intelligence, 34:8528–8535, 2020.

[Petroni et al., 2019] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander Miller. Language models as knowledge bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473, Hong Kong, China, 2019.

[Qu et al., 2020] Meng Qu, Tianyu Gao, Louis-Pascal AC Xhonneux, Jian Tang. Few-shot relation extraction via bayesian meta-learning on relation graphs. In International Conference on Machine Learning, 2020.

[Radford et al., 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, 2018. Improving language understanding by generative pre-training (2018).

[Roberts et al., 2020] Adam Roberts, Colin Raffel, Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5418–5426, Online, 2020.

[Song et al., 2018] Linfeng Song, Yue Zhang, Zhiguo Wang, Daniel Gildea. N-ary relation extraction using graph-state LSTM. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2226–2235, Brussels, Belgium, 2018.

[Sui et al., 2020a] Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao. Distantly supervised relation extraction in federated settings. arXiv preprint arXiv:2008.05049, 2020a.

[Sui et al., 2020b] Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuantao Xie, Weijian Sun. Feded: Federated learning via ensemble distillation for medical relation extraction. In Proceedings of the

2020 conference on empirical methods in natural language processing (EMNLP), pages 2118–2128, 2020b.

[Sui et al., 2021] Dianbo Sui, Chenhao Wang, Yubo Chen, Kang Liu, Jun Zhao, Wei Bi. Set generation networks for end-to-end knowledge base population. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9650–9660, Online and Punta Cana, Dominican Republic, 2021.

[Takanobu et al., 2019] Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, Minlie Huang. A hierarchical framework for relation extraction with reinforcement learning. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19, 2019.

[Verga et al., 2021] Pat Verga, Haitian Sun, Livio Baldini Soares, William Cohen. Adaptable and interpretable neural MemoryOver symbolic knowledge. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3678–3691, Online, 2021.

[Wan et al., 2021] Hai Wan, Manrong Zhang, Jianfeng Du, Ziling Huang, Yufei Yang, Jeff Z Pan. Fl-msre: A few-shot learning based approach to multimodal social relation extraction. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 13916– 13923, 2021.

[Wang et al., 2020] Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, Limin Sun. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In Proceedings of the 28th International Conference on Computational Linguistics, pages 1572–1582, Barcelona, Spain (Online), 2020.

[Wei et al., 2020] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, Yi Chang. A novel cascade binary tagging framework for relational triple extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1476–1488, Online, 2020.

[Xie et al., 2021] Chenhao Xie, Jiaqing Liang, Jingping Liu, Chengsong Huang, Wenhao Huang, Yanghua Xiao. Revisiting the negative data of distantly supervised relation extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3572–3581, Online, 2021.

[Yang et al., 2019] Kaijia Yang, Liang He, Xin-yu Dai, Shujian Huang, Jiajun Chen. Exploiting noisy data in distant supervision relation classification. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguis- tics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3216–3225, Minneapolis, Minnesota, 2019.

[Yao et al., 2021] Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, Maosong Sun. CodRED: A cross-document relation extraction dataset for acquiring knowledge in the wild. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4452–4472, Online and Punta Cana, Dominican Republic, 2021.

[Yao et al., 2019] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 764–777, Florence, Italy, 2019.

[Ye et al., 2020] Hongbin Ye, Ningyu Zhang, Shumin Deng, Mosha Chen, Chuanqi Tan, Fei Huang, Huajun Chen. Contrastive triple extraction with generative transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 14257–14265, 2020.

[Ye & Ling, 2019] Zhi-Xiu Ye, Zhen-Hua Ling. Multi-level matching and aggregation network for few-shot relation classification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2872–2881, Florence, Italy, 2019.

[Zeng et al., 2015] Daojian Zeng, Kang Liu, Yubo Chen, Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of EMNLP, pages 1753–1762, 2015.

[Zeng et al., 2014] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao. Relation classification via convolutional deep neural network. In Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, pages 2335– 2344, 2014.

[Zeng et al., 2020a] Daojian Zeng, Ranran Haoran Zhang, Qianying Liu. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. Proceedings of the AAAI Conference on Artificial Intelligence, 34:9507–9514, 2020.

[Zeng et al., 2020b] Shuang Zeng, Runxin Xu, Baobao Chang, Lei Li. Double graph based reasoning for document-level relation extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1630–1640, Online, 2020b.

[Zeng et al., 2018a] Xiangrong Zeng, Shizhu He, Kang Liu, Jun Zhao. Large scaled relation

extraction with reinforcement learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 5658–5665, 2018a.

[Zeng et al., 2018b] Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, Jun Zhao. Extracting relational facts by an end-to-end neural model with copy mechanism. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 506–514, Melbourne, Australia, 2018b.

[Zhang et al., 2017] Meishan Zhang, Yue Zhang, Guohong Fu. End-to-end neural relation extraction with global optimization. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1730–1740, Copenhagen, Denmark, 2017.

[Zhang et al., 2020a] Ranran Haoran Zhang, Qianying Liu, Aysa Xuemo Fan, Heng Ji, Daojian Zeng, Fei Cheng, Daisuke Kawahara, Sadao Kurohashi. Minimize exposure bias of Seq2Seq models in joint entity and relation extraction. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 236–246, Online, 2020a.

[Zhang et al., 2020b] Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, Jun Zhao. MIE: A medical information extractor towards medical dialogues. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6460–6469, Online, 2020b.

[Zhao et al., 2021] Jun Zhao, Tao Gui, Qi Zhang, Yaqian Zhou. A relation-oriented clustering method for open relation extraction. In Proceedings of the 2021 Conference on Empiri- cal Methods in Natural Language Processing, pages 9707–9718, Online and Punta Cana, Dominican Republic, 2021.

[Zheng et al., 2017] Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, Bo Xu. Joint extraction of entities and relations based on a novel tagging scheme. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1227–1236, Vancouver, Canada, 2017.

[Zhong & Chen, 2021] Zexuan Zhong, Danqi Chen. A frustratingly easy approach for entity and relation extraction. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 50–61, Online, 2021.

[Zhou et al., 2021] Wenxuan Zhou, Kevin Huang, Tengyu Ma, Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 14612–14620, 2021.

第五章 事件知识获取

丁效

哈尔滨工业大学 计算学部，哈尔滨 150001

一、任务定义、目标和研究意义

信息抽取任务随着互联网信息爆炸式的增长越来越凸显其重要性，而事件抽取又是信息抽取中至关重要的一个研究点。它旨在将无结构化文本中人们感兴趣的事件以及事件所涉及到的时间、地点、人物等元素准确地抽取出来并以结构化的形式存储下来，以供自动文摘、人机对话、情感分析、话题检测等自然语言处理上层技术的使用和用户方便的查看。本章重点介绍事件抽取、事件表示学习及事理图谱构建的相关研究工作。

1. 任务定义

根据美国国家标准技术研究所组织 ACE（Automatic Content Extraction）的定义[Doddington et al. 2004]，事件由事件触发词（Trigger）和描述事件结构的元素（Argument）构成，因此事件抽取任务主要包括以下两个步骤：

（1）事件类型识别：触发词是能够触动事件发生的词，是决定事件类型的最重要特征词。一般情况下，事件类型识别任务需要预先给定待抽取的事件类型。对于每一个检测到的事件还需要给其一个统一的标签以标识出它的事件类型。ACE 2005/2007 定义了 8 种事件类别以及 33 种子类别，如表 1 所示。

（2）事件元素识别：事件的元素是指事件的参与者，ACE 为每种类型的事件制定了模板，模板的每个槽值对应着事件的元素。

表 1 ACE 事件的类别

Types	Subtypes
Life	Be-Born, Marry, Divorce, Injure, Die
Movement	Transport
Transaction	Transfer-Ownership, Transfer-Money
Business	Start-Org, Merge-Org, Declare-Bankruptcy, End-Org
Conflict	Attack, Demonstrate
Contact	Meet, Phone-Write
Personnel	Start-Position, End-Position, Nominate, Elect
Justice	Arrest-Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon

图 1 给出了 ACE 2005 中定义的 Business 大类，Merge-Org 子类事件的一个详细描述的例子，“购并”是这类事件的一个触发词。该事件由三个元素组成，“雅虎公司”、“9 号”、“奇摩网站”分别对应着该类（Business/Merge-Org）事件模板中的三个角色标签，即：Org、Time-Within 以及 Org。

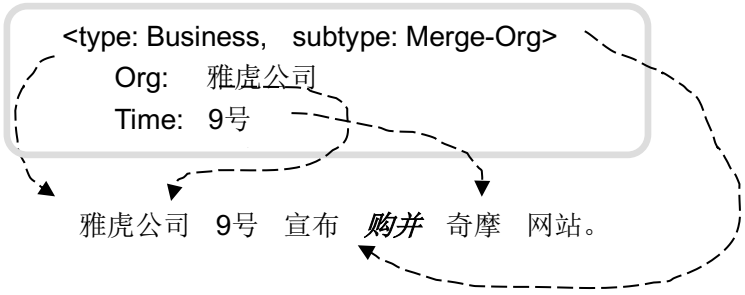


图 1 “购并”事件的基本组成要素

1）公开评测和数据集

- 国际评测和相关语料资源

最早开始信息抽取评测的是由美国国防高级研究计划委员会资助的 MUC (Message Understanding Conference) 会议（1987~1998）连续举办七届，MUC 会议不仅举办论文宣读，poster 展示等形式的学术交流活动，还额外组织多国参加消息理解评测比赛。正是有了这一会议的大力支持，信息抽取的研究达到了高潮。随后 MUC 会议停办，两年后美国国家标准技术研究所组织 ACE（Automatic Content Extraction）会议，目前该会议已经成功举办八次信息抽取评测（2000~2008）。

值得一提的是，从 ACE 2003 开始引入中文事件抽取的相关评测，至今已经举行了四次评测。但是很可惜，从 ACE 2008 年开始中文事件抽取评测不再是其中的一项评测。这个很可能是因为目前的 ACE 事件抽取语料并不规范且任务过于复杂，很难有大的突破所致。

MUC 会议和 ACE 会议定义了信息抽取研究中应有的各项任务，以及对这些任务的性能评测方法。并且还组织大量人力标注语料供参赛者进行训练和测试。

每一届 MUC 会议都会针对某个特定的场景提供训练语料和测试语料。在最开始的四届评测中（MUC-1 到 MUC-4）只提供英文语料。随着非英语系国家的加入，MUC 会议逐渐认识到多国语言的重要性，在第五届评测会议（MUC-5）中增加了对日文的评测。作为全世界使用人数最多的汉语未能入选 MUC 会议应该算是一种遗憾，因此第六届评测会议（MUC-6）中增加了中文的评测。从已发表的研究来看，MUC-6 语料使用的最多，一方面是因为中文语料的引入，另一方面也是因为有了前五届的积累，语料的标注愈发正规和成熟。

两年后 ACE 会议接力 MUC 会议，继续组织信息抽取的评测。ACE 会议从早期只有英

语、阿拉伯语和中文的语料发展到现在融合了西班牙语系的评测语料。虽有补充，但每年补充的语料幅度不大，ACE 2005 年的中文评测语料仅有 633 篇文章，共计 30 万词左右。而 ACE 2007 语料并没有任何的增加，基本上是沿用 2005 的语料。

MUC 会议和 ACE 会议所提供的语料基本上是针对通用领域，还有一些特定域的语料也引起了学术界的重视。

卡耐基梅隆大学标注了 485 个电子板报构成的学术报告通知数据集：其中包含报告人、时间、地点等相关信息。国内的北京语言大学也标注了 4 类突发事件（地震、火灾、中毒、恐怖袭击）文本，每类事件标注 20 篇文本，共计 80 篇突发性事件语料。

- 评价方法

MUC 会议，对系统总体性能的评价是通过衡量该系统的各个子任务的抽取结果来反映的。MUC 的评价指标总结如下：准确率(Precision，P) 、召回率(Recall，R)和 F 值（F-Measure）。F 值是综合考虑准确率和召回率后对系统性能的综合评价指数。具体的准确率、召回率和 F-Measure 计算公式如下面公式（1）、（2）和（3）所示。

$$\mathrm{P}=\frac{\text{系统正确抽取结果的数目}}{\text{系统抽取结果的总数}} \quad (1)$$

$$\mathrm{R}=\frac{\text{系统正确抽取结果的数目}}{\text{语料中标注结果的总数}} \quad (2)$$

$$\text{F-Measure}=\frac{2\mathrm{P}\mathrm{R}}{\mathrm{P}+\mathrm{R}} \quad (3)$$

MUC 会议的评价标准相对而言比较简单、直观、透明、易于理解。ACE 在此基础上采用了基于错误代价的评价策略，对系统的各部分错误赋予一定的权重分值，且不同的错误对应不同的权重分值，然后从最大分值中减去错误的分值。通过对各个子任务分值的叠加得到系统整体性能的分值，因此系统的各个子任务都会影响最后的得分。例如：事件识别与跟踪（VDR，Event Detection and Recognition）的评价体系中，ACE 官方认为事件元素识别的还会对系统的影响最大，因此，赋予，事件元素识别错误的惩罚分值也最高。该评价体系还可以单独看评价当前测试模块的结果，不考虑其他模块的影响。下面介绍一下 ACE 中 VDR 子任务的具体评价方法。

$$VDR\_Value=\sum Value(sys\_token_i)/\sum Value(ref\_token_i) \quad (4)$$

其中，*VDR_Value* 即为系统的最终得分，$\sum_i Value(sys_token_i)$是系统关于事件抽取任务的得分，$\sum Value(ref_token_i)$标注语料给出的总分，起到归一化因子的作用。

其中，$Value(sys_token_i)$由公式（5）计算得到。

$$Value(sys\_token_i)=Element\_Value(sys\_token_i)\cdot Arguments\_Value(sys\_token_i) \quad (5)$$

由此可以看出，系统的得分有两部分计算所得，一部分取决于事件属性的识别，另一部

分取决于事件元素的识别，只是这两部分所占最后总分的权重会有所不同。

二、研究内容和关键科学问题

事件知识获取是自然语言处理领域一项非常具有挑战性的工作，当前的研究热点已经不局限于对于句子级事件类型的识别以及元素的抽取，其研究内容变得越来越丰富，包括但不限于篇章级事件抽取、事件表示学习、事件/事理知识库构建、事件预测等研究任务。

三、技术方法和研究现状

1. 事件模式归纳

通常情况下，事件抽取任务的事件类型以及每种事件类型对应的事件论元角色是预先定义好的，如 ACE（Automatic Content Extraction）2005 评测[Doddington et al. 2004]共包括了 8 大类 33 小类事件，每类事件都定义了一定数量的事件论元角色。然而通过人工归纳并定义事件类型及其所含事件论元角色不仅需要各个领域的专家知识，还需要耗费非常大的时间和人力成本。因此，如何自动发现新的事件类型以及定义相应的事件论元角色有着重大的社会价值和巨大的挑战性。本章介绍自动归纳事件类型及事件论元角色的研究，包括任务定义以及相关解决方法。这种任务一般被称为事件模式自动归纳（Event Schema Induction）。

1）事件模式自动归纳概述

事件模式自动归纳，简称事件模式归纳（Event Schema Induction），事件模式归纳指从无标注的文本中学习复杂事件及其实体角色的高级表示任务 [Chambers 2013]。

表 2 事件模式实例

事件类型	交通	威胁
句子 1	法庭官员称 2008 年 3 月 18 日梁女士对法官说她被一名不认识的人用 15000 美元雇来向澳大利亚运输海洛因。	哥伦比亚政府感到震惊，因为铀是生产大规模毁灭性武器的主要基础。
事件 1	运输	震惊
事件 1 元素	不认识的人（施事） 澳大利亚（目的地） 海洛因（受事）	哥伦比亚政府（当事人）
句子 2	官方媒体没有确认扎赫丹被绞死的 2 名罪犯的身份，但表示他们犯有运输 5.25 公斤海洛因的罪名。	人权组织批评集束炸弹，是因为它们的无差别危害性，而且未爆炸的炸弹对平民构成了类似于地雷的威胁。
事件 2	运输	威胁
事件 2 元素	他们（施事） 海洛因（受事）	炸弹（原因） 平民（当事人）

现有的事件模式自动归纳研究可以分为两大类：模板型事件模式自动归纳和叙述型事件模式自动归纳。模板型事件模式自动归纳主要建模事件的类型及对应的事件论元角色，归纳出的事件模式可用于指导事件抽取；叙述型事件模式主要建模事件之间的关系。本章主要介绍的是模板型事件模式自动归纳。狭义上讲，模板型事件模式即描述某类事件的通用模板，包括该类事件的事件类型及其对应的事件论元角色。例如，对于“选举”事件的事件模式来说，事件类型为“选举”，相应的事件论元角色包括：“日期”、“地点”、“胜者”、“败者”、“职位”。

事件模式中的事件类型名称及事件论元角色名称是人为定义，然而新闻中关于事件的描述往往并不包含具体的事件类型名称及事件论元角色名称，由新闻文本直接精确归纳出这些名称较为困难，但新闻中往往会包含可以描述事件类型及事件论元角色的隐含信息，如语料中与事件相关的动词集合可以描述事件类型，事件论元对应的实体集合及其上下文中的语义句法信息等可以描述事件论元角色。因此，目前学术界在事件模式归纳研究中对事件模式的定义并不是简单的“事件类型名称+事件论元角色名称”的形式，而是“事件类型表示+事件论元角色表示”的形式。具体的，事件类型的表示形式主要包括：事件类型对应的事件触发词集合、事件类型的隐向量表示等；事件论元角色的表示形式主要包括：事件论元角色对应的实体集合、事件论元角色语义语法表达式、事件论元角色的隐向量表示等。目前，事件模式自动归纳仍然是一个极具挑战性的任务。

归纳得到的事件模式与一些自然语言处理研究有很多关联，如框架（Frame）、脚本（Script）以及信息抽取（Information Extraction）等。此外，事件模式还可以帮助事件抽取任务定义待抽取的事件类型和相应的事件论元角色集合，以及作为技术基础来辅助构建以结构化事件为节点的事理图谱。自动归纳的事件模式可以无需大量领域专家知识以及不需要耗费大量的人力物力，从文字信息中自动归纳得到若干包括事件类型及相应的事件论元角色集合的事件模式，如表 2 所示，相比于人工构建事件模式，自动归纳的事件模式能快速地迁移到新领域。

（1）基于概率图的事件模式归纳

概率图模型（Probabilistic Graphical Model）是指利用图表达概率相关关系的一类模型方法来表示模型相关的一些变量的联合概率分布，是一种比较通用的对于不确定性知识的表示和处理方法，贝叶斯网络、马尔科夫模型、主题模型等基于概率图的方法也应用于各种自然语言处理问题中。概率图模型的研究方法基于端到端的概率模型，可以对隐含的事件结构进行建模，将事件类型及事件论元角色建模并表示为概率模型的隐变量，进一步对事件类型的隐含表示进行较好的建模可以得出不同类型事件的聚类。在解决事件模式归纳任务时，很多学者借鉴了主题模型的方法，加以利用和改进后应用到这一任务上。主题模型（Topic Model），

是以无监督学习的方式对文章的隐含语义结构进行聚类的统计方法，其常被用于文本收集、文本分类与聚类、降维等研究中，其中，隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）是一种常见的主题模型。

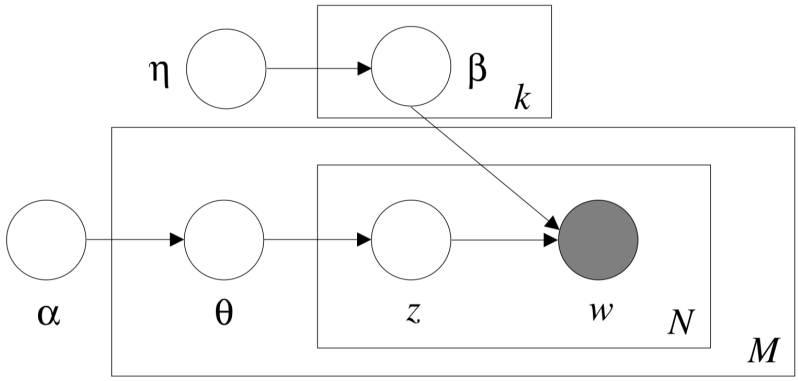


图 2 隐含狄利克雷分布的图模型表示

主题模型主要探索语料中主题与词分布的关系，隐含狄利克雷分布采用贝叶斯流派的思想，认为模型中需要估计的参数即主题分布以及词分布不是常数，而是服从狄利克雷分布的随机变量，在观测语料库中的样本后再对先验的狄利克雷先验分布的参数加以修正进而得到后验分布，图 2 展示的是隐含狄利克雷分布的图模型表示。整个语料库的生成过程可以看为对语料库中的每一篇文档获取到主题分布和词分布，然后从主题分布和词分布中对主题和词进行采样，隐含狄利克雷分布方法需要求得主题分布和词分布的期望，所以可以通过吉布斯采样等方法不断迭代计算获得主题分布和词分布的期望值。在给定主题数量这个超参数的前提下，主题模型背景下的文档聚类可以很好地根据文章主题将文档分成不同的类型。简单来说，主题模型假设语料库中每个文档的主题服从一定的分布，而对于每个主题，每个词语也服从一定的分布，因而可以通过文章中词语出现的概率计算其属于某种主题的概率。类似地，对于事件模式，可以类比认为语料库中文本所包含的事件类型也服从一定的分布，每个事件类型中，每个事件论元同样服从一定的分布，由此，事件模式归纳任务可以看为对事件类型、事件论元词等分布的期望计算过程。

受启发于上述主题模型，Chambers 等人 [Chambers & Jurafsky 2011] 在 2011 年尝试将朴素的隐含狄利克雷分布方法用于聚类事件，尽管在其研究工作中证明基于词汇距离的层次聚类在聚合事件的效果上会更佳，但这种尝试为事件模式归纳工作打开了思路。而后，2013 年 Cheung 等人[Cheung et al. 2013]将隐马尔科夫模型引入框架归纳（frame induction）研究工作，将框架、事件、事件参与者看做隐变量并学习其中的转移过程。同年，Chambers [Chambers 2013]首次将基于概率图生成模型的方法应用于事件模式归纳，通过实体的共指将

事件论元链条化，并同时考虑语料中词汇的词法与句法关系，使生成模型首先选择谓词而后预测其他的事件论元，实现了比隐马尔科夫更好的性能并且只需要更少的训练数据，但是其上述工作，只采用了实体的核心词（head word）来代表实体，然而忽略了同样会传递重要信息的对实体修饰限制的形容词等词，所以 Nguyen 等人 [Nguyen et al. 2015] 在其 2015 年的工作中认为，前人工作仅仅依靠实体核心词进行事件类型或事件论元角色聚类的方法会导致一些语义不明确的词汇所对应类型难以区分，如“士兵”在“袭击”事件中，可能存在“士兵”是施事者也有可能是受事者的上下文，因此引入实体核心词周围的上下文来实现对实体的消歧。近年来，深度神经网络的广泛应用也同样吸引了事件模式归纳工作的学者，Liu 等人 [Liu et al. 2019] 在 2019 年将基于神经网络的方法引入概率图模型，利用预训练语言模型和神经变分推断，并同时考量了新闻数据集中天然存在的冗余报道，提升了事件模式自动归纳的连贯性和模式匹配指标。

（2）基于表示学习的事件模式归纳

上一节介绍了基于概率图模型的事件模式自动归纳方法，在聚合同类事件时除了基于概率图的类主题模型方法外，在深度学习被广泛应用的当下，神经网络拥有强大的表示能力，可以表示任意的文本。因此，通过神经网络可以对词语、事件或文本进行稠密的向量表示，基于词语、事件或文本等的表示可以实现事件类型和事件论元角色的聚类（自动归纳）。在向量化表示前，早期的一些研究基于词语共现的统计学方法，例如在 2013 年 Balasubramanian 等人 [Balasubramanian et al. 2013] 通过 Open IEv5 工具抽取得到关系三元组（元素 1，关系，元素 2）并通过共现统计得到事件模式。在向量化表示被提出后，自然语言的向量化表示在比较文本之间的相似度、计算文本间的相关性的效果上相比独热编码 （One-hot Vector） 有着显著提升，而对于聚类同类事件，将事件和事件论元通过向量表示后计算事件或者是事件论元之间的相似度是很直观的想法，同时，同一事件中的各种论元在这一事件中共现，不同事件中同一论元也可能多次存在，因此，所有论元作为节点，若在同一事件共现则可形成节点间的边，进而可以组成一张图，如对上述图结构进行分割，每个分割后结构可视为一个事件模式，在这样的思路下，Sha 等人 [Sha et al. 2016] 于 2016 年借用图像分割的归一化分割的方法实现对事件论元节点的聚类，此外模型通过词嵌入以及点互信息计算实体间的内部相关性，并通过句中的存在性约束同时抽取模式和槽信息。在自然语言处理的多年发展过程中，语言学家等领域专家对自然语言建立了相对完备的知识库，如 FrameNet、PropBank 等，其中包括了谓词的各种语义角色信息，Huang 等人 [Huang et al. 2016] 也在 2016 年利用流水线式的方法结合上述外部知识库和自然语言处理工具等，实现触发词与事件论元的联合聚类，并通过距离度量选择中心词作为事件类型名并从外部信息中选择事件论元角色名。

（3） 事件图模式归纳

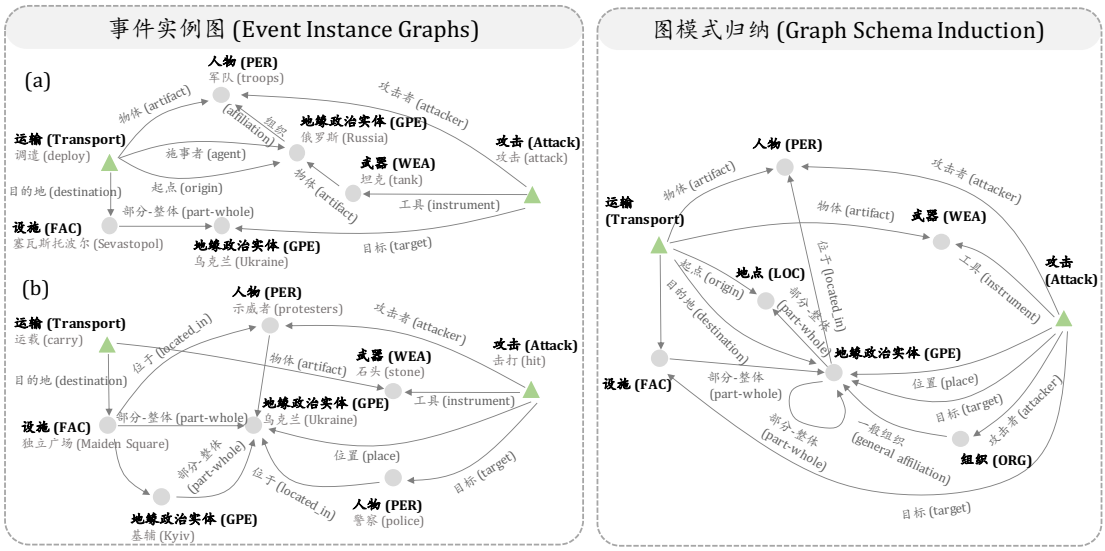


图 3 事件图模式归纳

事件图模式是在 2020 年由 Li 等人 [Li et al. 2020] 提出的一个新研究任务，既往的事件模式归纳仅仅关心同一个事件类型下的事件模式，然而在实际的文字信息尤其是新闻信息会包括多于一种类型的事件，而同篇文章中的不同类型的事件会共享一些事件论元，事件图模式即针对两种事件类型构建一篇文档的事件模式路径的有向无环图，图中存在两个事件类型节点和若干事件论元节点，两事件类型节点分别指向其事件中存在的事件论元节点，事件论元节点之间通过一些关系连接，继而从一个事件类型节点出发，到另一个事件类型节点停止，可以得到若干路径。如图 3 所示，(a) 和 (b) 是两个事件实例图，是分别从两个不同的文档中获取到的，每个图都包含了 2 个类型的事件：“运输”和“攻击”，每个事件都有一系列的事件论元角色以及对应的值，例如“攻击”事件的论元角色“武器”的值是“坦克”。由于“运输”类型事件和“攻击”类型的事件在同一篇文章中会有事件论元的联系，比如“运输”事件的目的地是“攻击”事件的目标，所以两个事件类型会形成一个有向无环图的结构，通过“运输”事件类型和“攻击”事件类型组成的多个图，希望能归纳出一个 (c) 所示的事件图模式，事件图模式中包含了 2 种事件类型以及它们的事件论元角色。Li 等人[Li et al. 2020]首先使用现有的信息抽取工具或者是人工标注的方式，得到实体、实体间的关系，事件以及事件论元，进行实例图的构建，然后经过处理得到显著的且连贯的路径，接着训练一个路径语言模型 (Path Language Model) 实现对某一路径进行打分，某一路径的得分构成是自身得分和邻居路径得分的加权，最后对于两个不同的事件类型，他们选取路径得分前 K% 的路径来构成两个事件类型之间的图模式。Li 等人 [Li et al. 2021] 在 2021 年进一步提出时间复杂事件模式 (Temporal Complex Event Schema) 的新概念：一种基于图的模式表示，

包括事件、时间元素、时间连接和事件论元关系。并且他们发布了一个新的事件图模式学习的语料库，人工事件图模式的黄金标准。最后通过模式匹配和实例图的复杂度进行内在评估，证明了他们的概率图模式与线性表示相比拥有更高的质量。

2. 事件识别和抽取

1）句子级事件识别和抽取方法

句子级事件抽取主流研究可以分为四个主要阶段：（1）早期发展阶段（上世纪 90 年代之前），以语言学家或领域专家手动编写规则和模板为基础的基于知识工程的方法的信息抽取，代表人物有 Riloff 和 Yangarber。（2）90 年代初到 2005 年，这段时间研究者们在不断反思基于规则的信息抽取系统的弊端：很难胜任大规模复杂类型数据集上的信息抽取任务。因此，基于统计和机器学习的方法被提出并开始在信息抽取领域广泛使用。（3）2005 年开始，以 Heng Ji 为代表的一系列信息抽取研究集中在跨文档事件抽取方面的研究，这种方法为信息抽取系统引入了更多的背景知识和语义知识，使得该系统功能更加丰富和智能。（4）为了克服限定域事件抽取类型、数目有限且需要固定的模板槽等局限性，2007 年华盛顿大学 Oren Etzioni 等人提出了开放域信息抽取方法。下面先分四个部分介绍这四个阶段典型的代表研究工作。

- 基于模式匹配方法的事件抽取

模式是对信息表述的一种描述性抽取规则。模式可以分为平面模式和结构模式。一般来讲，平面模式主要是基于词袋（bag-of-words）等字符串特征构成模式，由于不考虑相关句子结构和语义特征，因此被称为平面模式。而结构模式则是相对于平面模式而言，这种模式更多的考虑了句子的结构信息，融入句法分析特征。采用模式匹配方法的事件抽取系统工作流程基本上要分两个步骤：模式的获取和模式的匹配。

在模式的挖掘和构建过程中，非常重要的就是要找到高质量的模式，使得挖掘回来的模式既能准确地召回事件所涉及的事件元素，又不过多的引入噪声。在应用该方法进行抽取前，会将挖掘回来的模式进行打分排序，质量高的模式会获得一个更高的分数，从而在进行匹配时会优先进行匹配。该方法如果需要获得比较高的召回率，需要挖掘出尽可能多的模式并且将大部分的模式都用于事件元素的抽取；但是这样做的副作用就是排在后面的质量不是特别高的模式在提高了召回率的同时，也会抽取出一些无关的噪声数据，从而降低了事件元素抽取的准确率。

在模式获取方面的研究，早年的学者尝试了各种方案。Riloff 1993 年提出了 AutoSlog 系统[Riloff 2013]，基于知识工程的信息抽取系统在当时看来虽然取得了很大的成功，但是其中有一个很大的问题就是这种方法过于依赖人工构造的领域词典，然而这些领域词典的构建

过程并不是十分简单甚至会花费大量人力物力。因此，AutoSlog 系统通过 13 个启发式方法获得 13 个模板，然后再用这些模板去匹配文本，从而自动构建出领域词典，值得一提的是 AutoSlog 系统是世界上第一个使用机器学习方法进行信息抽取系统模式获取的系统。

Kim 和 Moldovan 1995 年提出了 PALKA 系统[Kim & Moldovan 1995]。这套系统也是基于人工标注语料的信息抽取模式学习系统。这套系统成功的融入了 WordNet 词典语义信息，从而使其更加擅长处理开放域信息抽取问题，而不仅仅局限于特定域的信息抽取。

Riloff 和 Shoen1995 年在 AutoSlog 系统的基础上提出了 AutoSlog-TS 系统 [Riloff & Shoen 1995]。这个系统与 AutoSlog 系统最大的不同或改进就在于，AutoSlog 系统需要人工标注的语料作为训练语料，然而构建这种语料时也是需要大量时间的。而 AutoSlog-TS 系统则不需要人工标注的语料，它仅仅需要人工把语料进行一个分类即可，最终的结果与 AutoSlog 系统相当，却节省了大量人工标注工作量。

Joyce Yue Chai 1998 年提出了 TIMES 系统[Joyce 1998]，这是一个基于 WordNet 和标注语料的信息抽取模式学习系统。WordNet 与人工标注语料共同使用确实起到了很好的效果，其系统抽取结果要好于以往的信息抽取系统，并且对于特定域与开放域语料均可以处理，但是由于需要作为输入的外部资源过多也限制了其应用。

Yangarber 2001 年提出了 ExDisco 系统[Yangarber 2017]，这个系统是基于种子模式的自举信息抽取模式学习系统。系统首先给定一个初始化的手工构造质量较高的种子模板，然后根据已有的模板在语料库上增量式的学习新的模板，经过几轮迭代后就获得了大量高质量模板。

姜吉发 2004 年在其博士论文中使用了一种称之为“GenPAM”的模板学习方法[姜吉发 2004]。它的优势在于完全的无指导学习模板，对于标注语料几乎没有需求。这里人工干预的部分在于给出要抽取的事件类型、事件元素及其所属角色。最后再人工地对模板的抽取质量进行评价。经过以上步骤，事件抽取模板便可以自动学习出来。这对于模式学习来讲，大大减少了人工工作量。

- 基于机器学习方法的事件抽取

随着各大企业逐渐认识到信息抽取的重要作用，以及它们对信息产业的迫切需求，大力推动了相关领域语料库的构建，有了这些语料库后，人们开始将研究重点转向基于统计和机器学习的方法进行信息抽取。一些经典的统计模型被引入，这些模型有隐马尔科夫模型（Hidden Markov Model，HMM）、朴素贝叶斯模型（Naïve Bayes Model, NBC）、最大熵模型（Maximum Entropy Model，ME）、最大熵隐马尔科夫模型（Maximum Entropy Hidden Markov Model， MEMM）、支持向量机模型（Support Vector Machine，SVM）等。这种基于统计模

型的机器学习方法将信息抽取看成是分类问题，其重点在于挑选合适的特征使得分类器更加准确。另外，核（kernel）的引入也使得分类器的效果有了很大的提升，也有研究者分析和开发新的核。

H. L. Chieu 和 H. T. Ng 2002 年在进行事件元素抽取的研究中，大胆尝试引入最大熵分类器，将事件元素的识别看成是一个分类问题。[Chieu & Ng 2002]。这套系统在 MUC 2002 评测中讨论发表会事件和工作交接事件抽取任务中获得了较好的结果。Chieu 在他的分类器中采用了 unigram、bigram、命名实体、短语等简单特征，最终在卡内基梅隆大学标注的语料库上进行实验验证，取得了 86.9%的 F 值，超过了当时的最好结果。

Ralph Grishman 参加了 ACE 2005 的事件抽取任务评测，在参赛的系统中他们使用了最大熵模型[Grishman 2005]。他们的系统共有四个模块（即四个分类器）：（1）基于事件触发词分类的事件类型识别模块；（2）事件元素识别模块；（3）事件元素角色识别模块；（4）整合已有的事件类型识别模块，事件元素识别模块，事件元素识别模块，并依据各个模块的输出结果最终判定输入的句子是否为事件。

Ahn 2006 年在提出了进行事件触发词及类别识别和事件元素识别这两个事件抽取主要任务的研究中，尝试性地在其事件抽取系统中整合了 Timbl 和 MegaM 两种机器学习方法[Ahn 2006]。Ahn 把事件类型识别看成事件触发词的识别，首先对输入的句子进行分词（就英文而言只需根据空格分词），对每一个词抽取相关的词法特征、上下文词特征、WordNet 词典特征以及上下文相关实体及其类型等特征，然后首先使用 MegaM 分类器对当前词进行二元分类来判断其是否是触发词。如果当前词被判定为触发词，则使用多元分类器 Timbl 指定当前词所属的事件类别及子类别。Ahn 的系统在 ACE2005 英文语料库上进行测试，实验结果显示事件类别识别的 F 值达到了 60.1%，这一结果超过了分别单独使用 MegaM 和 Timbl 分类器的方法。另外，针对事件元素识别任务，这套系统把句子中出现的每一个实体都看作是候选事件元素，抽取与实体相关的词法特征、事件属性特征、实体的修饰特征、依存句法路径特征等，并为每一种事件训练一个分类模型，专门用来确定事件元素的角色。该系统在 ACE 2005 英文语料上进行事件元素识别的测试，结果为：F 值达到了 57.3%。

Z. Chen 2009 年打破原有的将事件抽取看做分类问题的思维模式，而是将事件类型识别及元素识别看做序列标注问题，采用最大熵隐马尔科夫模型（Maximum Entropy Hidden Markov Model，MEMM），选择一般特征和中文独有的特征，在 ACE 2005 中文语料上测试，其 F-Measure 高于当前最好的中文事件抽取系统[Chen 2009]。

- 基于跨文档方法的信息抽取

传统的基于模式匹配的方法与基于统计机器学习的方法，实际上都是在做句子级的信息

抽取，这里很少考虑篇章和丰富的背景知识。在基于“One Trigger Sense for Cluster”和“One Argument Role for Cluster”的思想基础上，Heng Ji 于 2008 年提出了跨文档事件抽取系统框架[Heng & Grishman 2008]。在这个框架下，对于一个句子级的抽取结果不仅要考虑当前的置信度，还要考虑与这个待抽取文本相关的文本对它的影响。作者共设置了 9 条推理规则定量的度量相关文本对当前抽取结果的影响，从而帮助人们修正原有的句子级事件抽取结果。这个系统最后在 ACE 2005 英文语料上进行评测，事件类型识别最终 F 值达到 67.3%，事件元素识别最终 F 值达到 46.2%，均超过了目前最好的英文事件抽取系统。

Heng Ji 的这项研究一经发表后，引起了很多人的关注，后来学者借鉴她成功的引入篇章和背景知识的思想，相继出现了跨语言事件抽取系统 [Heng 2009]，跨文本事件抽取的改进 [Liao & Grishman 2010]，跨实体事件抽取系统[Hong et al. 2011]等相关研究。

- 开放域事件抽取

为了解决大规模语料信息抽取的问题，开放域事件抽取任务被首次提出，其主要抽取的是事件三元组（施事，事件词，受事）。在开放域事件抽取这一研究方向，华盛顿大学人工智能研究组做出了很多杰出的工作，并且开发出了一系列开源信息系统：TextRunner，WOE 和 ReVerb 等。TextRunner 是第一个对于关系名称进行抽取的开放域信息抽取系统，它首先利用启发式规则从语料库中获取句法特征，然后训练分类器判断两个元组之间是否存在某种语义关系，再利用海量互联网数据帮助评估抽取到的三元组是否正确。 WOE 则充分利用 Wikipedia 中大量人工填写的 InfoBox 信息，从中获取大量训练语料，从而训练信息抽取器抽取更多的信息三元组。 ReVerb 在 TextRunner 基础上提出了句法和词汇的限制条件，进而提高了三元组的抽取精度，使其更加实用，并且值得一提的是 ReVerb 用动词词组描述两个元组之间的语义关系，这非常符合事件的定义。

2）篇章级事件识别和抽取方法

篇章级事件抽取任务的目标是在文档中识别预先指定类型的事件及相对应的事件元素。近年来，随着金融、法律、公共卫生等各个领域数字化进程的发展，文档级事件抽取已成为这些领域业务发展的越来越重要的加速器。以金融领域为例，持续的经济增长见证了数字化金融文本的爆炸式增长，例如对特定股票市场中的大量金融公告文档进行文档级事件抽取，能够帮助人们提取有价值的结构化信息，预知风险并及时发现获利机会。同时，为促进信息检索和文章摘要等下游应用的发展，对文档级的事件抽取技术展开研究也是必不可少的。

传统的基于模式匹配的方法与基于统计机器学习的方法，实际上都是在做句子级的信息抽取，很少考虑篇章和丰富的背景知识。在基于“One Trigger Sense for Cluster”和“One Argument Role for Cluster”的思想基础上，Heng Ji 于 2008 年提出了跨文档事件抽取系统框

架[Heng & Grishman 2008]。在这个框架下，对于一个句子级的抽取结果不仅要考虑当前的置信度，还要考虑与这个待抽取文本相关的文本对它的影响。作者共设置了 9 条推理规则定量的度量相关文本对当前抽取结果的影响，从而帮助人们修正原有的句子级事件抽取结果。这个系统最后在 ACE 2005 英文语料上进行评测，事件类型识别最终 F 值达到 67.3%，事件元素识别最终 F 值达到 46.2%，均超过了目前最好的英文事件抽取系统。Heng Ji 的这项研究一经发表后，引起了很多人的关注，后来学者借鉴她成功的引入篇章和背景知识的思想，相继出现了跨语言事件抽取系统 [Heng 2009]，跨文本事件抽取的改进 [Liao & Grishman 2010]，跨实体事件抽取系统[Hong et al. 2011]等相关研究。

此外，最近的部分工作探索了采用 Pipeline 框架来解决文档级事件抽取任务，该结构为每种类型的事件及事件元素训练单独的分类器，并通过上下文来增强模型性能，以学习事件类型识别及事件元素抽取策略。GLACIER [Patwardhan & Riloff 2009]在概率模型中同时考虑了跨句信息以及能够作为依据的名词短语以提取角色填充物。TIER [Huang & Riloff 2011]则提出首先使用分类器确定文档类型，然后在文档中识别事件相关的句子并填充事件元素槽。2012 年 Riloff 等人[Huang & Riloff 2012]则提出了一种自下而上的方法，该方法首先根据词汇句法模式特征来识别候选的事件元素，然后通过基于语篇特征的分类器来移除与事件无关的句子中的候选事件元素。

上述方法存在跨不同 Pipeline 阶段的错误传播问题，同时需要大量的特征工程（例如，用于候选事件元素发现的词汇句法模式特征、用于在文档级别检测与事件相关的句子的语篇特征），而且这些特征需要针对特定领域手动设计，又有一定的领域专业知识门槛。然而神经端到端模型已证明在命名实体识别、ACE 句子级事件抽取等句子级信息提取任务上表现出色。

因此，Du 等[Du et al. 2020]于 2020 年提出将文档级事件抽取任务作为端到端神经序列标注任务来解决。作者认为文档级事件抽取任务无法利用句子层面的抽取方法得到解决，其最主要的原因是一个事件的论元分散在了不同的句子当中，因此如何获取跨句子信息就显得较为重要。由于文档的长序列特点，捕获长序列中的远距离依存关系是文档级神经端到端事件抽取的一项基本挑战，该工作对输入的上下文长度与模型性能之间的关系进行了研究，找到了最合适的长度来学习文档级事件抽取任务。此外该工作还提出了一种新颖的多粒度特征抽取器，以动态汇总在不同粒度（例如句子级和段落级）学习到的神经表示所捕获的信息。在 MUC-4 事件提取数据集所提出的方法上比以前的工作表现更好。

文档级事件抽取的另一个主要障碍是培训数据的缺乏。由于基于远程监督技术来自动生成训练数据的方法已取得了大量进展，一些研究试图通过远程监督来缓解该问题。例如，考

虑到经典的事件抽取任务所要求的触发词信息在知识库中并没有出现 Chen 等[Chen et al. 2017]采用额外的语言资源及预先定义的词典来标记触发词。

在金融领域，文档级事件抽取技术可以帮助用户获得竞争对手的策略，预测股票市场并做出正确的投资决策，然而在中文金融领域中，没有待标记的文档级事件抽取语料库。Yang 等[Yang et al. 2018]则针对中文金融领域文档级事件抽取的文档级建模及数据缺乏两大挑战展开研究。该工作提出了 DCFEE 框架，该框架将文档级事件抽取任务视为序列标注任务，基于远程监督技术自动生成大量带伪标签的数据，并通过关键事件检测模块和事件元素填充策略，从财务公告中提取文档级事件。

对于财务文档以及许多其他业务领域中的文档而言，事件元素分散和多事件的特点给文档级事件抽取带来了挑战。第一个挑战是一个事件的事件元素可能散布在文档的多个句子中，而另一个是一个文档可能包含多个事件的信息。Zheng 等[Zheng et al. 2019]针对上述挑战提出了一种新颖的端到端模型 Doc2EDAG，Doc2EDAG 的关键思想是将事件信息转换为基于实体的有向无环图，该形式可以将原本的表格填充任务转换为更易于处理的多路径扩展任务。为了有效地生成 EDAG，Doc2EDAG 对文档中的实体基于上下文进行编码，设计了一种适用于路径扩展任务的存储形式。此外该工作还改进了文档级事件抽取的标记体系，删除了触发词标记，这种无须触发词的设计不依赖任何预先定义的触发词集或启发式方法来筛选触发词，并且不改变文档级事件抽取的最终目标。其整体模型分四个模块：预处理模块、文档级信息融合模块、文档级信息记忆模块、路径扩展模块。首先预处理模块利用 transformer 编码器将输入文本转换为词向量序列，并添加 CRF 层，利用经典的 BIO 标注方案训练模型进行实体识别。其次，文档级信息融合模块为了有效地解决论元分散的挑战，利用全局上下文来更好地识别一个实体是否扮演特定的事件角色，该模块的训练目标是上下文对预处理中提取的实体提及进行编码，并为每个实体提到的内容生成实体向量，为了提高对文档级上下文的认识，作者使用了第二个 transformer 模块，以方便所有实体和句子之间的信息交换。模型中还增加了句子的嵌入位置来指示句子的顺序。在这个模块之后，获得了文档级上下文相关的实体和句子表示，并对每种事件类型进行了事件触发分类。然后，文档级信息记忆模块考虑到依次生成基于实体的有向无环图时必须同时考虑文档级上下文和路径中已经存在的实体，采用了一种内存记忆机制，更新图结构时需要追加已经识别的实体嵌入。最后，路径扩展模块在扩展事件路径时对每个实体进行二分类，结合当前路径状态、历史上下文和当前角色信息判断是否对当前实体进行展开。在由大规模的财务公告组成的真实数据上 Doc2EDAG 的表现超过了以往的工作。

3. 事件关系获取

事件是由特定人、物、事在特定时间和特定地点相互作用的客观事实。然而，事件的发生往往不是孤立现象，一个事件的发生必然存在与之相关的其他事件，例如与该事件相关的原因事件、结果事件、并发事件等。事件与其相关事件之间相互依存和关联的逻辑形式，称之为事件关系。事件关系抽取以事件为主题元素，通过分析事件文本的结构信息及语义特征，挖掘事件之间深层的逻辑关系，进而辅助事件的衍生、发展以及信息的推理与预测。本章主要对以下几种公认的事件关系即事件因果关系、事件时序关系、子事件关系和事件共指关系进行介绍。

1）事件因果关系获取

事件因果关系不仅是语篇理解的重要组成部分，对于问答等各种自然语言处理应用也具有重要意义。它包括两个部分，原因和结果。例如:“公共汽车没能出现。因此，我开会迟到了”。这里的原因是“公共汽车没有出现”，而结果是“我开会迟到”。因果关系可以是显式的，也可以是隐式的。通常，显式因果模式可以包含相关的触发词，如原因（cause）、结果（effect）、结果（consequence），也可以包含模糊的触发词，如生成（generate）、诱导（induce）等。隐式因果关系比较复杂，涉及基于语义分析和背景知识的推理。一个隐式因果关系的例子：“飓风卡特里娜星期一早上沿着墨西哥湾海岸向海岸肆虐。早些时候有报道说沿岸有建筑物倒塌”，这里飓风的“肆虐”导致了建筑物“倒塌”。因此，因果关系的抽取极其复杂和困难[1]。该任务常用的评价指标有：准确率（Acc）、精确率（P，precision）、召回率（R，recall）、F1 值。

当前，已有工作涵盖基于监督/无监督的抽取方法，包含针对语言模式、统计方法和监督分类器等建模方式，从文本语料中获取事件因果关系的知识。例如 Kaplan 等人[Kaplan & Rogghe 1991]提出基于手工编码的、特定领域的知识推理从文本中提取句子间隐含的因果关系，但在实际应用中较难扩展。Khoo 等人[Khoo et al. 2000]使用预定义的语言模式（linguistic patterns）从商业和医学报纸文本中识别明确的因果关系，而不需要任何基于知识的推理。Girju 等人[Girju et al. 2003]设计出了一种自动检测表达因果关系的词汇句法模式的方法。使用名词-动词-名词的词汇-句法模式来捕捉“蚊子引起疟疾”这样的例子，其中提到的因和果是名词，不一定是事件。Do 等人[Do et al. 2011]设计了一种最小监督方法，利用因果线索和事件间的统计关联识别语境中的事件因果关系。基于 Do 等人的工作，Riaz 和 Girju 等人[Riaz

[1]该任务涉及到许多因素，如事件的上下文特征(如词汇项（lexical items）、动词时态、动词的论元等)、事件的语义和语用特征、背景知识、世界知识、常识等。

& Girju 2013]探究了哪些类型的知识有助于动词（事件）间的因果关系识别。他们提出了一种无监督方法，基于一套知识丰富的度量来学习动词（事件）之间因果关系。利用这些度量标准，能够自动生成一个知识库(KB)，其中标识三种类型的动词对:强因果的、模糊的和强非因果的。和 Do 等人[Do et al. 2011]提出的 CEA 相比，Riaz 和 Girju 等人[Riaz & Girju 2013]引入了知识丰富的关联度量指标，利用自动生成的训练语料库的监督来学习因果关系。同时，针对无监督方法，定义了 3 种涵盖显式、隐式因果关联的评价指标。Hashimoto 等人[Hashimoto et al. 2014]提出一种利用事件的词汇语义信息建模的有监督方法（基于大量的手工特征训练有无因果关系的二分类器）。利用该方法能够从互联网上抽取得到如“从事刀耕火种的农业”导致“加剧沙漠化”的因果关系。这些关系可被看作是未来可能发生的事件进而帮助人类实现情景规划（scenario planning）。Gao 等人[Gao et al. 2019]**错误!未找到引用源。**针对文档级别的因果关系进行建模，抽取了包含句内和跨句的所有因果关系。因果具有方向性，文中仅识别两个事件是否存在因果关系，并不对二者间的方向做判断。针对事件因果关系的稀疏问题且很少显式表达，使用 ILP 分别从全局和细粒度方面对因果结构进行建模（产生约束条件）。具体地，全局建模基于一个观察，即因果关系（尤其跨句）往往涉及文档中的一两个主要事件（可看作故事的焦点，通常在文章标题中提到，并在整个文档中反复提到）。细粒度方面分别从特定的句子句法关系、篇章关系、事件因果关系和事件相关关系角度进行建模。由于句内因果和句间因果在本质上的不同，分别构建两个独立的分类器分别用于句内/句间的因果关系检测。

现有工作仅利用了标注数据，缺乏使用有助于该任务的相关外部知识的能力，通常对新的、以前未见过的数据表现不佳。针对这个问题，Liu 等人[Liu et al. 2020]提出带知识感知的因果推理机（knowledge-aware causal reasoner），利用 ConceptNet 引入外部知识进行推理，很大程度丰富事件表示。又由于知识库本身具备不完备的缺陷，Liu 等人提出指称掩码推理机（mention masking reasoner）挖掘与事件无关的，基于特定上下文的模式，能够大幅增强模型处理新的，之前未见过的数据的能力。这里基于一种假设：在包含因果关系的表述中，往往包含事件无关的语言模式，这对识别新事件的因果关系很有帮助。在此基础上，提出细心哨兵模块（attentive sentinel）对以上两个推理机进行权衡，是一个句子级别的两两事件间的因果抽取模型。

除了基于外部知识库作为知识源，另一种常被作为知识源的是被广泛使用的语言模型。Kadowaki 等人[Kadowaki et al. 2019]提出一种基于 BERT 的方法抽取事件因果关系，作为基于大语料进行预训练的语言模型，BERT 在预训练过程中可以学习到一些事件因果关系的背景知识。此外，在标注事件因果关系时，关系标签的确定通常需要对多个标注结果（来自多

个标注者）依照多数投票方式确定。这种标注方式忽略了每个标注者的独立判断结果。通过训练多个分类器捕捉每个注释者的标注策略，结合产生的分类器输出来预测最终标签能够进一步提升模型性能。Li 等人[Li et al. 2021]提出预训练模型 CausalBERT，通过将因果知识注入预训练语言模型，使预训练模型具备因果推理能力。具体地，通过设计因果对分类任务实现为 BERT 等预训练模型注入因果知识。利用 CausalBank 语料[Li et al. 2020]，构建正负例因果对，并采用合页损失函数作为训练目标。

2）事件时序关系获取

事件时序关系抽取是一项重要的自然语言理解任务，对后续任务如问答、信息检索和叙事生成等都有重要的作用。该任务可以被建模为针对给定文本构建一个图结构，图中节点表示事件，边被相应地标记为事件时序关系，如图 4 所示。已有工作一般将该任务分为两个独立的子任务，即事件抽取和事件时序关系分类。这种做法假设在训练关系分类器时，已经给定了正确抽取的事件结果。该任务包含以下三种常用的评价指标：

- 准确率（Acc）
- 精确率（P，precision）、召回率（R，recall）、F1 值
- 时序意识得分（temporal awareness score）：从精度（precision，P）、召回率（recall，R）和 F1 值方面捕捉标注的时序意识（temporal awareness），能够更好的捕捉事件时序图有多“有用”。精度，召回率计算公式如下所示：

$$P = \frac{\left|G_{sys}^{-} \cap G_{true}^{+}\right|}{\left|G_{sys}^{-}\right|}, \qquad R = \frac{\left|G_{true}^{-} \cap G_{sys}^{+}\right|}{\left|G_{true}^{-}\right|}$$

其中G^{+}表示图 G 的闭包，G^{-}表示图 G 的约简，即去掉图中的冗余关系。∩表示两图中时序关系的交集，$|G|$表示图 G 中边的数量（时序关系的个数）。给定两个系统 1 和 2，如果系统 2 仅为系统 1 的传递闭包，两个系统会产生相同的评价结果。这里，时序间的模糊关系（vague）常被视作不存在的时序边，且在评价过程中不被考虑在内。

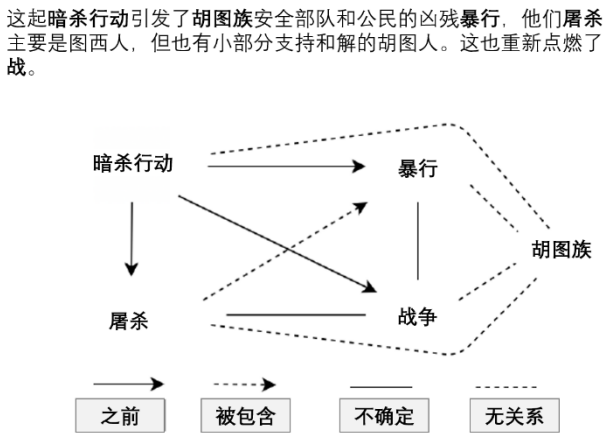


图 4 事件时序关系图

近年来，事件时序关系抽取在自然语言处理领域引起了广泛关注。该任务的一个标准数据集是基于 TimeML 标准[1]标注的 TimeBank（TB）语料。在此之后，一系列的时序关系数据集被收集起来，包括但不限于 Bethard 等人[Bethard et al. 2007]利用动词从句对 TB 的扩展，TempEval1-3 数据集，TimeBank-Dense（TB-Dense）数据集，EventTimeCorpus 数据集，MATRES 数据集以及同时包含时序关系和其他类型关系的多标注数据集（例如，包含事件共指关系和因果关系）如 CaTeRs，RED 等。

现有的标注方法均采用事件在时序上的区间表示，令$\left[t^1_{\text{开始}},\ t^1_{\text{结束}}\right]$和$\left[t^2_{\text{开始}},\ t^2_{\text{结束}}\right]$分别表示两个事件对应的事件区间（隐含$t_{\text{开始}} \le t_{\text{结束}}$的假设）。在两个区间之间共包含 13 种时序关系，如图 4 所示。为了进一步缓解标注负担，一些工作经常仅使用 13 种关系约简后的集合。

3）子事件关系获取

给定事件对（A, B），如果事件 B 是事件 A 的子事件，需要满足以下条件：（1）A 是一个复杂的活动序列，大部分由相同(或兼容的)代理（agent）执行;（2）B 是活动序列中的一个；（3）B 与 A 发生在同一时间和地点。这里 A 扮演了一种事件集合的角色。这种关系使得不同的事件间形成了一个典型的事件序列（或脚本）。例如：“伊斯梅尔说，这场持续了几天的战斗加剧了，因为效忠伊戈尔的伊萨克人哈巴尔·阿瓦尔部族的部队**袭击**(E12)了他主要的反对派对手的一个民兵据点，…。声称自己是国防军的伊加勒民兵说，他们**占领**(E15)了反对派的两个哨所，**杀死**(E16)了和**打伤**(E17)了许多战士，**摧毁**(E18)了三辆技术车(武装皮卡)，**没收**(E19)了大炮和各种弹药。”进一步，将例子中的事件构建了一个事件图，如图所示。图 5 中，事件 E15 是事件 E12 的子事件。事件 E15，E16，E17，E18，E19 在它们的父事件 E12 下形成了一个聚类。箭头表示从父事件指向子事件的一个子事件关系。该任务常用的评

[1] 查看 http://www.timeml.org 获取语言规范和注释指南

价指标有：BLANC、精确率（P，precision）、召回率（R，recall）、F1 值。

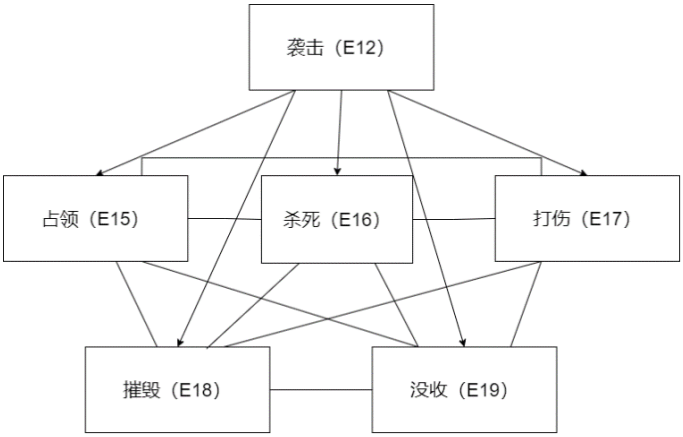


图 5 一个子事件关系的例子

子事件关系抽取常用的评估语料有 HiEve 语料，IC 语料，SeRI 语料。HiEve 语料关注于新闻故事中的子事件关系。由于新闻故事中包含大量表示不同时空粒度的真实事件，新闻故事中的叙述通常描述一些粗糙的具有空间、时间粒度的现实世界事件及其子事件。Glavaš 等人[Glavaš et al. 2014]基于新闻故事，提出了 HiEve 语料，一个识别事件之间时空包容关系的语料库。在 HiEve 中，叙事被表示为基于时空包容关系(即父事件-子事件关系)的事件层次，事件关系主要包含：父子事件关系（SUPERSUB），表示事件对中的第一个事件在空/时间上包含第二个事件；子父事件关系（SUBSUPER），和父子事件关系对称；共指关系（COREF），表示两个事件指称表示了现实世界中的同一事件；无关系（NORELATION），表示两个事件既无空时包含，也无共指关系。语料中包含了 100 篇文档，包含 1354 个句子，33273 个词。Hovy 等人[Hovy et al. 2013]标注了一个情报系统(intelligence community, IC)语料库，包含暴力事件领域(爆炸、杀戮、战争等)的文本。鉴于部分共指类型的稀疏性，语料中注释了事件完全共指、子事件和成员关系的实例。除了新闻领域等限定域，Ge 等人[Ge et al. 2018]基于英文维基百科中特有的关系模板（partof）及规则构建一个 SeRI 语料，包含了 3917 篇事件文章，共 7373 个候选子事件对。该语料中共包含三种关系：父子事件关系，子父事件关系，无关系。可以用做从百科全书中挖掘子事件关系的模型的训练及评估语料。

4. 事件表示学习

由于传统的 One-hot 高维特征表示方式会使得事件特征异常稀疏，从而不利于后续的研究和应用，因此，Ding 等人提出了两种全新的事件表示方式。第一种离散模型是基于语义词典对事件元素，进行泛化，进而缓解事件的稀疏性。第二种连续向量空间模型则为每一个事件学习一个低维、稠密、实数值的向量进行表示，从而使得相似的事件具有相似的向量表示，在向量空间中相邻。

由于历史上发生的事件大多数都很难以再次发生，因此会导致事件具有严重的稀疏性，离散模型的目标是对同一事件的不同表达进行归一和泛化。例如，“微软以 72 亿美元价格吞并诺基亚移动手机业务”和“微软出资 72 亿美元收购诺基亚移动手机业务”表达的是同一事件。为了完成这一目的，可以利用几个广泛应用的语义词典 WordNet、HowNet 和 VerbNet 等对事件元素进行泛化。具体而言，泛化过程包含两个步骤。首先，从 WordNet 中找到事件的施事者和受事者中名词的上位词将其泛化。例如，利用“微软”的上位词是“IT 公司”将其替换掉。随后，找到事件元素中的动词，并用 VerbNet 中该动词所属类别的名词替换掉改动词，从而对其进行泛化。例如，“增加”在 VerbNet 中所属的动词类别名称为 multiply。下面给出一个事件泛化的完整例子，给定句子“Instant view: Private sector adds 114,000 jobs in July.”，可以抽取出事件（Privatesector, adds, 114,000 jobs）将其泛化后的结果是（sector, multiply class, 114,000job）。类似方法也曾被 Radinsky[126]提出用来做因果事件预测任务上。

离散模型方法简单且有效，但是也存在着两个重要的局限性： 其一，WordNet， VerbNet 等语义词典词覆盖有限，很多词难以在语义词典中找到相应记录。其二，对于词语的泛化具体到哪一级不明确，对于不同应用可能会有不同要求，很难统一。此外，即使对事件进行了泛化还是无法解决 One-hot 的特征表示带来的维度灾难（curse of dimensionality）问题。例如，假设词典中有 10,000,000 个词，那么就需要用 10,000,000 维特征表示一个词。由此带来的特征稀疏问题，会导致后续的应用难以取得较好结果。并且超高维度的特征空间也会消耗大量的实验时间和空间存储，增加了计算成本。

为了解决这一问题，Bengio 首先提出了为词汇学习一个分布式表示（即 word embedding），用低维、稠密、实数值向量表示一个词汇。为了学习这样一个词汇向量（向量维度一般是 30，60，100，200 等），Bengio 训练一个神经网络模型将该词汇的大规模上下文语义信息都融入到词汇向量中。由于语义上相似的两个词汇应该会有相似的上下文，因此，相似的词汇也应该会学习到相似的词汇向量。

事件的分布式表示学习动机与词汇的分布式表示学习动机是一样的，Ding 等人提出学习低维、稠密、实数值事件向量表示，从而相似的事件在向量空间中具有相邻的位置。该任务与知识库中的多元关系数据分布式表示学习相近似，关系数据的分布式表示学习是为关系三元组(e_1, R, e_2)学习一个连续向量，其中 e_1 和 e_2 是命名实体，R 是这两个命名实体之间的关系类型。然而，事件的表示学习与关系的表示学习也有着显著的不同之处，主要体现在两个方面。

第一，知识库中的关系类型数量有限。因此大多数关系数据的分布式表示学习模型都将某一个特定关系类型用一个矩阵或者张量建模学习。然而，抽取的是开放式事件元组，因此，

事件类型是开放的，也就是无限的，这样就导致无法用一个矩阵或张量建模某一个事件类型，因为这样的代价太高了。为了解决这一问题，Ding 等人[Ding et al. 2015]将事件词 P 也表示成与施事者 O_1 和受事者 O_2 具有相同纬度的向量，从而摆脱了事件类型无限多的限制。

第二，关系的表示学习目的是能够指出两个命名实体(e_1, e_2)是否具有某一确定的关系 R。当 R 是一个正定矩阵时，命名实体是可以互换位置的，也就是说这时候关系是没有方向性的。然而，事件元素都是有特定角色的，其具有很强的方向性，谁是事件的施事方，谁是受事方是不可以随便变化的，一旦改变则事件就完全不同。

基于以上的分析，Ding 等人[Ding et al. 2015]设计了一个全新的张量神经网络来学习事件的结构化向量表示，事件的每一个元素及其所扮演的角色都会被显式地建模学习。如图 9-2 所示，两个张量 T_1 和 T_2 被分别用来建模学习施事者 O_1 与事件词 P 以及受事者 O_2 与事件词 P 之间的关系。 O_1T_1P 和 PT_2O_2 则被用来分别生成两个事件角色相关的向量 R_1 和 R_2。第三个张量 T_3 则被用来将 R_1 和 R_2 进行最后的语义合成并生成事件 E = (O_1, P, O_2)最终的向量 U。

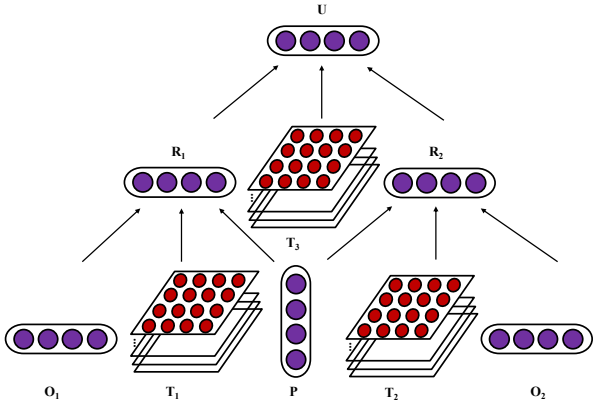


图 6 基于张量神经网络的事件表示学习模型

张量神经网络（Neural Tensor Network， NTN）的输入是词向量，输出是事件向量。可以利用 Mikolov 提出的 Word2Vec 模型中的 skip-gram 算法，从大规模的新闻语料中学习到最初始的词向量（维度为 d＝100）。由于事件元素可能会包含多个词汇，可以采用各个词汇向量的平均值来生成最终的事件元素初始向量，这样做的好处是可以让无论是短语还是单一词汇都具有同样维度的向量表示(例如， 诺基亚移动手机业务和诺基亚)。

从图 6 中可以看出， $R_1 \in \mathbb{R}^d$是由下式计算得到：

$$R_1 = f(O_1^T T_1^{[1:k]} P + W \begin{bmatrix} O_1 \\ P \end{bmatrix} + b)$$

其中， $T_1^{[1:k]} \in \mathbb{R}^{d \times d \times k}$是一个张量，并且双线性张量乘积$O_1^T T_1^{[1:k]} P$结果是一个向量$\mathrm{r} \in \mathbb{R}^k$，其中每一个向量维度都由一片张量计算得到($r_i = O_1^T T_1^{[i]} P, i = 1, \cdots, k$)。张量神经网络中的其他参数都是反向传播神经网络中的常规参数，其中$\mathrm{W} \in \mathbb{R}^{k \times 2d}$是权重矩阵， $\mathrm{b} \in \mathbb{R}^k$是偏

置向量，$f = tanh$ 是激活函数。R_2 和 U 的计算方式与 R_1 完全一致。

参考文献

[Doddington et al. 2004] Doddington D, Mitchell A, Przybocki M, et al. The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. LREC 2004, Proceedings of the Fourth International Conference on Language Resources and Evaluation. 2004.

[Chambers 2013] Chambers N. Event schema induction with a probabilistic entity-driven model. EMNLP 2013, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013.

[Chambers & Jurafsky 2011] Chambers N, Jurafsky D. Template-Based Information Extraction without the Templates[C]// ACL 2011, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011.

[Cheung et al. 2013] Cheung J, Poon H, Vanderwende L. Probabilistic Frame Induction. NAACL 2013, Proceedings of NAACL-HLT 2013.

[Nguyen et al. 2015] Nguyen K H, Tannier X, Ferret O, et al. Generative Event Schema Induction with Entity Disambiguation. ACL 2015, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. 2015.

[Liu et al. 2019] Liu X, Huang H, Zhang Y. Open Domain Event Extraction Using Neural Latent Variable Models. ACL 2019, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[Balasubramanian et al. 2013] Balasubramanian N, Soderland S, Mausam, et al. Generating Coherent Event Schemas at Scale. EMNLP 2013, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013.

[Sha et al. 2016] Sha L, Li S, Chang B, et al. Joint Learning Templates and Slots for Event Schema Induction. NAACL 2016, Proceedings of NAACL-HLT 2016. 2016.

[Huang et al. 2016] Huang L, Cassidy T, Feng X, et al. Liberal Event Extraction and Event Schema Induction. ACL 2016, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016.

[Li et al. 2020] Li M, Zeng Q, Lin Y, et al. Connecting the Dots: Event Graph Schema Induction with Path Language Modeling. EMNLP 2020, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[Li et al. 2021] Li M, Li S, Wang Z, et al. The Future is not One-dimensional: Complex Event Schema Induction by Graph Modeling for Event Prediction. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 5203-5215.

[Riloff 2013] Riloff E. Automatically constructing a dictionary for information extraction tasks. AAAI. 1993, 1(1): 2.1.

[Kim & Moldovan 1995] Kim J T, Moldovan D I. Acquisition of linguistic patterns for knowledge-based information extraction. IEEE transactions on knowledge and data engineering, 1995, 7(5): 713-724.

[Riloff & Shoen 1995] Riloff E, Shoen J. Automatically acquiring conceptual patterns without an annotated corpus. Third Workshop on Very Large Corpora. 1995.

[Joyce 1998] Chai J Y, Biermann A W, Guinn C I. Syntactic Generalization and Two-dimensional Generalization in Information Extraction. 1998.

[Yangarber 2017] Yangarber R. Scenario customization for information extraction. DEFENSE ADVANCED RESEARCH PROJECTS AGENCY ARLINGTON VA, 2001.

[姜吉发 2004] 姜吉发. Research on the information extraction pattern of free text. Chinese Academy of Sciences, 2004.

[Chieu & Ng 2002] Chieu H L, Ng H T. A maximum entropy approach to information extraction from semi-structured and free text. Aaai/iaai, 2002, 2002: 786-791.

[Grishman 2005] Grishman R, Westbrook D, Meyers A. Nyu's english ace 2005 system description. ACE, 2005, 5.

[Ahn 2006] Ahn D. The stages of event extraction. Arte'06 Proceedings of the Workshop on Annotating & Reasoning About Time & Events, 2006:1-8.

[Chen 2009] Chen Z, Ji H. Language specific issue and feature exploration in Chinese event extraction. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. 2009: 209-212.

[Heng & Grishman 2008] Heng Ji and Ralph Grishman. 2008. Refining Event Extraction through Unsupervised Cross-document Inference. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, pages 254-262. Ohio, USA.

[Heng 2009] Heng Ji. 2009. Cross-lingual Predicate Cluster Acquisition to Improve Bilingual Event Extraction by Inductive Learning. In Proceedings of the NAACL HLT Workshop on

Unsupervised and Minimally Supervised Learning of Lexical Semantics, pages 27-35. Boulder, Colorado.

[Liao & Grishman 2010] Shasha Liao and Ralph Grishman. 2010. Filtered Ranking for Bootstrapping in Event Extraction. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 680–688. Beijing, China.

[Hong et al. 2011] Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou and Qiaoming Zhu. 2011. Using Cross-Entity Inference to Improve Event Extraction. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon.

[Patwardhan & Riloff 2009] Siddharth Patwardhan and Ellen Riloff. 2009. A unified model of phrasal and sentential evidence for information extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 151–160, Singapore. Association for Computational Linguistics.

[Huang & Riloff 2011] Ruihong Huang and Ellen Riloff. 2011. Peeling back the layers: Detecting event role fillers in secondary contexts. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1137–1147, Portland, Oregon, USA. Association for Computational Linguistics.

[Huang & Riloff 2012] Ruihong Huang and Ellen Riloff. 2012. Modeling textual cohesion for event extraction. In Twenty-Sixth AAAI Conference on Artificial Intelligence.

[Du et al. 2020] Du X, Cardie C. Document-Level Event Role Filler Extraction using Multi-Granularity Contextualized Encoding. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8010-8020.

[Chen et al. 2017] Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. In ACL 2017.

[Yang et al. 2018] Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. DCFEE: A document-level chinese financial event extraction system based on automatically labeled training data. In Proceedings of ACL 2018, System Demonstrations.

[Zheng et al. 2019] Zheng S, Cao W, Xu W, et al. Doc2EDAG: An End-to-End Document-level Framework for Chinese Financial Event Extraction. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 337-346.

[Kaplan & Rogghe 1991] R.M. Kaplan, and G. Berry-Rogghe. Knowledge-based acquisition of

causal relationships in text. In Knowledge Acquisition, 3(3), 1991.

[Khoo et al. 2000] C. Khoo, S. Chan and Y. Niu. Extracting Causal Knowledge from a Medical Database Using Graphical Patterns. In Proceedings of ACL, Hong Kong, 2000.

[Girju et al. 2003] Girju, Roxana. "Automatic detection of causal relations for question answering." Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering. 2003.

[Do et al. 2011] Do, Quang, Yee Seng Chan, and Dan Roth. "Minimally supervised event causality identification." Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011.

[Riaz & Girju 2013] Riaz, Mehwish, and Roxana Girju. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. Proceedings of the SIGDIAL 2013 Conference. 2013.

[Hashimoto et al. 2014] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, Istv´an Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 987–997.

[Gao et al. 2019] Gao, Lei, Prafulla Kumar Choubey, and Ruihong Huang. "Modeling document-level causal structures for event causal relation identification." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019.

[Liu et al. 2020] Liu, Jian, Yubo Chen, and Jun Zhao. Knowledge Enhanced Event Causality Identification with Mention Masking Generalizations. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence Main track. Pages 3608-3614.

[Kadowaki et al. 2019] Kadowaki, Kazuma, et al. "Event causality recognition exploiting multiple annotators' judgments and background knowledge." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[Li et al. 2020] Zhongyang Li et al. "Guided Generation of Cause and Effect.." International Joint Conference on Artificial Intelligence (2020).

[Li et al. 2021] Zhongyang Li et al. "CausalBERT: Injecting Causal Knowledge Into Pre-trained

Models with Minimal Supervision.” arXiv: Computation and Language (2021).

[Bethard et al. 2007] Steven Bethard, James H Martin, and Sara Klingenstein. 2007. Timelines from text: Identification of syntactic temporal relations. In IEEE International Conference on Semantic Computing (ICSC). Pages 11–18.

[Glavaš et al. 2014] Goran Glavaš, Jan Šnajder, Parisa Kordjamshidi, and Marie-Francine Moens. 2014a. Hieve: A corpus for extracting event hierarchies from news stories. In Proceedings of 9th Language Resources and Evaluation Conference (LREC), pages 3678–3683.

[Hovy et al. 2013] Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity. In Proceedings of the Workshop on Events: Definition, Detection, Coreference, and Representation, pages 21–28, Atlanta, Georgia.

[Ge et al. 2018] Ge, Tao, et al. "SeRI: A Dataset for Sub-event Relation Inference from an Encyclopedia." CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2018.

[Ding et al. 2015] Xiao Ding, Yue Zhang, Ting Liu, Junwen Duan. Deep Learning for Event-Driven Stock Prediction. In Proc. IJCAI 2015.

第六章 知识融合

胡伟[1]，漆桂林[2]

1. 南京大学 计算机软件新技术国家重点实验室，南京 210023

2. 东南大学 计算机科学与工程学院，南京 211189

一、任务定义、目标和研究意义

知识图谱以符号化的方式描述真实世界中的实体及其属性和相互关系，并将它们组织成事实三元组的结构。时至今日，知识图谱已成为各类知识驱动人工智能方法的重要资源，涵盖了包括社交网络、生物医学、地理信息、电子商务、电影音乐等众多领域，支撑语义搜索、智能问答、推荐系统、大数据分析等智能应用。

知识图谱可能由不同的机构和个人构建，同时，构建知识图谱的数据可能有各种来源，导致不同的知识图谱之间存在多样性和异构性。例如，对于不同的相关领域（甚至是相同领域），通常会存在多个不同的实体指称真实世界中的相同事物。

知识融合旨在将不同知识图谱融合为一个统一、一致、简洁的形式，为使用不同知识图谱的应用间的交互建立互操作性。知识融合常见的研究内容包括：本体匹配（也称为本体映射）、实体对齐（也称为实例匹配、实体消解）、真值发现（也称为真值推断）以及实体链接等，面临的核心挑战主要包括大规模、异构性、低资源等问题。

知识融合是知识图谱研究中的一个核心问题。知识融合研究有助于提升基于知识图谱的信息服务水平和智能化程度，推动人工智能、自然语言处理、语义网、数据库等相关领域的技术进步，具有重要的理论价值和广泛的应用前景，可以创造巨大的社会和经济效益。

二、研究内容和关键科学问题

图 1 展示了一个知识融合的常见流程，下面将分别概述主要研究内容和近期发展趋势。

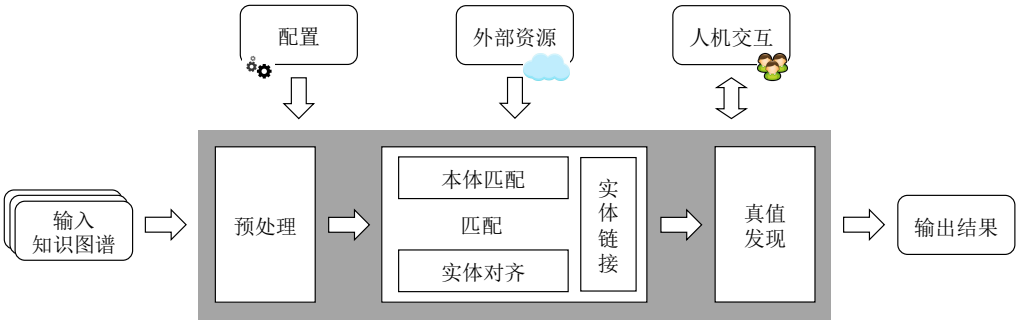


图 1 知识融合的常见流程

- 预处理主要包括预先对输入的知识图谱进行清洗和后续步骤的准备。清洗主要为了解决输入的质量问题，而后续步骤的准备通常使用分块（blocking）技术，通过对索引的设计，可以避免在匹配环节达到知识图谱规模的平方级复杂度。这里的一个关键问题是对分块大小和数量的权衡，在尽量不丢失可能结果的情况下使分块尽可能的小。
- 根据匹配对象的不同，匹配一般分为本体匹配、实体对齐以及实体链接等方面。本体匹配侧重发现知识图谱模式层的等价或相似的类、属性或关系，实体对齐侧重发现指称真实世界相同个体的实例，而实体链接则将自然语言文本中的实体提及（mention）链接到知识图谱中的实体节点。如何从语义上消解对象之间的异构性是匹配环节待解决的关键科学问题。
- 在匹配的基础上，真值推断的主要目标是从不一致的数据中推测出真值，以实现多源异构知识的关联与合并，最终形成一个一致的结果。研究的关键在于如何综合判断数据源的可靠性和数据值的可信度。

三、技术方法和研究现状

受限于篇幅，本节仅介绍知识融合方向的近期研究动向和一些代表性技术方法，更早的工作请参见《知识图谱发展报告（2018）》以及其他研究综述。

1. 本体匹配

本体匹配的目标是建立不同本体概念之间的语义映射[Euzenat & Shvaiko, 2013]。近年来，关于本体匹配的研究进展不多。早期的一些代表性工作包括 RiMOM [Li et al., 2008]、Falcon-AO [Hu & Qu, 2008]等。值得一提的是，LogMap [Jiménez-Ruiz & Grau, 2011]获得了2021 年语义网科学联盟（SWSA）颁发的十年最具影响力论文奖。LogMap 是一个高度可扩展的本体匹配系统，总体流程如图 2 所示。它可以高效地匹配包含数万（甚至数十万）类别的本体，也可以利用复杂的推理和修复技术来减少逻辑不一致性的数量，还可以在匹配过程中支持用户的可视化干预。近年来，LogMap 也将表示学习技术集成到本体匹配任务中。

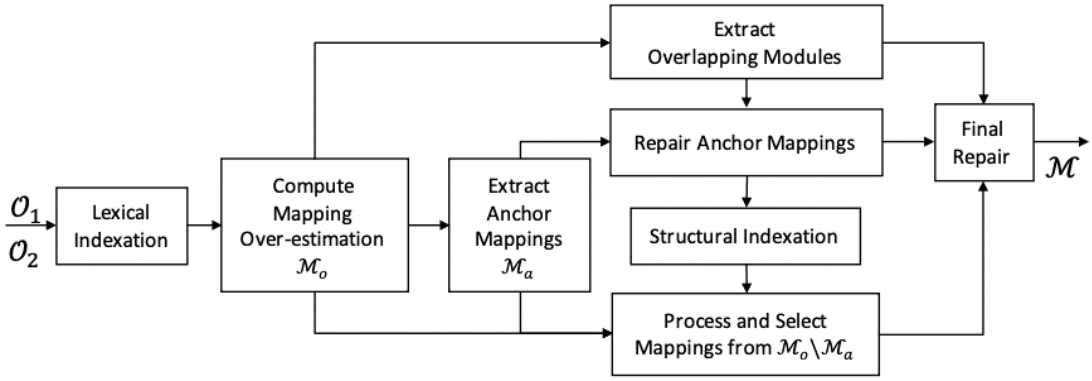


图 2 本体匹配方法 LogMap [Chen et al., 2021]

2. 实体对齐

1）基于表示学习的实体对齐

近年来，以知识图谱表示学习为基础的实体对齐方法逐渐成为主流。如图 3 所示，基于表示学习的实体对齐框架主要包含 2 个主要模块[Sun et al., 2020; Zhao et al., 2020]：表示学习模块将单个知识图谱嵌入到向量空间，多数方法采用基于几何运算的模型，也有工作使用图神经网络等。对齐模块使用先验知识或人工标注得到少量先验对齐进行训练，再使用常用的向量度量函数对齐实体的表示，或者寻找全局最优的集体实体对齐结果。还有一些工作采用迭代的方式不断选择新发现的实体对齐来扩充训练样本。表示学习模块与对齐模块之间存在两种典型的交互方式：一种是将不同知识图谱嵌入到统一的向量空间，另一种则是学习不同知识图谱向量空间之间的映射关系。

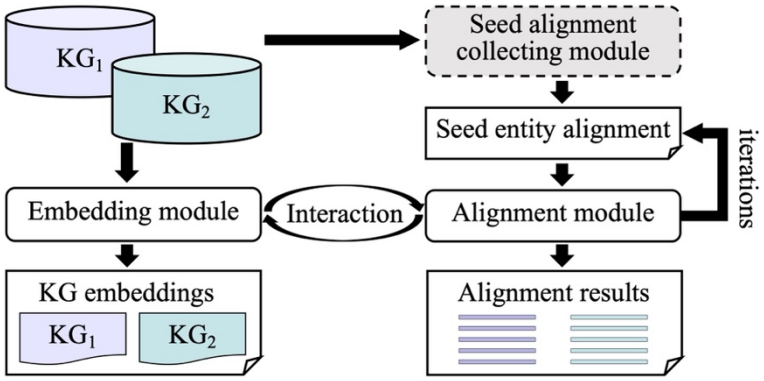


图 3 基于表示学习的实体对齐框架[Sun et al., 2020]

Dual-AMN [Mao et al., 2021] 是近期的一个代表性方法，其在降低模型计算复杂度的情况下保持了对知识图谱内和知识图谱间信息的建模。具体地，Dual-AMN 设计了一个基于关系型注意力的卷积层用于捕捉单个知识图谱内的结构信息。针对知识图谱间的对齐信息，Dual-AMN 设置了一组代理向量隐式地表示图谱之间的对齐关系，并通过代理匹配注意力机制来捕捉。

除了面向常规实体对齐场景的方法，一些研究工作也尝试考虑更具挑战性的新场景。DiNGAl [Yan et al., 2021]首次提出了动态实体对齐任务，改变了常规场景中知识图谱是静态的假设，认为图谱事实是会动态演变的，因此表示学习模型需要针对不断变化的图结构信息对实体表示进行更新。针对该挑战，DiNGAl 先基于拓扑无关的掩码门控机制得到静态的实体表示，再采用局部更新策略对动态过程中受影响的实体表示进行修正。由于动态过程中也会出现新的先验对齐，DiNGAl 将这部分新的对齐作为正例进行训练，从而对所有实体表示进行更新，避免了从头训练的开销。

知识图谱中的事实具有时效性，而现有的实体对齐方法完全忽视了时间信息。针对该问题，TEA-GNN [Xu et al., 2021]提出了面向时序知识图谱的实体对齐任务，使用开始时间戳和结束时间戳表示时间信息，并基于图神经网络将不同知识图谱中的实体、关系、时间戳嵌入到统一的向量空间中，整体框架如图 4 所示。TEA-GNN 首先为关系和时间戳分配不同的正交矩阵用于获得实体的邻居信息，然后在聚合时使用了一种时间感知的注意力机制来区分不同邻居的重要性。为了进一步集成时间信息，TEA-GNN 还将实体表示和相邻的时间表示之和进行拼接，从而得到最终的实体表示。

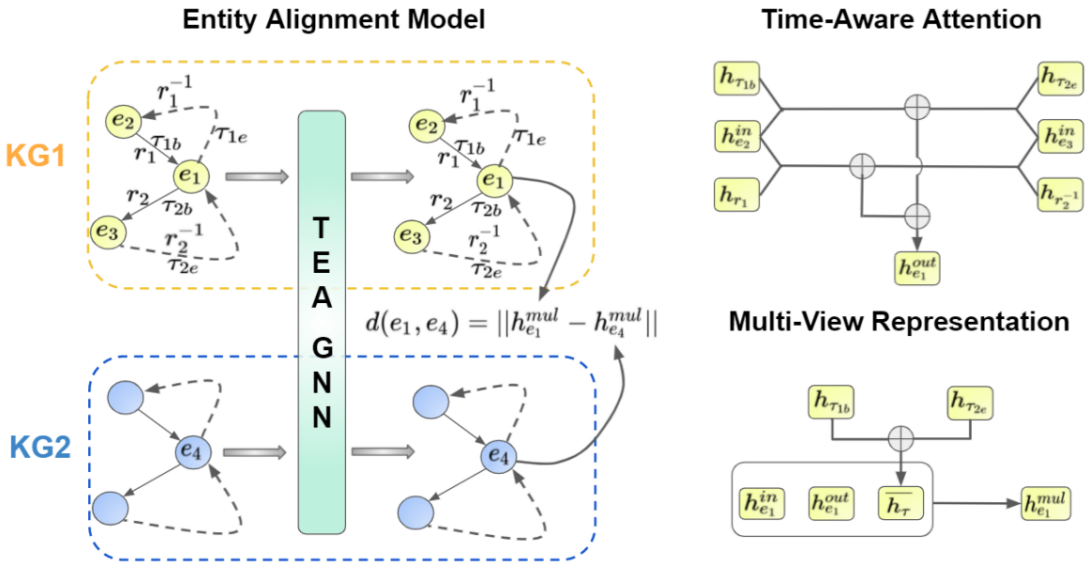


图 4 面向时序知识图谱的实体对齐方法 TEA-GNN [Xu et al., 2021]

2）基于人机协作的实体对齐

基于人机协作的实体对齐方法通过付出较小的人工代价来获得丰富的标注数据，从而提高模型的性能。

常见方法先构建实体对标签的推断结构，然后由用户标注推断效用最大的未知实体对，并进行推断。Power [Chai et al., 2018]计算每对实体在不同属性上的相似度并将它们拼接成相似度向量，通过向量划分算法构造偏序结构，让用户标注偏序中前驱和后继总数最多的实

体对。Remp [Huang et al., 2020]将实体对用对齐好的关系连接构成实体消解图，再基于实体对之间的关系建立概率传播模型，通过错误容忍的真值推断策略以及最优化问题选择算法来最大化收益期望。

近年来，一些工作也尝试将深度神经网络和人机协作方法相结合。DTAL [Kasai et al., 2019]基于迁移学习初始化模型参数，并根据深度模型输出的熵挑选出候选对齐用于标注。ActiveEA [Liu et al., 2021a]提出了一种结构感知的不确定性采样策略，用于度量每个实体的对齐不确定性以及对周围邻居的影响程度。考虑到有些孤立实体在对应知识图谱内不存在可与之对齐的实体，ActiveEA 还设计了一种孤立实体识别器，从而减少对这部分实体采样而造成的偏差。

RAC [Zeng et al., 2021] 进一步探索了深度强化学习与主动学习技术的结合，整体框架如图 5 所示。基于度数、PageRank 值和信息熵，RAC 设计了 3 种查询策略。考虑到不同迭代轮次中不同查询策略的重要性会有所不同，且单个查询策略不能满足所有数据集的需要，RAC 采用多臂老虎机策略自适应地混合 3 种查询策略，并挑选出最优查询供人工标注。

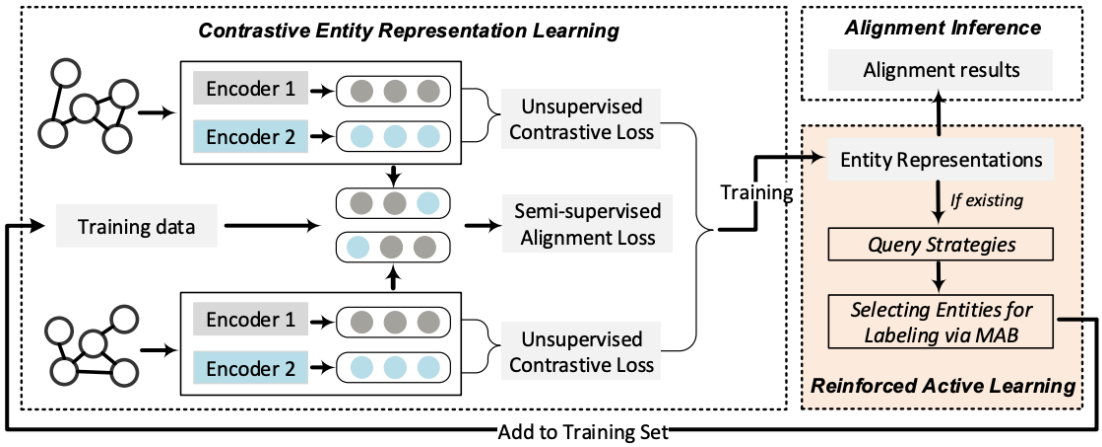


图 5 基于人机协作的实体对齐方法 RAC [Zeng et al., 2021]

考虑到潜在的人工标注成本，有工作开始探索不利用任何标签信息的实体对齐方法。SelfKG [Liu et al., 2022] 设计了一种自监督实体对齐算法，其利用预训练语言模型将不同知识图谱中的实体映射到一个统一的向量空间中，并以此捕捉实体的语义相似度。为了避免利用标签信息，SelfKG 拉远随机采样到的负例实体对的表示，以此达到拉近潜在正例实体对的效果。为了避免随机采样出假负例，其只在实体所在的知识图谱中进行负例采样。在基准数据集上，该方法优于众多监督方法，展现了将自监督学习应用于实体对齐的潜力。

3）多模态实体对齐

考虑到图像特征可以在一定程度上帮助消歧，近期一些工作引入图像模态，并将多种模态的信息进行融合，基于多模态的实体对齐逐渐成为一个新的研究热点。

MMEA [Chen et al., 2020] 较早地在实体对齐中考虑了图像特征空间。总体框架如图 6 所示，主要包含两个模块：多模态知识嵌入用于获得实体在不同模态下的向量表示，其中使用 TransE 生成结构特征，使用 VGG16 获得图像特征。MMEA 还额外考虑了数值型属性，并利用径向基函数（radial basis function）神经网络生成该模态的向量表示。在多模态知识融合模块，MMEA 认为每个模态下的向量表示来自于不同的特征空间，因而设置了一个公共特征空间，并要求不同模态下的向量表示与公共空间下的向量表示尽可能接近，以此实现不同模态信息的互补。

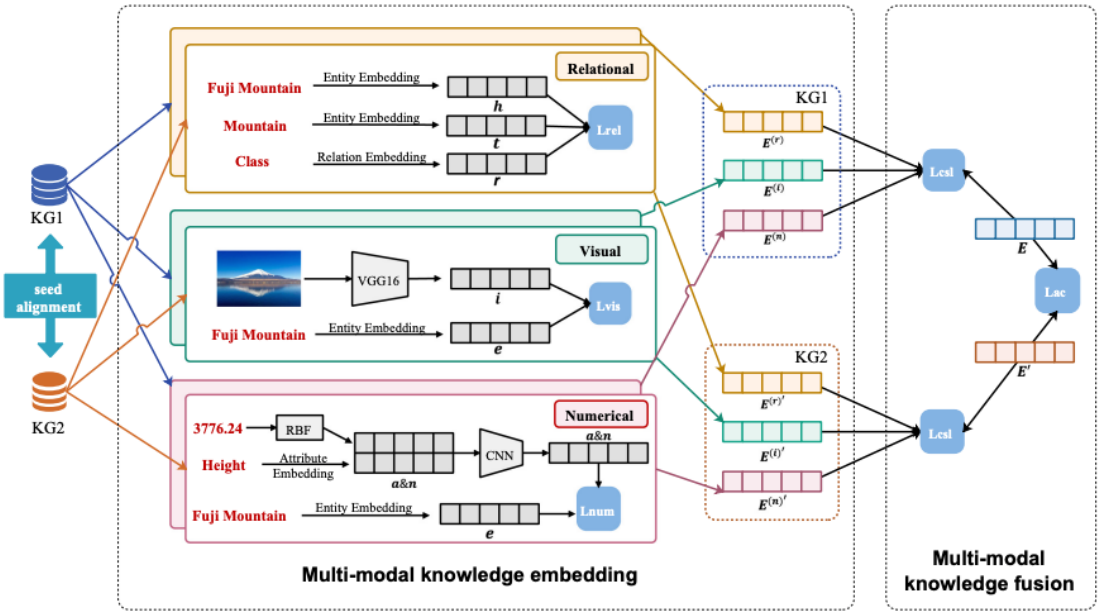


图 6 基于多模态的实体对齐方法 MMEA [Chen et al., 2020]

EVA [Liu et al., 2021b] 采取了类似的建模思路，使用 ResNet-152 对图像特征进行初始化，并基于 HMAN [Yang et al., 2019a]得到关系特征与属性特征。进一步地，EVA 设计了一种基于注意力机制的多模态加权策略以实现多模态信息融合。此外，EVA 还探索了多模态技术在无监督实体对齐场景下的可能性，实验结果表明仅利用图像相似度生成初始实体对的性能能够逼近有监督场景下的表现。

3. 真值发现

真值发现一般通过冲突检测、真值推断等技术消除知识融合过程中的冲突，再对知识进行关联与合并，最终形成一个一致的结果。如何处理多源数据中的冲突是真值发现的主要研究问题[Li et al., 2015]。例如，不同数据源可能对珠穆朗玛峰的高度有不同的描述，其中有些可能是不准确的，需要推断。常见的方法包括 3 类：第一类是迭代方法，例如 TruthFinder [Yin et al., 2008]、Investment [Pasternack & Roth, 2010]和 ACCU [Dong et al., 2009]，其将数据来源纳入考量，迭代评估数据源的可靠性与数据值的可信度直至收敛。第二类是优化方法，

例如[Li et al., 2014a; Li et al., 2014b; Aydin et al., 2014]，其通过最小化带权整体推断误差，使得真值向可靠性高的数据源所提出的值靠近，同时距离较远的数据源会在优化过程中被分配较小的权重作为其可靠性。最后一类是概率图模型，例如 SimpleLCA [Pasternack & Roth, 2013]和 OKELE [Cao et al., 2020]，其对影响数据源可靠性的潜在因素进行假设并利用贝叶斯网络等模型对随机变量及其依赖关系进行建模。由于迭代和优化方法中的一系列计算规则以及概率图方法中的各种影响因素需要人为设置，常常不能真实反映各种场景下的潜在数据分布与影响。

近年来一些工作运用深度学习探索真值推断问题。CASE [Lyu et al., 2019]基于数据源—数据值、数据源—数据源以及真值—数据值之间的关联来构建异构信息网络，将真值发现建模成异构信息网络的表示学习问题，即通过节点的表示来拟合节点之间边的存在性。同时，CASE 根据数据源的表示来建模它们在不同目标上数据值的相似性，并使用 beta 分布来解决数据稀疏性问题。最终，CASE 利用已知真值进行半监督学习得到网络元素的表示，将与真值的表示最接近的数据值选作真值。

BAT [Liu et al., 2021c] 将数据源和推断目标及其之间的关联建模成二部图，基于图自编码器和数据源之间的关联性得到数据源的初始特征，基于预训练文本或图像信息编码器得到带推断目标的初始特征。BAT 先通过注意力机制计算节点之间的关联性，再使用二部图卷积网络同时聚合这些信息得到数据源、推断目标和边的信息。最后，基于图卷积网络聚合的信息预测推断目标的真值，并通过真值进行训练。

此外，还有工作针对批量或流式数据研究快速更新数据源可靠性和实体真值的方法。EvolveT [Zhi et al., 2018]注意到同一推断目标在不同时间点的真值之间具有关联性，因此引入了马尔可夫模型，即下一时刻的真值可以通过当前真值和一个固定的转移矩阵来确定。EvolveT 基于卡尔曼滤波与平滑器设计了一种线性时间的在线参数估计算法，实现快速高效地估计真值。

4. 实体链接

实体链接通常建立在实体识别任务之上，需要预先识别文本中的命名性实体的提及文本，然后根据该提及枚举知识图谱中可能的候选实体，并利用排序的方式从中挑选出最符合当前语境的实体作为链接结果[Shen et al., 2014]。由于自然语言的多样性和模糊性，实体的表述往往具有较高的歧义性，这使得实体链接方法通常需要处理“一词多义”和“多词同义”两种歧义性问题。“一词多义”是指同一个实体名称可以表示多个实体的情况，例如，给定自然语言文本“苹果发布了最新的手机产品 iPhone 13”，实体链接方法需要将其中的“苹果”链接到实体“苹果 Apple（企业）”，而非实体“苹果（水果）”。“多词同义”则是指一

个实体可以用多个名称来表示的情况，例如，“自然语言处理”和“NLP”都可以用来表示“自然语言处理（领域）”这个实体。

一个完整的实体链接方法通常包括 4 个步骤：（1）实体提及识别，即利用字符串比较、机器学习等方法，从给定的文本序列中识别出描述实体的单词或短语。（2）候选实体生成，即根据已识别出的实体提及，从海量的实体集合中选出有限数量的候选实体，可以划分为基于字符串匹配、基于资源扩展别名以及基于先验概率计算 3 种方法。（3）候选实体排序，即结合上下文语境，对实体提及和候选实体进行相似度判断，并按照相似度得分进行排序，可以划分为基于统计的方法和基于深度学习的方法。（4）不可链接提及预测。由于知识图谱的不完备性，部分实体在知识图谱中并不存在，因此需要判断实体提及是否链接到不存在的实体。下面介绍近期的一些代表性方法：

BLINK [Wu et al., 2020] 是由 Facebook 提出的一种两阶段零样本实体链接模型。其首先使用双向编码器来编码文本提及和实体描述，并使用两个独立的 BERT 来分别获得提及和实体的表示向量，将二者的点积作为候选实体得分。接着，使用一个基于 BERT 的交叉编码器来同时编码提及和实体，随后接入一个线性层计算出最后的实体得分并进行排序，取得分最高的候选实体作为预测的链接结果。

CHOLAN [Ravi et al., 2021] 使用 Transformer 编码器来进行端到端的实体链接，其架构如图 7 所示。CHOLAN 认为现有的预训练模型（例如 BERT）虽然在大型语料库上进行了预训练，但是在具体任务中仍需考虑额外的上下文信息。CHOLAN 首先利用 BERT 识别输入句子中的提及，然后利用工具 Falcon [Sakor et al., 2019] 和 DCA [Yang et al., 2019b] 为每个提及生成为知识库中的实体候选，最后将实体提及、句子、实体候选以及 Wikipedia 中关于实体的描述信息拼接起来输入另一个 BERT，从而预测出链接的实体。

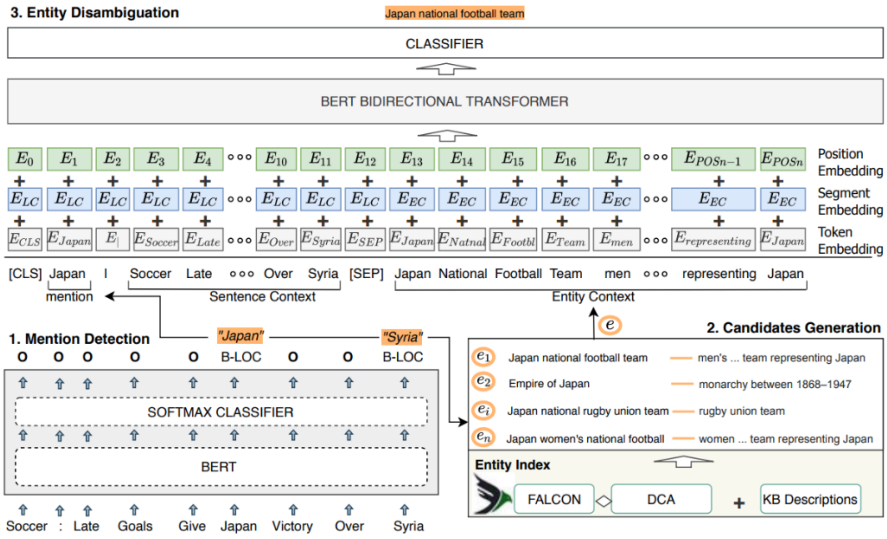


图 7 端到端实体链接方法 CHOLAN [Ravi et al., 2021]

REL [van Hulst et al., 2020] 利用先进的命名实体识别模型 Flair [Akbik et al., 2018] 来识别实体提及。针对候选生成，REL 首先利用 Wikipedia 和 CrossWikis 的超链接数量来预估每个(提及, 实体)对的先验概率，然后根据该概率选取排名靠前的实体作为候选实体。之后，再利用相似度度量函数，从提及的附近单词中选取相似度最大的几个实体作为候选实体。最后，基于先验的重要程度、上下文相似度以及文档中其他实体链接的一致性，对所有候选实体进行排序与消歧。EntQA [Zhang et al., 2022] 将提及检测和实体消歧两个子任务的顺序进行颠倒，并将整个链接任务建模为一个开放域问答任务。EntQA 采用 Retriever-Reader 的框架，利用知识图谱中实体的标题和描述来建模实体。Retriever 模块计算文本片段和实体之间的相似性评分，快速地生成多个候选实体；Reader 模块以文档、文本片段和候选实体为输入，建模出候选实体对应于提及的概率以及该候选实体为正确实体的概率，进而预测出实体链接结果。

5. 工具软件和评测数据集

就本体匹配而言，一些常见的本体匹配工具和系统可以从 OAEI[1]（Ontology Alignment Evaluation Initiative）网站上获得。面向实体对齐，OpenEA[2]是一个最新的基于表示学习的实体对齐开源软件库，总体框架如图 8 所示。OpenEA 目前集成了 12 种代表性实体对齐方法，同时它使用了一个灵活的软件架构，可以较容易地集成大量现有的表示学习模型。另一个类似的开源软件库是 EAkit[3]。而面向真值发现，CrowdTruthInference[4]集成了 17 种真值推断算法，支持是否判断、单项选择和数值估计 3 种类型任务的真值推断。

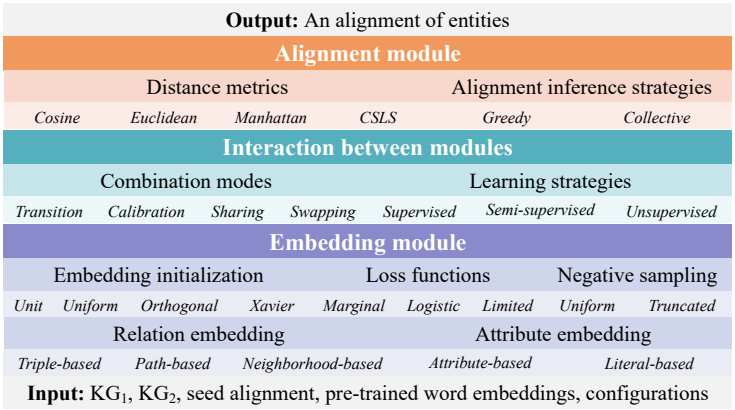


图 8 基于表示学习的实体对齐开源软件库 OpenEA [Sun et al., 2020]

[1] http://oaei.ontologymatching.org/

[2] https://github.com/nju-websoft/OpenEA

[3] https://github.com/THU-KEG/EAkit

[4] https://zhydhkcws.github.io/crowd_truth_inference

标准的评测数据集对于知识融合也十分重要，它们提供了一个横向比较各种方法性能优劣的平台。随着知识融合研究的蓬勃发展，除了传统的 OAEI 评测数据集，也出现了一些新的数据集，简单介绍如下：

- 面向实体对齐，DBP15k 数据集包含 3 个从多语言版本 DBpedia 构建的跨语言数据集，分别是中文到英文、日语到英文以及法语到英文。DYW100k 则包含两个从 DBpedia、Wikidata 和 YAGO3 抽取出的大规模数据集 DBP-WD 和 DBP-YG。由于上述这些数据集缺乏悬挂实体[Sun et al., 2021]，一个新的基于多语言版本 DBpedia 的实体对齐数据集 DBP 2.0 被构建。
- 实体链接技术的重要性和实用性得到了工业界和学术界的广泛关注，通过 AIDA、AQUAINT、ACE 等评测竞赛构建了 AIDA CoNLL-YAGO、TAC KBP 等经典数据集以及 WNED-CWEB、WNED-WIKI 等新数据集，同时也催生出 TagMe [Ferragina & Scaiella, 2010]、AGDISTIS [Usbeck et al., 2014]、REL [van Hulst et al., 2020]等优秀的开源实体链接框架。
- 另外，一批公开的面向图像、文本、数值等不同领域和任务类型的真值发现数据集可以从如下网站访问：http://dbgroup.cs.tsinghua.edu.cn/ligl/crowddata。

四、技术展望

在过去的几年里，表示学习技术被广泛运用于知识融合相关研究，未来可能的研究方向包括：

- 预训练语言模型在自然语言处理领域中取得了巨大成功。受此启发，针对大规模知识图谱进行预训练成为了未来的一个潜在研究方向，预训练得到的知识同样可以迁移至下游诸多任务。例如在实体对齐中，大规模知识图谱的表示学习可以得到实体的通用知识信息，一定程度缓解了下游实体对齐中知识不充分的问题，如实体缺失部分模态信息。同样地，在多语言实体链接中，预训练得到的高资源语言知识可以间接帮助低资源语言的实体链接。然而相关技术的探索也存在一定挑战，例如如何利用知识融合技术对异构的知识图谱进行融合从而在更大规模的知识图谱上开展预训练很值得研究。
- 知识融合的研究问题近年来也有了一些新设定。例如，知识可能会随着时间变化，未来的工作可以考虑面向流式数据的动态实体对齐和真值发现技术，得到更多准确的事实，用来补充动态知识图谱。又如，也可以考虑利用动态知识图谱表示学习技术为动态真值发现提供真值的先验知识，以提高真值发现的准确性。

- 在评测数据集方面，现有的研究工作主要基于一些小规模数据集进行评测，比如实体对齐的 DBP15K 数据集、实体链接的 TAC KBP 数据集等。然而，这些数据集的构建已有一段时间，已经显现出一定的滞后性；同时数据集的规模较小，覆盖面较窄，与真实世界存在一定的差别。因此，未来需要考虑如何结合现阶段的研究进展，针对诸如多模态实体对齐、复杂事实真值推断、跨语言实体链接等新任务，开发出规模更大、质量更高的大规模评测数据集，从而更专业、更全面地评测知识融合领域的工作。

参考文献

[Akbik et al., 2018] Alan Akbik, Duncan Blythe, Roland Vollgraf. Contextual string embeddings for sequence labeling. In: COLING, 1638-1649, 2018

[Aydin et al., 2014] Bahadir Ismail Aydin, Yavuz Selim Yilmaz, Yaliang Li, Qi Li, Jing Gao, Murat Demirbas. Crowdsourcing for multiple-choice question answering. In: AAAI, 2946-2953, 2014

[Cao et al., 2020] Ermei Cao, Difeng Wang, Jiacheng Huang, Wei Hu. Open knowledge enrichment for long-tail entities. In: WWW, 384-394, 2020

[Ceccarelli et al., 2013] Diego Ceccarelli, Claudio Lucchese, Raffaele Perego, Salvatore Orlando, Salvatore Trani. Dexter: An open source framework for entity linking. In: ESAIR, 17-20, 2013

[Chai et al., 2018] Chengliang Chai, Guoliang Li, Jian Li, Dong Deng, Jianhua Feng. A partial-order-based framework for cost-effective crowdsourced entity resolution. Proceedings of the VLDB Endowment, 27(6):745-770, 2018

[Chen et al., 2020] Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, Enhong Chen. MMEA: Entity alignment for multi-modal knowledge graph. In: KSEM, 134-147, 2020

[Chen et al., 2021] Jiaoyan Chen, Ernesto Jiménez-Ruiz, Ian Horrocks, Denvar Antonyrajah, Ali Hadian, Jaehun Lee. Augmenting ontology alignment by semantic embedding and distant supervision. In: ESWC, 392-408, 2021

[Dong et al., 2009] Xin Luna Dong, Laure Berti-Equille, Divesh Srivastava. Integrating conflicting data: The role of source dependence. Proceedings of the VLDB Endowment, 2(1):550-561, 2009

[Euzenat & Shvaiko, 2013] Jérôme EuzenatPavel Shvaiko. Ontology matching. Springer, 2013

[Ferragina & Scaiella, 2010] Paolo Ferragina, Ugo Scaiella. TAGME: On-the-fly annotation of short text fragments (by wikipedia Entities). In: CIKM, 1625-1628, 2010

[Hu & Qu, 2008] Wei Hu, Yuzhong Qu. Falcon-AO: A pratical ontology matching system. Journal

of Web Semantics, 6(3):237-239, 2008

[Huang et al., 2020] Jiacheng Huang, Wei Hu, Zhifeng Bao, Yuzhong Qu. Crowdsourced collective entity resolution with relational match propagation. In: ICDE, 37-48, 2020.

[Jiménez-Ruiz & Grau, 2011] Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau. LogMap: Logic-based and scalable ontology matching. In: ISWC, 273-288, 2011

[Kasai et al., 2019] Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, Lucian Popa. Low-resource deep entity resolution with transfer and active learning. In: ACL, 5851-5861, 2019

[Li et al., 2008] Juanzi Li, Jie Tang, Yi Li, Qiong Luo. RiMOM: A dynamic multistrategy ontology alignment framework. IEEE Transactions on Knowledge and Data Engineering, 21(8):1218-1232, 2008

[Li et al., 2014a] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. Proceedings of the VLDB Endowment, 8(4):425-436, 2014

[Li et al., 2014b] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, Jiawei Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: SIGMOD, 1187-1198, 2014

[Li et al., 2015] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, Jiawei Han. A survey of truth discovery. ACM SIGKDD Explorations Newsletter, 17(2):1-16, 2015

[Liu et al., 2021a] Bing Liu, Harrisen Scells, Guido Zuccon, Wen Hua, Genghong Zhao. ActiveEA: Active learning for neural entity alignment. In: EMNLP, 3364-3374, 2021

[Liu et al., 2021b] Fangyu Liu, Muhao Chen, Dan Roth, Nigel Collier. Visual pivoting for (unsupervised) entity alignment. In: AAAI, 4257-4266, 2021

[Liu et al., 2021c] Jiacheng Liu, Feilong Tang, Jielong Huang. Truth inference with bipartite attention graph neural network from a comprehensive view. In: ICME, 1-6, 2021

[Liu et al., 2022] Xiao Liu, Haoyun Hong, Xinghao Wang, Zeyi Chen, Evgeny Kharlamov, Yuxiao Dong, Jie Tang. SelfKG: Self-supervised entity alignment in knowledge graphs. In: WWW, 860-870, 2022

[Lyu et al., 2019] Shanshan Lyu, Wentao Ouyang, Yongqing Wang, Huawei Shen, Xueqi Cheng. Truth discovery by claim and source embedding. IEEE Transactions on Knowledge and Data Engineering, 33(3): 1264-1275, 2019

[Mao et al., 2021] Xin Mao, Wenting Wang, Yuanbin Wu, Man Lan. Boosting the speed of entity

alignment 10×: Dual attention matching network with normalized hard sample mining. In: WWW, 821-832, 2021

[Pasternack & Roth, 2010] Jeff Pasternack, Dan Roth. Knowing what to believe (when you already know something). In: COLING, 877-885, 2010

[Pasternack & Roth, 2013] Jeff Pasternack, Dan Roth. Latent credibility analysis. In: WWW, 1009-1020, 2013

[Ravi et al., 2021] Manoj Prabhakar Kannan Ravi, Kuldeep Singh, Isaiah Onando Mulang', Saeedeh Shekarpour, Johannes Hoffart, Jens Lehmann. CHOLAN: A modular approach for neural entity linking on Wikipedia and Wikidata. In: EACL, 504-514, 2021

[Sakor et al., 2019] Ahmad Sakor, Isaiah Onando Mulang', Kuldeep Singh, Saeedeh Shekarpour, Maria Esther Vidal, Jens Lehmann, Sören Auer. Old is gold: Linguistic driven approach for entity and relation linking of short text. In: NAACL, 2336-2346, 2019

[Shen et al., 2014] Wei Shen, Jianyong Wang, Jiawei Han. Entity Linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering, 27(2):443-460, 2014

[Sun et al., 2020] Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, Chengkai Li. A benchmarking study of embedding-based entity alignment for knowledge graphs. Proceedings of the VLDB Endowment, 13(11):2326-2340, 2020

[Sun et al., 2021] Zequn Sun, Muhao Chen, Wei Hu. Knowing the no-match: Entity alignment with dangling cases. In: ACL, 3582-3593, 2021

[Usbeck et al., 2014] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, Andreas Both. AGDISTIS - Graph-based disambiguation of named entities using linked data. In: ISWC, 457-471, 2014

[van Hulst et al., 2020] Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, Arjen P. de Vries. REL: An entity linker standing on the shoulders of giants. In: SIGIR, 2197-2200, 2020

[Wu et al., 2020] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In: EMNLP, 6397-6407, 2020

[Xu et al., 2021] Chengjin Xu, Fenglong Su, Jens Lehmann. Time-aware graph neural network for entity alignment between temporal knowledge graphs. In: EMNLP, 8999-9010, 2021

[Yan et al., 2021] Yuchen Yan, Lihui Liu, Yikun Ban, Baoyu Jing, Hanghang Tong. Dynamic

knowledge graph alignment. In: AAAI, 4564-4572, 2021

[Yang et al., 2019a] Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, Xu Sun. Aligning cross-lingual entities with multi-aspect information. In: EMNLP, 4430-4440, 2019

[Yang et al., 2019b] Xiyuan Yang, Xiaotao Gu, Sheng Lin, Siliang Tang, Yueting Zhuang, Fei Wu, Zhigang Chen, Guoping Hu, Xiang Ren. Learning dynamic context augmentation for global entity linking. In: EMNLP-IJCNLP, 271-281, 2019

[Yin et al., 2008] Xiaoxin Yin, Jiawei Han, Philip S. Yu. Truth discovery with multiple conflicting information providers on the Web. IEEE Transactions on Knowledge and Data Engineering, 20(6):796-808, 2008

[Zeng et al., 2021] Weixin Zeng, Xiang Zhao, Jiuyang Tang, Changjun Fan. Reinforced active entity alignment. In: CIKM, 2477-2486, 2021

[Zhang et al., 2022] Wenzheng Zhang, Wenyue Hua, Karl Stratos. EntQA: Entity linking as question answering. In: ICLR, 2022

[Zhao et al., 2020] Xiang Zhao, Weixin Zeng, Jiuyang Tang, Wei Wang, Fabian Suchanek. An experimental study of state-of-the-art entity alignment approaches. IEEE Transactions on Knowledge and Data Engineering, early access, 2020

[Zhi et al., 2018] Shi Zhi, Fan Yang, Zheyi Zhu, Qi Li, Zhaoran Wang, Jiawei Han. Dynamic truth discovery on numerical data. In: ICDM, 817-826, 2018

第七章 知识推理

张小旺[1]、李炜卓[2]、张文[3]、漆桂林[4]

1. 天津大学 智能与计算学部，天津市 300350

2. 南京邮电大学 现代邮政学院，江苏省 南京市 210003

3. 浙江大学 软件学院，浙江省 宁波市 315048

4. 东南大学 计算机科学与工程学院，江苏省 南京市 211189

一、任务定义、目标和研究的意义

知识图谱推理在每个知识图谱的发展演变过程中有重要的作用。随着知识图谱研究的深入，人们发现图谱在实际应用中仍存在两类主要的问题：

一类是知识图谱的不完备性问题，即知识图谱中有些关系会缺失或者有些属性缺少值，比如说一个人的职业信息缺失。这类问题可能是因为构建知识图谱的数据本身就是不完备的，也可能是信息抽取算法无法识别到一些关系或者抽取到属性值。

另一类则是知识图谱中存在噪声问题，即错误的事实声明，比如人物知识图谱中可能包含错误的人物关系。这类问题可能是因为构建知识图谱的数据存在错误，也可能是因为知识图谱构建时采用了基于统计的方法，而统计方法很难保证学习的知识是绝对正确的。

这两类问题对于智能问答等应用有较大影响，对于问答来说，前者会导致提出的问题没有答案, 而后者会导致系统给出的答案是错误的。知识图谱之所以被认为是实现人工智能的一个重要研究方向，是因为知识图谱上的推理使之能够支撑人工智能的很多应用，而这也是知识图谱区别于传统关系数据模型的关键所在。

知识图谱推理指的是从给定知识图谱推导出新知识或者检测知识图谱的逻辑冲突。它的核心技术手段主要可分为两大类，即：演绎系列，如：基于描述逻辑语言、逻辑规则的符号推理；归纳系列，如：基于嵌入表示学习、规则学习的统计推理。基于符号的推理技术被广泛用于生物医学中术语定义和概念分类、电商数据的不一致检测和查询重写等应用，有助于消除知识图谱中的噪声。基于统计的推理技术则对知识图谱进行补全，有效地缓解知识图谱中存在的不完备问题。近期研究表明，两类技术的相互融合可以有效地提升知识图谱推理方法的鲁棒性、可迁移性、可解释性、可应用性等，进一步支持智能问答等图谱应用。

二、技术发展脉络和进展

1. 基础知识

描述逻辑：描述逻辑（Description logics, 简称 DLs）是一类被广泛研究和应用于知识表示和推理的逻辑。DLs 是标准 Web 本体语言 OWL 和 OWL 2 的基础。在 DLs 中，域中的元素被编译成概念(对应于一阶语言中的一元谓词)，它们的属性通过角色的方式被结构化(对应于一阶语言中的二元谓词)。复杂的概念和角色表达式由原子概念名和原子角色名组成。这些名称由合适的构造符连接起来。可用构造符的集合取决于特定描述逻辑的语义。描述逻辑包含的构造符越丰富，描述逻辑可以捕获的语义就越复杂。描述逻辑知识库$\mathcal{K}$由 TBox ($\mathcal{T}$)和 ABox ($\mathcal{A}$)组成。TBox 由一系列公理组成，描述概念和角色间的包含关系。TBox 的语义受构造符的影响。在 ABox 中，可以表达对象的概念和对象间的角色关系[1]。

逻辑规则：逻辑规则的形式是 $H \leftarrow B_1 \wedge B_2 \wedge ... \wedge B_n$ 其中，H 代表规则的头部（Rule Head），$B_1 \wedge B_2 \wedge ... \wedge B_n$ 是规则的体（Rule Body），它是带有原子（atom）的合取范式。一种典型的逻辑规则是路径规则，其中规则的体是从规则的头部的头变量到尾部变量的路径；比如 $r(X, Z) \leftarrow r_1(X, Y) \wedge r_2(Y, Z)$是以从变量 X 到 Z 的路径为主体的路径规则，这里规则体中原子数量也称为规则的长度。在知识推理中，我们熟知的三元组(h, r, t)均可以转换为规则原子实例化后的形式即 r(h, t)用于激活规则来进行推理。

2. 研究内容

1）基于本体物化的推理

本体物化算法是一种基于前向链的本体推理算法。物化算法根据本体对原始数据集进行演绎推理，将隐含的本体信息表达为显式推理得到的新知识，扩充原始输入数据集。本体物化的目标是使用高表达性的本体语言尽可能地丰富原有数据集，从而在后续基于本体的查询问答技术中得到更完整的答案。不同的本体语言，如 OWL，相较于 RDFS 等语言拥有更强的知识表达能力，更适合于复杂域的描述。因为知识表达语言的表达能力不同，导致基于不同语言的物化算法的物化效率不同。

一般地，随着语言表达能力的增强，物化的时间复杂度会显著增加，因此如何在完备性与高效率之间达到平衡是目前面临的较大挑战；除此之外，当本体中表达循环依赖关系时，

[1] 对于一个知识图谱的而言，可以将其数据分为术语层 TBox 和实例层 ABox，其中术语层(Schema)包含了当前知识图谱中的概念层次以及关系约束等抽象知识，用于指导实例层数据的构建，实例层包含了描述了用三元组表示的实体和实体之间关系，其中实体则对应概念的实例。

基于物化算法计算得到的 RDF 数据集可能是无穷的。查询改写[Calvanese et al. 2017]通过使用虚拟 RDF 图谱技术来避免无穷物化，但是查询重写方法需要线上的查询改写时间，并且查询改写中的映射输入需要人工干预；gOWL[Meng et al. 2018]提出了一种部分物化方法避免无穷物化。但 gOWL 只解决不包含循环结构的非布尔合取查询，不支持实际应用中广泛存在的布尔查询和循环查询,并且具有较高的时间和空间复杂度。因此如何提出高效的物化算法来处理无穷物化问题也是一个较大的挑战。

2）基于神经网络和本体表示学习的知识推理

基于神经网络和本体表示学习的知识图谱可以分为作用于实例层的神经网络推理以及作用于本体层的本体表示学习推理，其中基于神经网络的推理又可分为基于图神经网络的知识图谱推理以及基于知识图谱嵌入和预训练的推理。下面简单介绍一下这三类研究技术的问题定义：

- 基于知识图谱嵌入与预训练的推理：给定知识图谱的实例层$ABox = \{E, R, T\}$，其中E, R和T分别表示实体、关系和三元组的集合，知识图谱嵌入与预训练的目标是通过学习实体和关系嵌入矩阵$\boldsymbol{E}$和$\boldsymbol{R}$，通过设定的打分函数f对三元组进行真值打分，同时使得$\boldsymbol{E}$和$\boldsymbol{R}$在向量空间中捕捉潜在的不同实体和关系之间的相似性和逻辑蕴含性。知识图谱嵌入方法多用于知识图谱补全，而预训练方法则多用于知识图谱相关的下游任务
- 基于图神经网络的知识图谱推理：给定一个三元组(h, r, t)以及实体的邻居信息N_h/N_t，基于图神经网络的知识图谱推理方法通过聚合函数g基于实体的邻居信息得到实体的表示$\boldsymbol{h}/\boldsymbol{t}$，然后基于聚合得到的实体表示和关系表示，通过打分函数f对当前的三元组进行真值打分，从而完成对未知三元组的真值判断。由于考虑了实体周围的邻居信息，基于图神经网络的推理方法比基于嵌入和预训练的方法往往具有更好的可解释性潜力。
- 基于本体表示学习的知识图谱推理：给定知识图谱的实例层和术语层，可以表示为一个本体$O = \{ABox, TBox\}$，本体表示学习对本体O中包含的概念、关系、属性等进行表示学习，使得表示学习结果不仅需要满足 ABox 中的三元组真值判断，还满足 TBox 中定义的该概念层次以及公理约束。因此本体表示学习方法的逻辑表达能力要求比基于嵌入和预训练的推理方法以及基于图神经网络的推理方法更高。

3）基于符号逻辑与嵌入表示的混合推理

符号逻辑推理与嵌入表示推理一直是知识图谱推理的主流技术。前者将问题形式化为语义框架，通过一些预定义的规则推出图谱中潜在的知识，后者则设计合适的统计模型来适配

已有的数据，通过训练得到的参数模型来预测出知识图谱中实体之间的潜在关系。两者在实际的知识图谱应用中各具优缺点。符号逻辑推理依赖于规则和本体这类难获取的知识，嵌入表示推理这种数据驱动的方法无法得到精确的预测同时无法提供良好的解释。为此，基于知识图谱的混合推理应运而生，致力于让两种方法优势互补[Li et al., 2020] [Chen et al., 2020] [Zhang et al., 2022] 。本节将基于知识图谱混合推理的研究技术大致分为以下三类:

- 嵌入表示模型中融入逻辑规则的混合推理：这类方法是将符号知识（如：规则、路径等）约束融入到嵌入表示模型中，以此来增强嵌入表示模型在知识图谱中推理的效果。
- 逻辑推理中融入嵌入表示的混合推理：这类方法是将嵌入模型训练得到的实体与关系向量表示应用到符号推理过程中，将符号推理的过程进行“软化”，以此来缓解知识图谱自身不完备所导致的推理链中断问题。
- 其他知识图谱混合推理：这类方法主要汇总了多跳推理(Multi-hop reasoning)、模式归纳(Schematic induction)、流推理(Streaming reasoning)等相关混合推理技术，它们更多是将广义符号推理与统计推理的方法进行融合，为下游任务提高性能的同时，也为推理结果提供有效的解释。

不同的符号逻辑推理方法与嵌入表示推理方法在近年来的知识图谱推理发展中呈逐渐融合的趋势。而基于知识图谱的符号推理主要采用的是开放世界假设，而嵌入表示推理的方法则主要是依赖（半）封闭世界假设。如何在保证方法融合过程中精度提升的同时，使其具备更好的鲁棒性、可迁移性、可解释性则是当下知识图谱混合推理的重要挑战。

三、技术方法和研究现状

1. 基于本体物化的推理

1）基于部分物化算法的查询问答系统-SUMA

部分物化算法是一种为了解决物化算法中结果可能无穷尽的问题，通过计算多步通用模型得到有穷尽结果集的本体推理算法。但是，部分物化算法限制查询必须包含自由变量并且不能包含循环结构，因此不能支持包含循环结构的根查询和不包含自由变量的布尔查询。然而，根查询和布尔查询对于查询图中的循环结构和可满足性十分重要。

Qin 等人[Qin et al. 2021]提出了一种基于查询分析算法的扩展部分物化算法。该扩展部分物化算法能在$\mathrm{DL}-\mathrm{Lite}_{horn}^{\mathrm{N}}$中可靠完备地支持根合取查询和包含循环结构或者叉形结构的布尔合取查询。另外，该工作还通过应用改写和近似技术可靠地扩展部分物化算法支持高

表达性的本体语言 OWL 2 DL，并通过等价角色和逆角色改写算法，进一步优化物化效率和物化空间消耗。该物化算法和角色改写算法最终集成于查询问答系统 SUMA 中，如图 1 所示。该系统效果与 Pellet 齐平，并在部分测试数据上优于 PAGOdA。特别地，SUMA 是高度可扩展的。其物化流程如图 2 所示。

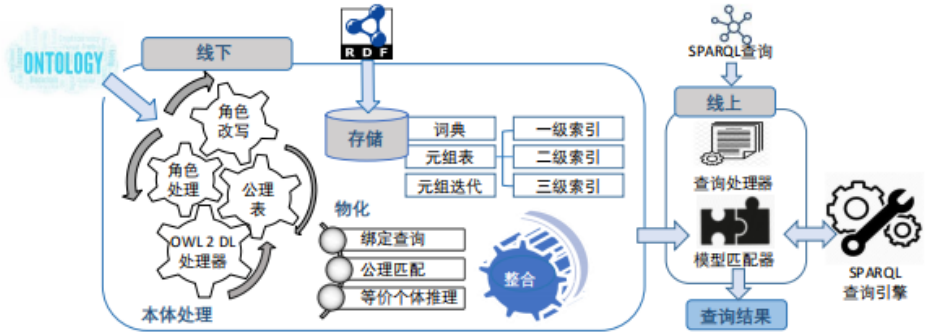


图 1 SUMA 系统架构图

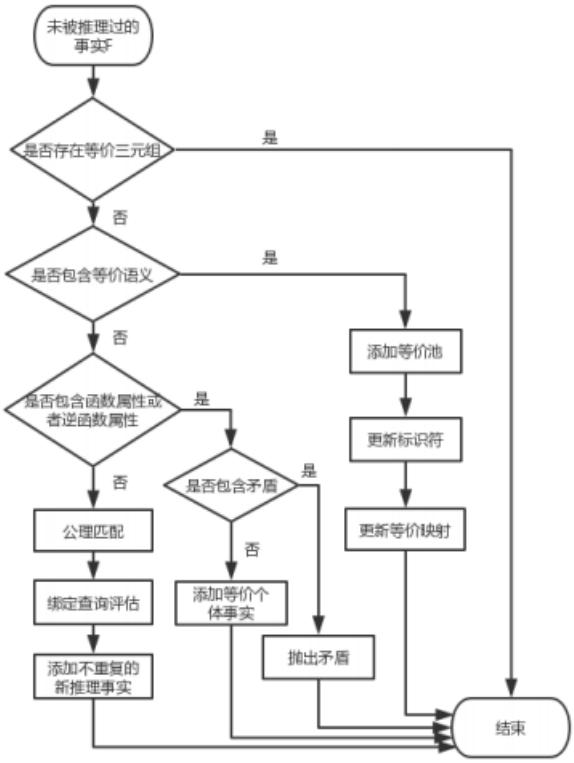


图 2 SUMA 物化流程图

2）基于查询改写的本体推理

查询重写[Xiao et al. 2019, Han et al. 2022]是一种解决物化算法结果无穷化问题的常用方法，一般基于虚拟知识图谱（VKG）实现，如图 3 所示。虚拟知识图谱的核心思想为：底层一般采用现有的商用关系型数据库对数据进行存储，顶层则为本体描述。二者通过映射层来

连接。其“虚拟”表现在，并没有实际存储如 RDF 等形式的图数据，而是通过本体和数据源之间的映射来描述图数据。因此在本体推理中，可以通过 VKG 来避免物化，取而代之的推理运算步骤如下：在本体层应用有本体推理能力的 OWL 2 QL 作为其本体语言，并通过本体层对 SPARQL 查询进行重写，继而利用映射层将查询转换为数据层支持的查询语言。得到结果后，通过映射层转换为 SPARQL 查询结果返回。即采用查询改写算法并不根据本体显式计算出所有隐含知识，而是通过本体和映射对查询进行重写，使得查询显式包含本体中的隐含信息。但是，查询重写方法需要线上的查询改写时间，并且查询改写中的映射输入需要人工干预。特别地，重写的查询可能是原始查询的指数级别大小。

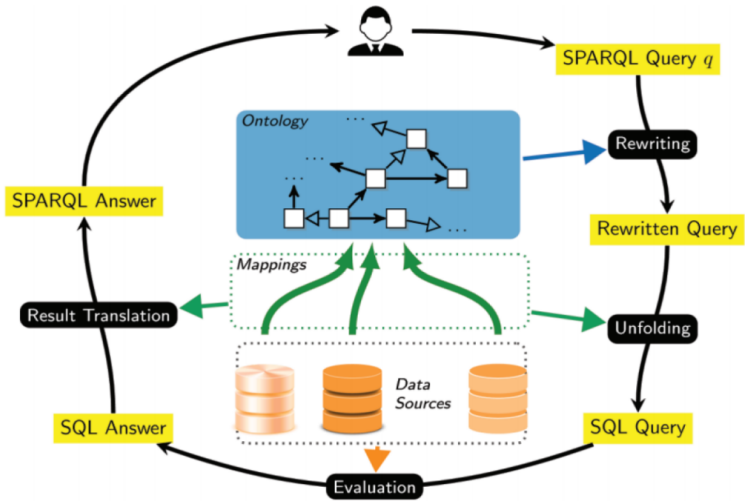


图 3 虚拟知识图谱系统实例

3）基于本体推理的本体中介查询回答技术

本体中介查询回答(OMQA-Ontology-Mediated Query Answering)是人工智能、数据库和语义 Web 领域的数据管理发展趋势，旨在回答知识库上的数据库查询。因为它是自动化推理和数据库查询评估的复杂组合，因此主要带来了性能上的挑战。本体介导的查询回答[Bienvenu et al. 2016](OMQA)相当于回答知识库(KB)上的数据库查询，即由建模应用程序领域背景知识的本体所描述的数据。它超越了数据库查询回答，后者只从数据中生成答案，通过在本体的帮助下对数据进行推理来识别额外的答案。OMQA 最近已经成为人工智能、数据库和语义 Web 社区的一个热门话题，特别是在连接查询（用描述逻辑表示的 KB 上的核心项目-连接数据库查询），在此基础上构建了 W3C 的 Web 本体语言（OWL 2），以及密切相关的数据日志与存在规则。

OMAQ 三种常见技术：通过重新制定或重写[Chortaras et al. 2011,Thomazo et al. 2013, Bursztyn et al. 2016]在查询中编译本体知识，通过饱和化或物化[Leone et al. 2019]，或通过组合或混合方法（如：[Kontchakov et al. 2010]）将 OMQA 简化为标准的关系数据库查询评估。

基于查询重构的 OMQA 技术是迄今为止研究最多和使用最多的一种技术。它被引入[Calvanese et al. 2007]并包括使用 KB 的本体将每个传入的联合查询(CQ) q 重新构造为联合查询(UCQ) q'，以便评估对(SQLized) q'的标准数据库在 KB 存储的数据上在关系数据库中产生对 q 的正确答案。至关重要的是，这样一个重构的查询 q' 在它的并集中列举了 q 相对于所有（最坏情况下成倍增加）专业本体论的知识。在实践中，重构查询的可能性很大，在这种情况下，关系数据库管理系统(RDBMS)，即使是现代的，也无法有效地回答它们（UCQ 中的所有 CQ 都被评估）。

线性时序逻辑中本体中介查询的一阶可重写性[Xiao et al.2021]：研究基于本体的数据访问时间数据，考虑在离散时间上解释的线性时间逻辑 LTL(Liner Temporal Logic)中给出的时间本体$(\mathbb{Z},<)$。查询以 LTL 或 MFO(<)，具有内置线性顺序的一元一阶逻辑给出。关注由时间本体和查询组成的本体中介查询（OMQ）的一阶可重写性。考虑本体中使用的时间运算符并区分完整 LTL 及其核心、Krom 和 Horn 片段中给出的本体，证明可重写为 FO(<)，一阶具有内置线性顺序逻辑，或 FO(<)，使用标准算数谓词 x≡0(mod n)，扩展了 FO(<)，对于任何固定的 n>1，或 FO(RPR)，它扩展了 FO(<)与关系原语递归。在复杂性方面，FO(<，≡)-和 FO(RPR)-可重写性保证 OMQ 在AC^0和NC^1的查询应答中统一。

4）基于本体物化推理的高效可扩展引擎-PAGOdA

PAGOdA[Zhou et al. 2015]是目前采用物化方式的推理引擎中的佼佼者，由牛津大学研发，使用 JAVA 语言开发。系统架构图如图 4 所示。

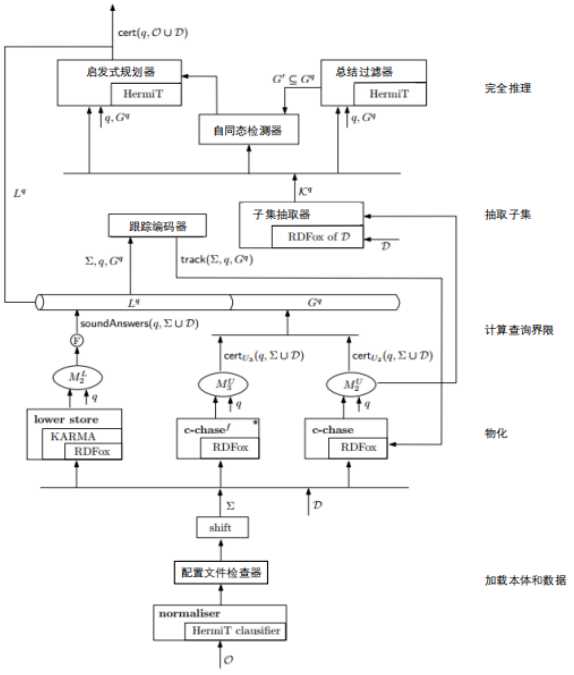


图 4 PAGOdA 系统架构图

为解决完全 OWL 2 DL [Horrocks et al. 2006]推理机可扩展性差的缺点，PAGOdA 采用了一种组合方式进行高效可扩展的本体推理。该引擎整合了 Datalog 推理机 RDFox [Nenov et al. 2015]和完全成熟的 OWL 2 推理机 HermiT [Motik et al. 2009]，使之成为“黑盒子”为 PAGOdA 提供功能支持实现细节对用户透明。PAGOdA 提升本体推理效率的主要核心思想是将大量的计算负载委托给高可扩展的 Datalog 引擎并且只有特定关键情况才会求助于运行代价较高的 OWL 2 推理机。

PAGOdA 完成整个查询问答的过程可以分为 5 个步骤: 加载本体和数据、物化处理、计算查询边界、提取子集以及完全推理。图中的每个框表示 PAGOdA 的一个组件，任何外部系统都可以调用该组件。原则上，PAGOdA 可以使用任何基于物化的 Datalog 推理器来完成 CQ 评估和事实的增量添加，以及使用任何完全成熟的 OWL 2 DL 推理器来支持事实蕴涵，也就是说，RDFox 和 HermiT 并非是唯一的选择。

给定知识库$\mathcal{K}$和查询q的前提下，PAGOdA 按照以下算法进行查询回答：

第一步通过 Datalog 推理机计算给定查询 q 的答案的一个下界，这个下界是可靠的但可能是不完备的，和一个上界完备的但可能不可靠。

第二步如果上界和下界都返回不满足，则推理机返回不满足。如果上界和下界推理机都返回查询可满足，并且上界和下界推理机返回答案一致，则推理机返回最终结果。其它情况，进入下一步处理。

第三步针对在上界中存在的而不在下界中存在的答案，通过 Datalog 推理机抽取相关知识库数据。

第四步对第三步中的每一个答案，通过 OWL 推理机来检测是否可满足。为了减少 OWL 推理机的计算负载，采用摘要技术(Summarization)有效减少备选答案数。最终返回所有可满足答案。

5）本体物化推理的其他应用

Ahmetaj 等人[Ahmetaj et al. 2021]采用来自语义 Web 社区和数据交换社区的方法和技术，开发了一个灵活的、开源的框架，用于关系数据库上的查询回答，使用来自数据交换社区的物化过程来实现一个通用的解决方案，可用于回答企业医疗数据库上的查询。在此过程中，确定了一类新的行为良好的无环 EL 本体，并扩展了角色层次结构、适当限制的功能角色和域/范围限制，它们涵盖了这篇工作的用例。证明了这样的本体物化过程在多项式时间内终止。

Li 等人[Li et al. 2022]提出了一种协同提升框架(CBF)，以迭代方式将数据驱动的深度学习模块和知识引导的本体推理模块结合起来。深度学习模块采用 DSSN 架构，将原始图像和推断通道的整合作为 DSSN 的输入。此外，本体推理模块由分类内推理和分类外推理组成。

万物互联和语义网可以通过为普遍系统提供更多智能来加入。为此，即使是资源非常有限的嵌入式设备也应该具有启用推理的能力。Ruta 等人[Ruta et al. 2022]提出了 Tiny-ME（Tiny Matchmaking Engine），它是一种用于 Web 本体语言 (OWL) 的匹配和推理引擎，采用紧凑且可移植的 C 内核设计和实现。主要特点是高资源效率和多平台支持，涵盖容器化微服务、桌面、移动设备和嵌入式板。 OWLlink 接口已扩展为在 Web、云和边缘计算中启用非标准推理服务以进行匹配。提出了原型评估，包括对 Pixhawk 无人机(UAV)自动驾驶仪和性能亮点的案例研究。

2. 基于神经网络和本体表示学习的知识推理

在本小节中，我们将从不同方法欲解决的推理问题角度对三类推理方法在近些年的进展进行展开介绍。

基于知识图谱嵌入与预训练的推理

与词向量的思想类似，知识图谱嵌入推理将实体和关系映射到向量空间，称为实体或关系的嵌入表示，嵌入表示支持通过计算获得实体或关系的语义信息，例如，实体的相似度、关系的性质以及实体和实体之间的关系等。作为最早的知识图谱嵌入表示学习方法之一TransE[Bordes et al., 2013]将头实体到尾实体的映射看作向量的平移翻译，模型简单有效，但对复杂的关系表达能力不足，例如，关系的自反性、可逆性、传递性以及组合性等，随后众多可编码关系多样语义的模型被提出[Wang et al., 2014][Théo et al., 2016]，改善了嵌入模型对关系的多对多关系、关系的对称型、关系的非对称性、关系的传递性等语义的表达能力。近年来，学者提出了一些表达能力更强的嵌入表示模型，探索了更加多样和丰富的关系语义。这里介绍两个代表方法 RotatE [Sun et al., 2019]和 BoxE[Ralph et al., 2020]。RotatE 将知识图谱嵌入到复数空间，并将实体表示为复数向量，将关系表示为对实体复数向量的空间旋转操作，将头实体向量旋转至尾实体的表示。BoxE 用两个向量表示实体，一个基本向量和一个平移向量，将关系表示为一个矩形，并假设如果三元组成立，那么基于头尾实体基本和平移向量的组合表示应落在关系矩形内。下表中展示了不同嵌入模型对不同推理模式的表达能力。

表 1. 不同嵌入模型对不同推理模式的表达能力

Inference pattern	BoxE	TransE	RotatE	DistMult	ComplEx
Symmetry: $r_1(x,y) \Rightarrow r_1(y,x)$	✓/✓	✗/✗	✓/✓	✓/✓	✓/✓
Anti-symmetry: $r_1(x,y) \Rightarrow \neg r_1(y,x)$	✓/✓	✓/✓	✓/✓	✗/✗	✓/✓
Inversion: $r_1(x,y) \Leftrightarrow r_2(y,x)$	✓/✓	✓/✗	✓/✓	✗/✗	✓/✓
Composition: $r_1(x,y) \wedge r_2(y,z) \Rightarrow r_3(x,z)$	✗/✗	✓/✗	✓/✗	✗/✗	✗/✗
Hierarchy: $r_1(x,y) \Rightarrow r_2(x,y)$	✓/✓	✗/✗	✗/✗	✓/✗	✓/✗
Intersection: $r_1(x,y) \wedge r_2(x,y) \Rightarrow r_3(x,y)$	✓/✓	✓/✗	✓/✗	✗/✗	✗/✗
Mutual exclusion: $r_1(x,y) \wedge r_2(x,y) \Rightarrow \perp$	✓/✓	✓/✓	✓/✓	✓/✗	✓/✗

从表中可以看出，目前常被考虑的推理模型有关系的对称性、关系的非对称性、逆关系对、关系的组合性、关系的上下位、关系的交集推理以及关系的互斥性。同时可以看出，目前的尚无模型可以覆盖所有的推理模式，因此尽管嵌入模型具有较好的推理能力，但就表达能力而言，依然弱于符号表示，表达能力有待进一步提升。

近年来，大规模预训练语言模型在自然语言处理领域取得卓越的进步，在文本分类、情感分析、关系抽取等任务均实现了突破性提升。其核心理念是在大规模数据上进行自监督的预训练，并在少量下游任务数据上进行微调，即可实现良好的下游任务效果。类似的，大规模知识图谱场常常被用于各种不同的下游任务。例如，金融知识图谱会广泛被用于智能客服、金融产品推荐、金融趋势分析等任务，因此，基于“预训练+服务”理念的知识图谱预训练模型被提出[Zhang et al., 2021]，其核心思想是在大规模知识图谱上进行嵌入学习，使得预训练模型具有知识图谱补全的能力，并在下游任务中为每个实体提供多个可以反映其关系和属性的服务向量，并设计了统一的在多种在下游任务中融合服务向量的方法，使得知识图谱的信息可以轻松融入并增强下游任务，实现了知识图谱嵌入推理方法的实用化，基于其补全和推理的能力为下游知识图谱应用任务提供更好的知识服务。

2）基于图神经网络的知识图谱推理

受到图神经网络在同构网络研究上的启发，应用于知识图谱推理的图神经网络主要基于知识图谱的图结构进行学习。与之不同的是，知识图谱推理还需要考虑节点和边的语义类型信息以支持更复杂的逻辑推理。对比可以隐含地捕获图结构的基于图谱嵌入的推理，基于图神经网络的推理会显式地对图结构以及节点特征进行编码，因而可以有效地利用实体的邻居实体信息和连接关系进行推理。此类典型的算法有 R-GCN[Schlichtkrull et al., 2018]以及 CompGCN[Vashishth et al., 2020]等。其中 R-GCN 为每个关系学习了一个聚合函数，用不同关系连接的邻居通过不同的聚合函数得到实体表示并融合为一个实体表示，R-GCN 采用编码解码器结构，将聚合邻居信息得到实体表示的过程看作是实体编码过程，解码器通常根据任务进行设定。例如，对于实体分类，解码器是一个以实体向量为输入的多分类器，对于链接预测，解码器是一个知识图谱嵌入模型。为了解决模型 R-GCN 过参的问题，CompGCN

设计了实体和关系的组合表示算子，在邻居聚合过程中允许各种实体和关系的交互，从而通过同样的聚合函数完成不同关系类型的邻居实体聚合，具有参数少但灵活的特点。

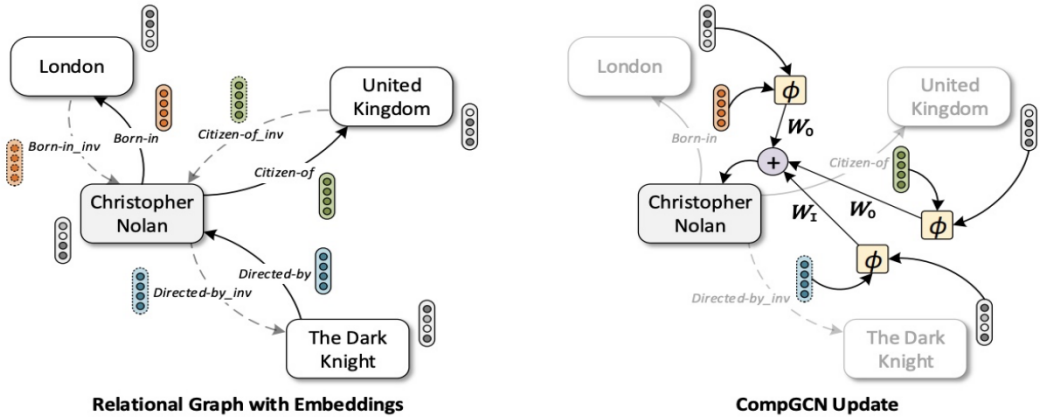


图 5 R-GCN 和 CompGCN 模型聚合过程示例

由于图神经网络可以通过实体的邻居得到实体的表示，具有一定的对零样本实体进行推理的能力，因此近年来被广泛应用于零样本实体的知识外推任务中。传统的链接预测假设测试时的实体在训练过程中都是见过的，但这个设定过于理想化，在实际应用中，随着时间的推移，知识图谱在不断变化和扩增，总是有新的实体被加入知识图谱中，也会有进行不同知识图谱间的知识迁移的需求，因此 inductive 关系推理任务近年来被广泛研究，致力于研究如何对训练过程中从未见过的实体进行关系预测。其中的代表性工作有 GraIL[Komal et al., 2020]和 CoMPILE[Sijie et al., 2021]。其核心思路分为三步，首先是子图抽取，根据要预测的实体对，抽取其在知识图谱中子图。其次是对实体进行特征表示，通过图上的特征，例如，距离头实体和尾实体的最短距离，为每个子图中的每个实体节点赋予一个特征向量。最后，通过图神经网络，根据子图节点的特征表示聚合得到实体和关系的表示并进行打分。CoMPILE 是 GraIL 模型的改进工作，在子图抽取过程中进一步考虑了被预测关系的方向性，并在子图聚合过程中也融合了目标关系的嵌入表示以及方向信息。

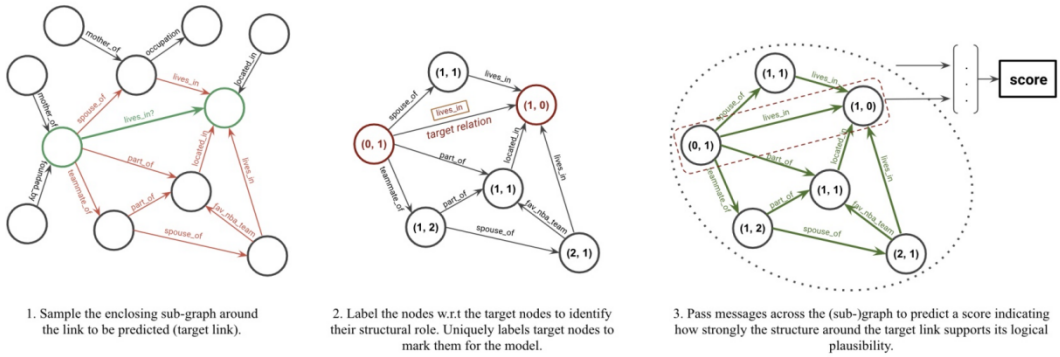


图 6 GraIL 和 CoMPILE 处理 inductive 关系推理的思路

基于图神经网络的知识图谱推理已广泛应用于长尾关系抽取、实体对齐、零样本图像识别、对话生成以及推荐系统等应用中。随着知识图谱规模不断扩大，大图数据的处理，也就是基于大数据的计算引擎的图计算也是需要深入研究和考虑的技术问题。

3）基于本体表示学习的知识图谱推理

三元组和图结构信息能较好地支持简单直接的知识图谱推理，而复杂的推理往往依赖于知识框架，又称为本体，其中的实体以概念与属性为主。本体嵌入表示主要侧重于将概念层次体系、概念之间的逻辑组合关系、属性的层次体系、概念和属性之间的逻辑组合以及属性自身的性质（如：传递性、对称性、自反性）等这类抽象的知识编码到稠密、连续的语义空间中，即如何将本体语义和逻辑表达进行向量化表示。典型的本体嵌入模型是 EL Embedding[Kulmanov et al., 2019]，该模型将轻量级的 EL 本体利用高维的球形空间来表示，用球心之间的位置来编码概念之间的关系。现有的本体表示学习研究较多的是概念之间的层次关系以及连接关系，对于更复杂的逻辑表达包括组合语义（或/且/非）、存在量词和全称量词涉及较少。

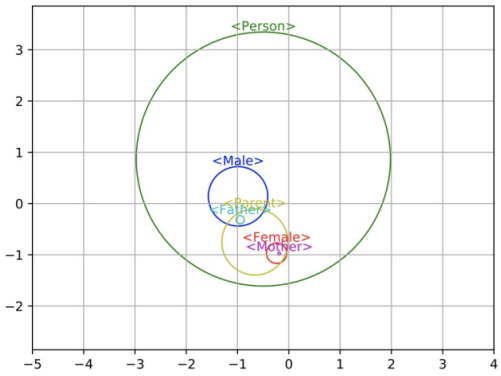


图 7 EL Embedding 可视化的概念表示

从上图中我们可以看出，本体表示学习结果可以捕捉到概念层次之间的关系，例如<Person>作为一级概念，又可细分出<Male>,<Female>,<Parent>,<Father>和<Mother>等的子概念，并且<Mother>又是<Female>的子概念，<Father>是<Male>的子概念，<Parent>与<Male>和<Female>均有交集。

近年来，受益于近些年来复杂知识库问句查询的研究[Ren et al., 2020a]，本体表示学习[Chen et al., 2021]在向量空间的逻辑表达能力得到了进一步的探索。

3. 基于符号逻辑与嵌入表示的混合推理

1）嵌入表示中融入逻辑规则的混合推理

尽管嵌入表示方法在一些下游的推理任务中取得了巨大的成功，但它们在复杂语义表示与稀疏数据的表现上仍存在不足。在嵌入表示的模型中融入逻辑规则的混合推理，不仅可以

增强模型在知识推理上表现，还可以有效地缓解知识图谱存在的数据稀疏等问题[Niu et al., 2020]。

根据规则融合的时间可以分成三个阶段。1) 模型训练之前：即在学习嵌入模型之前进行规则推理进行一定的知识补全。这样的逻辑规则推理通常会影响嵌入表示学习中训练样本，如：正确三元组与错误的三元组。2) 在模型训练之中：即在嵌入学习时注入逻辑规则。该方式会影响到嵌入模型自身损失函数定义，形成相应的逻辑约束，如：对具有逻辑性质的关系加上规则或关系语义性质约束。3）在模型训练之后：即在嵌入学习完成模型之后，将规则作为错误事实验证的过滤器，来进一步优化推理结果。下面我们主要对模型训练之中融入逻辑规则的方式进行介绍。

逻辑规则通常表示为 horn 子句，例如，∀x，y (x，首都，y) → (x，位于，y)声明由关系“首都”链接的任何两个实体也应该满足关系“位于”。Guo 等人[Guo et al., 2016]提出了一个联合模型 KALE，它将事实知识和逻辑规则嵌入到统一的框架，其中逻辑规则被解释为通过将基础原子与逻辑连接词(如：“∧”和“→”)相结合而构建的复杂公式，并采用 t-范数的模糊逻辑来度量每一个基础原子和规则实例化原子的真值分数。在此基础上，他们进一步改进了该模型，新提出的模型 RUGE [Guo et al., 2018]可以更好地对知识图谱中已有的三元组、带权重的逻辑规则以及借助规则推理出的三元组进行联合建模，并形成迭代。Zhang 等人[Zhang et al., 2019]针对知识图谱中存在数据稀疏现象提出了一种新型的迭代嵌入学习框架 IterE。它采用迭代方式使得模型通过实例化规则得到较多出未标记三元组，从而可以有效地缓解训练数据中实体与关系关联不足的问题。

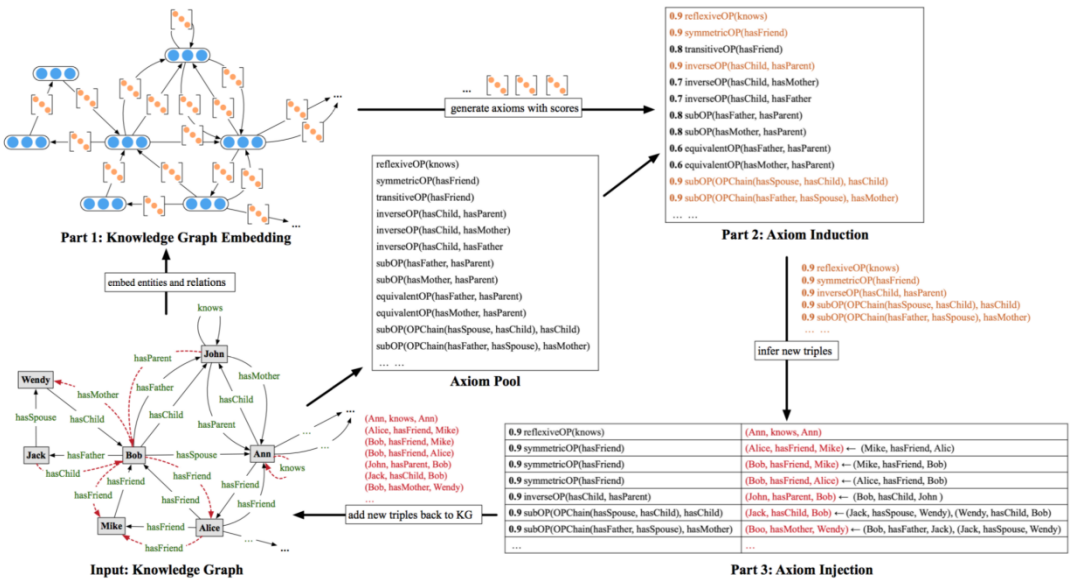


图 8 IterE 中详细的推理过程

除了规则的满足约束，Wang 等人[Wang et al., 2018]提出的 TARE 模型则更强调规则中

的传递性和不对称性，这使得模型中规则之间关系的顺序变得十分重要，作者通过分量式不等式对逻辑规则中关系类型的顺序进行了有效建模。

2）逻辑推理中融入嵌入表示的混合推理

尽管符号逻辑推理经过了多年的研究，具有较为广泛应用，它自身仍然存在一些局限。一方面，逻辑的构造通常依赖于领域专家，会耗费大量的时间和人力，加之真实的数据往往存在固有的噪声与不确定性，因此，大多数逻辑应用程序在实际推理的覆盖度上会存在一定的限制。另一方面，逻辑表达过于复杂化则会导致较高的推理复杂度，因此其推理的规模上也会面临可扩展性的问题。相对而言，基于嵌入表示的推理的特点是将实体与关系映射在连续、稠密的向量空间，且无需预定义的归纳推理逻辑，因此，将嵌入表示模型融入到逻辑推理中可以较好地弥补逻辑推理中不确定性、泛化性等不足的问题。目前将嵌入表示融入逻辑推理的方法，主要有两种策略。1) 在逻辑推理之前：即在逻辑推理任务之前将嵌入模型相应的预测结果加入。2) 在逻辑推理之中：即在逻辑推理过程加入嵌入表示模型，来学习逻辑中的推理模式并优化推理的效率，以此更好地完成推理任务。查询问答与定理证明是逻辑中常见的两项推理任务。

在查询问答方面，最初研究者仅考虑最简单的查询问句类型。Guu 等人[Guu et al., 2015]基于 TransE 和随机游走的策略实现了对软边遍历算子的向量表示，并递归地将其应用于预测合成路径查询。受益于嵌入表示思想，学者们对更多复杂的查询问句形式（如：合取逻辑查询，含存在量词的一阶问句）进行了深入的研究。Hamilton 等人[Hamilton et al., 2018]基于知识图谱嵌入表示的思想对逻辑问句进行了建模，提出了 GQE 模型。GQE 将实体作为向量嵌入，将关系看作实体嵌入的投影算子，同时将合取逻辑查询中的“∧”作为交集算子。模型将每个查询编码成一个向量，并根据查询和候选实体嵌入之间的相似性给出答案。Ren 等人提出的模型 Query2box[Ren et al., 2020b]可以通过将查询中的析取(∨)转换成析取范式(DNF)来进一步支撑知识图谱的查询推理，并且它为每种量化类型定义了向量空间运算符以达到更好的查询效果。Arakelyan 等人提出的模型 CQD [Arakelyan et al., 2021]采用嵌入表示模型 ComplEx [Trouillon et al., 2016]来定义投影算子，同时作者借助 t-范数的模糊逻辑来量化查询问句中的量词，以此来支持常用的一阶逻辑运算，包括合取(∧)、析取(∨)和求反(¬)。近期，Kotnis 等人提出的模型 BiQE[Kotnis et al., 2021]能够将连接查询翻译成序列，并通过双向的 Transformer Encoder 对其进行编码，这种双向注意力机制的嵌入表示可以有效地捕捉查询中所有元素之间的交互。刘等人[Liu et al., 2021]通过嵌入表示模型进一步支持了逻辑运算不等于(≠)的查询，在支持更多复杂的查询问题种类的同时可以保证较高的推理效率。此外，也有部分学者尝试用嵌入表示学习来提高逻辑推理的查询效率。例如，Wei 等人[Wei et

al., 2015]提出的 INS-ES 模型主要是采用数据驱动的推理算法 INS 与马尔科夫逻辑网进行符号推理，在此基础上通过 TransE 模型能够进一步缩小查询的候选集。

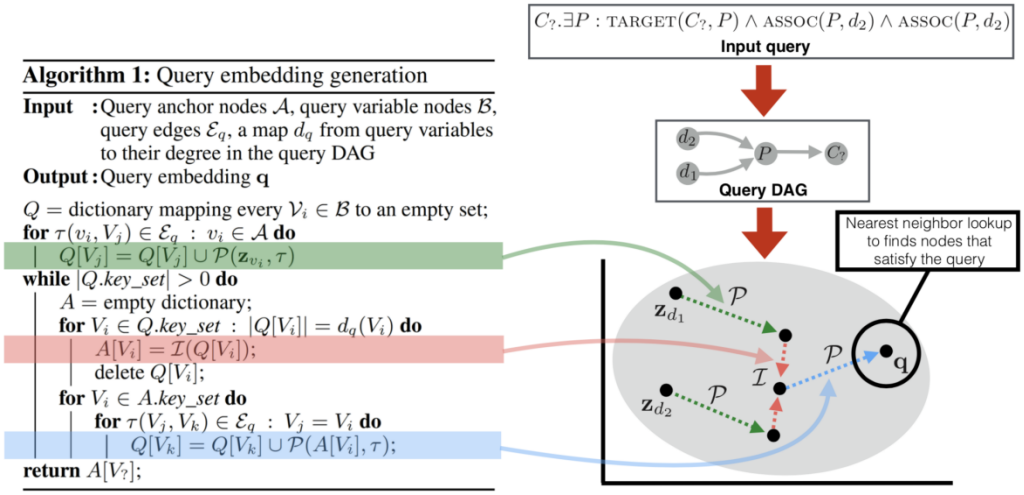


图 9 GQE 的算法框架概览

在定理证明方面，使用嵌入表示模型在一定程度克服了可微定理证明中符号证明器在推广到问句相似但符号不相同的查询限制。例如，Rockt¨aschel 与 Riedel 提出的模型 NTP[Rockt¨aschel and Riedel, 2017]可以使 Prolog 能够学习知识图谱中实体和关系的嵌入表示以及它们之间的相似性。它使变量绑定的符号遵循 Prolog 的推理语法，同时支持实体符号能够用与嵌入表示向量相似的实体进行替代。因此，NTP 可以在没有预定义的情况下学习特定于领域的规则，并进行无缝推理。为了提升 NTP 推理过程中枚举与评分所产生的效率问题。Minervini 等人则进一步提出了 GNTP 模型[Minervini et al., 2020a]，模型基于学习得到的事实的嵌入表示来选择用于证明子目标的顶部最近邻事实声明，并使用关系的嵌入向量表示来选择需要扩展的顶部规则。此外，他们还提出了另一种建模方法 CTP[Minervini et al., 2020b]。该模型是使用键-值记忆网络，将证明的目标、关系和常数的嵌入向量作为条件，在每一轮推理步骤中考虑动态地生成的极小规则集来提高推理效率。

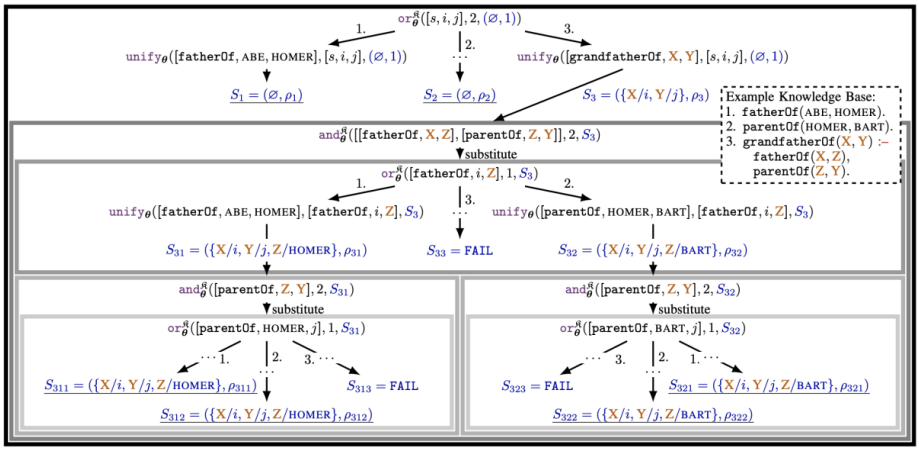


图 10 基于 NTP 计算图的示例性构造案例

3）其他知识图谱混合推理

本节主要对多跳推理、模式归纳、流推理等相关的混合推理技术进行汇总，它们更多是将广义符号推理与统计推理的方法进行融合，为下游任务提高性能的同时，也为推理结果提供有效的解释。

由于为问答定制的语义解析模型[Abujabal et al., 2017]和嵌入表示模型[Hao et al., 2017]均受限于数据标注和推理效率等局限，仍不足以解决多跳推理的任务场景。为此，不少学者尝试借助混合推理的来提高性能，并使这些结果可以解释。Zhang 等[Zhang et al., 2018]]提出了一种多跳问答的概率建模框架，模型可以处理不确定的主题实体，并实现问答的多跳推理。作者在知识图谱上引入了一种新的传播体系结构，使得逻辑推理可以在概率模型中执行。Zhou 等人设计了一个可解释的推理网络 IRN[Zhou et al., 2018]。它可以动态地决定在每一跳应该分析输入问题的哪一部分，并预测出对应于解析结果的关系。因此，IRN 在推理预测的中间实体和关系可以构建可追踪的推理路径为答案提供解释。Vakulenko 等人[Vakulenko et al., 2019]为复杂问答设计了一种基于无监督消息传递的方法。该方法通过利用一系列的稀疏矩阵乘法来模拟局部子图上的连接与消息的传递，通过解析输入问题，并将知识图谱中的术语与一组可能答案进行匹配来传播置信度得分，因此，它可以应用于 DBpedia 这种大型的知识图谱。此外，Saxena 等人提出了 EmbedKGQA 模型[Saxena et al., 2020]用于在稀疏知识图谱上回答多跳问答。模型的基本思想为：主题实体向量+关系路径向量≈主题实体向量+问题向量，即表示问题的语义在问答方法中，与关系路径的语义是大体一致的。基于该思想，作者引入嵌入表示模型 ComplEx [Trouillon et al., 2016]将知识图谱中的实体和关系投射到复数空间，并基于预训练模型 RoBERTa[Liu et al., 2021]建立前馈神经网络对问题进行编码。EmbedKGQA 通过将主题实体向量和问题向量相加，并在复数空间寻找与该向量最接近的实体向量，并将该实体为预测答案进行输出。

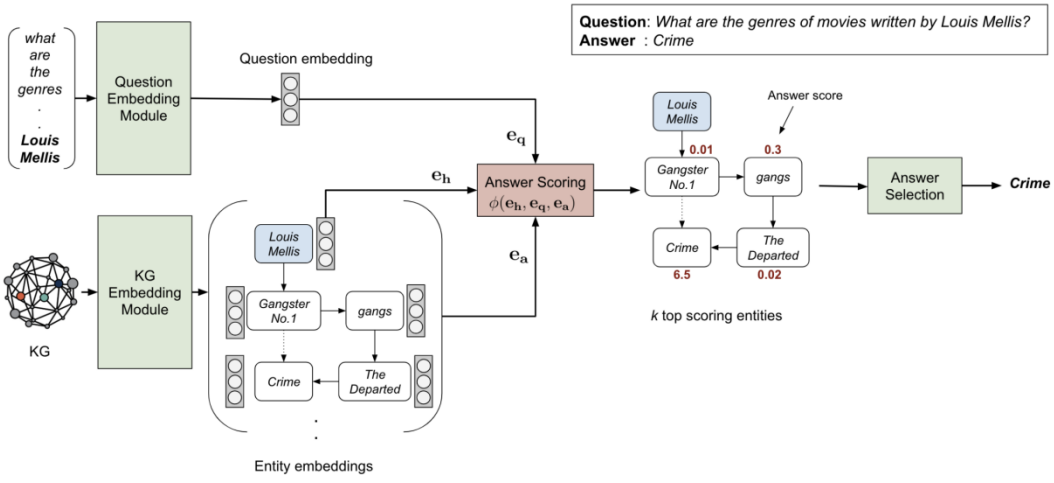


图 11 EmbedKGQA 的框架概览

模式归纳方法是从知识图谱中学习模式或者规则，并发现发现潜在的(概率的)逻辑。 例如，AMIE [Gal´arraga et al., 2015]和 Any-BURL [Meilicke et al., 2019]则是基于符号的典型模式归纳方法。然而，仅从显式三元组中学习较多高质量的模式与规则是存在一定困难的。嵌入表示模型的引入，可以有效地帮助传统模式归纳方法跳出图谱存在噪声以及自身不完全性的困境。Ho 等人采用嵌入表示方法的基础上提出了 RuLES 模型[Ho et al., 2018]。模型在原有知识图谱的基础上，扩展了带有一定可信度的三元组，并采用迭代方式从原始的知识图谱三元组与扩展带有可信度的三元组归纳出更多的规则。另一种方式则是通过在向量空间中，采用端到端可微的方式进行规则挖掘，该类方法使用知识图谱中“关系”的嵌入表示模块来学习使用规则中每个操作算子的权重[Sadeghian et al., 2019]。Yang 等人提出的模型 NeuralLP [Yang et al., 2017]将一阶逻辑规则的参数和结构学习共同建模在端到端可微分的模型中。作者设计了一种学习组合这些操作的神经控制器系统来将推理任务编译成可微操作的序列，以此实现一阶逻辑规则的规则学习。Cohen 等人提出的概率逻辑 Tensorlog[Cohen et al., 2020]则是将一阶规则的推理和稀疏矩阵乘法之间建立了联系。模型将某些类型的逻辑推理任务编码成一系列可微的数字矩阵运算，从而进一步学习到带概率的规则。Wang 等人[Wang et al., 2020]则围绕知识图谱定义了一种可微分的知识框架，该框架能挖掘具有数值特征的规则。此外，嵌入表示模型还可以进一步提升规则学习的效率。例如，Omran 等人提出的模型 RLvLR[Omran et al., 2018]利用嵌入表示模型 RESCAL[Nickel et al., 2011]在模式归纳的过程中起到了空间搜索剪枝的作用。

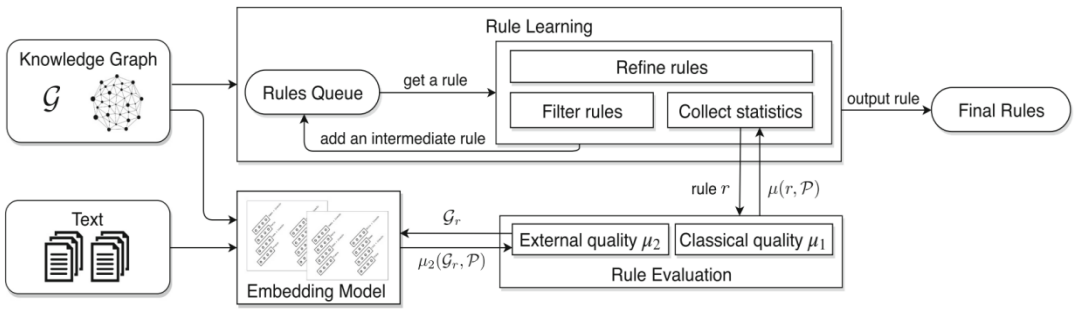


图 12 RuLES 的框架概览

为了解决概念漂移中学习和预测问题，Chen 等人[Chen et al., 2017]将嵌入表示模型生成的向量重新建模为语义特征(如：一致性向量与蕴涵向量)。这种嵌入表示模型生成的向量可以在监督流学习的环境中被利用来学习统计模型，对于概念漂移具备一定鲁棒性。此外，他们探索了一种基于本体的知识表示和推理框架[Chen et al., 2018]，用于迁移学习的解释。框架可以用表达力丰富的 OWL 本体对迁移学习中的学习域进行建模，并对学习域中的常识知识进行补充。作者设计了一个相关的推理算法来推断三种解释证据，以解释一个学习领域的

积极特征或消极迁移。Lécuéet 等人则讨论了迁移学习表达中语义设置等问题[Lécuéet al., 2019]。针对现有的基于实例的迁移学习方法，作者通过利用语义可变性、可迁移性以及一致性来处理迁移的对象以及迁移的时机。

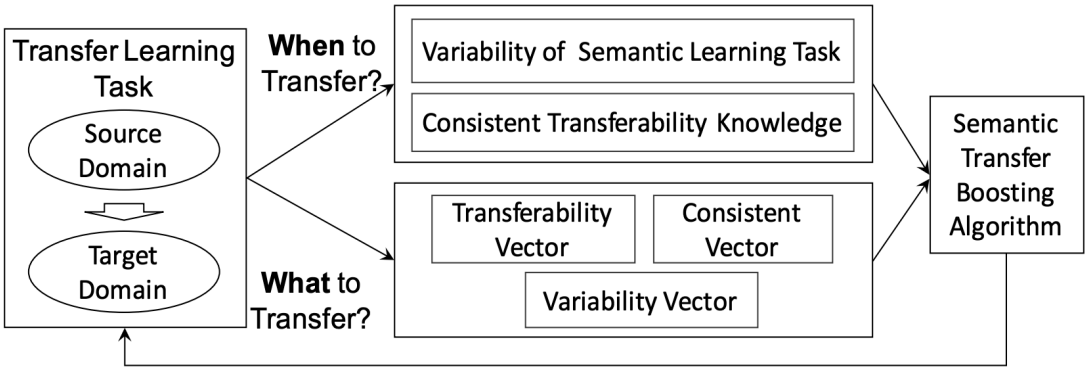


图 13 基于本体的迁移学习增强框架

四、技术展望与发展趋势

近十年，大数据和人工智能得到了快速发展，知识图谱和本体等知识表达模式逐渐在复杂语句问答、查询问答等研究领域被广泛应用。本章从知识图谱推理的应用层面出发，对基于本体物化推理、基于神经网络和本体表示学习的知识推理以及基于符号逻辑与嵌入表示的混合推理进行了介绍。虽然，这些知识图谱上的推理技术已经取得了很多进展，但是总体来说大部分推理技术离实际应用还有一段距离，存在一些问题需要完善解决。

1．本体物化推理技术的展望与发展趋势

在本体物化推理方法层面，部分物化算法和查询重写是解决物化过程中本体推理得到的结果集无穷问题的常用方法。通过查询应用改写和近似技术可靠地扩展部分物化算法支持表达性的本体语言。再通过等价角色和逆角色改写算法，提升物化效率。此外，查询分析算法扩展物化算法在保证物化方法的可靠完备性问题也是未来的研究热点。

在本体物化推理应用层面，将数据驱动的深度学习模块和知识引导的本体推理模块结合提升本体推理能力将是此类研究的热点问题。此外，在物联网边缘传感器中嵌入具有推理能力的感应器能大大提升设备感知能力，推进人工智能从感知阶段迈向认知阶段。

2．基于神经网络和本体表示学习的知识推理的展望与发展趋势

从推理方法来说，知识图谱嵌入与预训练模型可以学习实体和关系的表示，本体嵌入表示可以学习类的层次关系以及各种公理和规则，图神经网络方法可以充分捕获图的邻接信息以及子图结构，这些仍将是当前知识图谱推理的研究热点。同时为了完成更高层次的推理，

多种方法需要同时运用，进一步深挖更复杂的逻辑规则结构，因此，将神经网络方法与逻辑表示集成的混合推理方法也将是未来的研究热点。

从应用角度分析，目前的知识图谱推理方法在中等规模的标准数据集上取得了不错的效果和明显的进步，但面对超大规模、低资源以及人机协作的应用依然面临众多挑战。例如，在超大规模应用上提升推理方法的易用性和推理效率，在低资源应用中提升推理方法的鲁棒性和可迁移性，在人机协作应用中提升可解释性和人可介入性都是在把知识图谱推理实用化的道路上需要重点关注的问题。而在工业应用中，如何将不同的推理进行有效的融合，以此实现知识图谱中缺失知识补全、错误的信息矫正、高质量的规则挖掘，从而提升意图识别、实体链接、实体推荐等任务的实际效果，仍是诸多工业界关注的核心问题。

3．基于符号逻辑与嵌入表示的混合推理的展望与发展趋势

从推理方法来说，逻辑和嵌入表示集成的一个关键挑战在于符号逻辑的多样性。目前的大多数混合推理模型只考虑特定种类的逻辑规则与查询语句，因此，未来研究的热点一方面仍是探索更多的逻辑形式与嵌入表示进行混合推理。例如，探究如何利用嵌入表示模型来支撑常数的规则、全称量词(∀)、逻辑不一致检测的推理任务。另一方面，如何利用符号逻辑的表示与推理来增强嵌入表示模型的可解释性仍是未来研究的重点。只有当这些嵌入表示向量的含义被正确解释时，系统才具有更高的安全性、可信性和公平性，才能使得知识图谱推理方法在未来得到更广泛的应用。

从推理应用层面来看，相应算法的平台化和工具化也是未来知识图谱推理亟需完善的方向。即便现今存在较多前沿的工作，但我们发现部分工作仍缺少类似于 OpenKG 这样友好的平台来进行推广。例如，大多模式归纳算法的实现缺少友好的用户界面，也缺少相应的平台将其他的相似功能的方法进行整合，因此，期待有更多的团队能将这些知识图谱推理算法进行集成、维护，形成友好的项目工具，为知识图谱推理算法的应用与普及做出贡献。

参考文献

[Calvanese et al. 2017] Calvanese D, Cogrel B, Komla-Ebri S, et al. Ontop: answering SPARQL queries over relational databases. Semantic Web, 8 (3): 471–487(2017)

[Meng et al. 2018] Meng C, Zhang X, Xiao G, et al. gOWL: a fast Ontology-Mediated Query answering[C]. In ISWC (poster), 2018.

[Qin et al. 2021] Qin X, Zhang X, Yasin M Q, et al. SUMA: A Partial Materialization-Based Scalable Query Answering in OWL 2 DL. Data Science and Engineering, 6(2): 229-245 (2021)

[Xiao et al. 2019] Xiao G, Ding L, Cogrel B, et al. Virtual Knowledge Graphs: An Overview of

Systems and Use Cases, Data Intelligence, 1(3):23(2019)

[Han et al. 2022] Han X, Dell'Aglio D, Grubenmann T, et al. A framework for differentially-private knowledge graph embeddings. Journal of Web Semantics, 72: 100696(2022)

[Bienvenu et al. 2016] Bienvenu, M.: Ontology-mediated query answering: Harnessing knowledge to get more from data. In IJCAI, 2016: 4058-4061.

[Chortaras et al. 2011] Chortaras, A., Trivela, D., Stamou, G. Optimized query rewriting for OWL2QL. In CADE, 2011: 192-206.

[Thomazo et al. 2013] Thomazo, M. Compact rewritings for existential rules. In IJCAI, 2013: 1125-1131.

[Bursztyn et al. 2016] Bursztyn, D., Goasdou´e, F., Manolescu, I. Teaching an RDBMS about ontological constraints. In: VLDB Endowment 9(12): 1161-1172 (2016)

[Leone et al. 2019] Leone, N., Manna, M., Terracina, G., Veltri, P.: Fast query answering over existential rules. ACM Transactions on Computational Logic, 20(2): 12:1-12:48 (2019)

[Kontchakov et al. 2010] Kontchakov, R., Lutz, C., Toman, D., Wolter, F., Zakharyaschev, M.: The combined approach to query answering in dl-lite. In KR, 2010.

[Calvanese et al. 2007] Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Tractable reasoning and efficient query answering in description logics: The DL-Lite family. Journal of Automated Reasoning. 39(3): 385-429 (2007)

[Xiao et al. 2021] Guohui Xiao, Julien Corman, Ontology-Mediated SPARQL Query Answering over Knowledge Graphs, Big Data Research, 23: 100177 (2021)

[Zhou et al. 2015] Zhou Y, Grau B C, Nenov Y, et al. PAGOdA: Pay-As-You-Go ontology query answering using a datalog reasoner. Journal of Artificial Intelligence Research, 54: 309–367(2015)

[Horrocks et al. 2006] Horrocks I, Kutz O, Sattler U. The even more irresistible SROIQ[C]. In KR, 2006: 57–67.

[Nenov et al. 2015] Nenov Y, Piro R, Motik B, et al. RDFox: a highly-scalable RDF store[C]. In ISWC, 2015: 3–20.

[Motik et al. 2009] Motik B, Shearer R, Horrocks I. Hypertableau reasoning for description logics[J]. Journal of Artificial Intelligence Research, 36: 165–228(2009)

[Ahmetaj et al. 2021] Ahmetaj S, Efthymiou V, Fagin R, et al. Ontology-enriched query answering on relational databases. In AAAI, 2021: 15247-15254.

[Li et al. 2022] Li Y, Ouyang S, Zhang Y. Combining deep learning and ontology reasoning for

remote sensing image semantic segmentation. Knowledge-Based Systems, 243: 108469(2022)

[Ruta et al. 2022] Ruta M, Scioscia F, Bilenchi I, et al. A multi-platform reasoning engine for the Semantic Web of Everything. Journal of Web Semantics, 73: 100709(2022)

[Li et al., 2020] Weizhuo Li, Guilin Qi, and Qiu Ji. Hybrid reasoning in knowledge graphs: Combing symbolic reasoning and statistical reasoning. Semantic Web, 11(1): 53-62 (2020).

[Chen et al., 2020] Xiaojun Chen, Shengbin Jia, and Yang Xiang. A review: Knowledge reasoning over knowledge graph. Expert Systems with Applications 141: 112948 (2020).

[Zhang et al., 2022] Wen Zhang, Jiaoyan Chen, Juan Li, Zezhong Xu, Pan Jeff Z., and Huajun Chen. "Knowledge Graph Reasoning with Logics and Embeddings: Survey and Perspective." arXiv preprint arXiv:2202.07412 (2022).

[Niu et al., 2020] Guanglin Niu, Yongfei Zhang, Bo Li, Peng Cui, Si Liu, Jingyang Li, Xiaowei Zhang: Rule-Guided Compositional Representation Learning on Knowledge Graphs. AAAI 2020: 2950-2958.

[Guo et al., 2016] Shu Guo, Quan Wang, Lihong Wang, Bin Wang, and Li Guo. Jointly embedding knowledge graphs and logical rules. In EMNLP, 2016: 192-202.

[Guo et al., 2018] Shu Guo, Quan Wang, Lihong Wang, Bin Wang, and Li Guo. Knowledge graph embedding with iterative guidance from soft rules. In AAAI, 2018: 4816-4823.

[Zhang et al., 2019] Wen Zhang, Bibek Paudel, Liang Wang, Jiaoyan Chen, Hai Zhu, Wei Zhang, Abraham Bernstein, Huajun Chen: Iteratively Learning Embeddings and Rules for Knowledge Graph Reasoning. WWW 2019: 2366-2377.

[Wang et al., 2018b] Mengya. Wang, Erhu Rong, Hankui Zhuo, and Huiling Zhu. Embedding knowledge graphs based on transitivity and asymmetry of rules. In PAKDD, 2018: 141-153.

[Guu et al., 2015] K. Guu, J. Miller, and P. Liang. Traversing knowledge graphs in vector space. In EMNLP, 2015: 318-327.

[Hamilton et al., 2018] William L. Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. "Embedding logical queries on knowledge graphs." Advances in neural information processing systems 31 (2018).

[Ren et al., 2020a] Hongyu Ren, Jure Leskovec: Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs. NeurIPS 2020.

[Ren et al., 2020b] Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. In ICLR, 2020.

[Arakelyan et al., 2021] Erik Arakelyan, Daniel Daza, Pasquale Minervini, and Michael Cochez. Complex query answering with neural link predictors. In ICLR, 2021.

[Trouillon et al., 2016] T. Trouillon, J. Welbl, S. Riedel, E., and G. Bouchard. Complex embeddings for simple link prediction. In ICML, 2016.

[Kotnis et al., 2021] B. Kotnis, C. Lawrence, and M. Niepert. Answering complex queries in knowledge graphs with bidirectional sequence encoders. In AAAI, 2021: 4968-4977.

[Liu et al., 2021] Lihui Liu, Boxin Du, Heng Ji, ChengXiang Zhai, and Hanghang Tong. Neural-answering logical queries on knowledge graphs. In KDD, 2021: 1087-1097.

[Wei et al., 2015] Zhuoyu Wei, Jun Zhao, Kang Liu, Zhenyu Qi, Zhengya Sun, and Guanhua Tian. Large-scale knowledge base completion: Inferring via grounding network sampling over selected instances. In CIKM, 2015: 1331-1340.

[Rockt¨aschel and Riedel, 2017] T. Rockt¨aschel and S. Riedel. End-to-end differentiable proving. In NIPS, 2017: 3788-3800.

[Minervini et al., 2020a] Pasquale Minervini, Matko Bosnjak, Tim Rockt¨aschel, Sebastian Riedel, and Edward Grefenstette. Differentiable reasoning on large knowledge bases and natural language. In AAAI, 2020: 5182-5190.

[Minervini et al., 2020b] Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, and Tim Rockt¨aschel. Learning reasoning strategies in end-to-end differentiable proving. In ICML, 2020: 6938-6949.

[Abujabal et al., 2017] Abdalghani, Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum, QUINT: Interpretable Question Answering over Knowledge Bases, In EMNLP, 2017: 61-66.

[Hao et al., 2017] Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge, In ACL, 2017: 221-231.

[Zhang et al., 2018] Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J. Smola, and Le Song. Variational Reasoning for Question Answering with Knowledge Graph, In AAAI, 2018: 6069-6076.

[Zhou et al., 2018] Mantong Zhou, Minlie Huang, and Xiaoyan Zhu, An Interpretable Reasoning Network for Multi-Relation Question Answering, In: ACL, 2018: 2010-2022.

[Vakulenko et al., 2019] Svitlana Vakulenko, Javier David Fernandez Garcia, Axel Polleres, Maarten de Rijke, and Michael Cochez Message Passing for Complex Question Answering over

Knowledge Graphs, In CIKM, 2019: 1431-1440.

[Saxena et al., 2020] Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In ACL, 2020: 4498-4507.

[Liu et al., 2021] Zhuang, Liu, Lin Wayne, Shi Ya, and Zhao Jun. A Robustly Optimized BERT Pre-training Approach with Post-training. In CCL, 2021: 471-484.

[Gal´arraga et al., 2015] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M. Suchanek. Fast rule mining in ontological knowledge bases with AMIE+. VLDB Journal. 24(6): 707-730 (2015).

[Meilicke et al., 2019] Christian Meilicke, Melisachew Wudage Chekol, Daniel Ruffinelli, and Heiner Stuckenschmidt. Anytime bottom-up rule learning for knowledge graph completion. In IJCAI, 2019: 3137-3143.

[Ho et al., 2018] Vinh Thinh Ho, Daria Stepanova, Mohamed H. Gad-Elrab, Evgeny Kharlamov, and Gerhard Weikum. Rule learning from knowledge graphs guided by embedding models. In ISWC, 2018: 72-90.

[Sadeghian et al., 2019] Ali, Sadeghian, Mohammadreza Armandpour, Patrick Ding, and Daisy Zhe Wang. DRUM: end-to-end differentiable rule mining on knowledge graphs. In NeurIPS, 2019: 15321-15331.

[Yang et al., 2017] Fan Yang, Zhilin Yang, and William W. Cohen. Differentiable learning of logical rules for knowledge base reasoning. In NIPS, 2017: 2319-2328.

[Cohen et al., 2020] William Cohen, Fan Yang, and Kathryn Rivard Mazaitis. Tensorlog: A probabilistic database implemented using deep-learning infrastructure. Journal of Artificial Intelligence Research. 67: 285-325 (2020)

[Wang et al., 2020] Po-Wei Wang, Daria Stepanova, Csaba Domokos, and J. Zico Kolter. Differentiable learning of numerical rules in knowledge graphs. In ICLR, 2020.

[Omran et al., 2018] Pouya Ghiasnezhad Omran, Kewen Wang, and Zhe Wang. Scalable rule learning via learning representation. In IJCAI, 2018: 2149-2155.

[Nickel et al., 2011] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In ICML, 2011: 809-816.

[Chen et al., 2017] Jiaoyan Chen, Freddy Lécué, Jeff Pan, and Huajun Chen. Learning from Ontology Streams with Semantic Concept Drift, In AAAI, 2017: 957-963.

[Chen et al., 2018] Chen, Jiaoyan, Freddy Lécué, Jeff Z. Pan, Ian Horrocks, and Huajun Chen. Knowledge-Based Transfer Learning Explanation, In KR, 2018: 349-358.

[Lécuéet al., 2019] Freddy Lécué, Jiaoyan Chen, Jeff Z. Pan, and Huajun Chen. Augmenting transfer learning with semantic reasoning. In IJCAI, 2019: 1779-1785.

[Bordes et al., 2013] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, Oksana Yakhnenko: Translating Embeddings for Modeling Multi-relational Data. In NIPS, 2013: 2787-2795.

[Wang et al., 2014] Zhen Wang, Jianwen Zhang, Jianlin Feng, Zheng Chen: Knowledge Graph Embedding by Translating on Hyperplanes. In AAAI, 2014: 1112-1119.

[Théo et al., 2016] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, Guillaume Bouchard: Complex Embeddings for Simple Link Prediction. In ICML, 2016: 2071-2080.

[Sun et al., 2019] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, Jian Tang: RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In ICLR (Poster), 2019.

[Ralph et al., 2020] Abboud Ralph, Ismail Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. Boxe: A box embedding model for knowledge base completion. In NeurIPS, 2020: 9649-9661.

[Zhang et al., 2021] Wen Zhang, Chi Man Wong, Ganqiang Ye, Bo Wen, Wei Zhang, Huajun Chen: Billion-scale Pre-trained E-commerce Product Knowledge Graph Model. In ICDE, 2021: 2476-2487.

[Schlichtkrull et al., 2018] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, Max Welling: Modeling Relational Data with Graph Convolutional Networks. In ESWC, 2018: 593-607.

[Vashishth et al., 2020] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, Partha P. Talukdar. Composition-based Multi-Relational Graph Convolutional Networks. In ICLR, 2020.

[Komal et al., 2020] Teru Komal, Etienne Denis, and Will Hamilton. Inductive relation prediction by subgraph reasoning. In International Conference on Machine Learning, In ICML, 2020: 9448-9457

[Sijie et al., 2021] Mai Sijie, Shuangjia Zheng, Yuedong Yang, and Haifeng Hu. Communicative message passing for inductive relation reasoning. In AAAI, 2021: 4294-4302.

[Kulmanov et al., 2019] Maxat Kulmanov, Wang Liu-Wei, Yuan Yan, Robert Hoehndorf: EL Embeddings: Geometric Construction of Models for the Description Logic EL++. In IJCAI, 2019: 6103-6109.

[Chen et al., 2020] Jiaoyan Chen, Pan Hu, Ernesto Jiménez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, Ian Horrocks: OWL2Vec*: embedding of OWL ontologies. Machine Learning. 110(7): 1813-1845 (2021).

第八章 知识图谱的存储和查询

彭鹏[1] 邹磊[2]

1. 湖南大学 信息科学与工程学院，湖南省 长沙市 410082

2. 北京大学 王选计算机研究所，北京 100080

“知识图谱”就是以图（Graph）的方式来展现“实体”、实体“属性”，以及实体之间的“关系”。目前知识图谱普遍采用了万维网联盟（W3C）所发布的 RDF[1]（Resource Description Framework，资源描述框架）模型来表示数据。RDF 被设计为一种通用的网络资源描述方法，被广泛地应用在表示知识图谱。本章将从数据管理的角度去介绍基于 RDF 模型的知识存储和查询方面的研究和应用问题。

一、背景和任务定义

RDF 是用于描述现实中资源的 W3C 标准。现实中任何实体都可以表示成 RDF 模型中的资源，比如大学的名称、地点、建校时间以及校长。资源以唯一的 IRI（国际化资源标识符——Internationalized Resource Identifiers）来表示，不同的资源拥有不同的 IRI。这些资源可以用来作为知识图谱中对客观世界的概念、实体和事件的抽象。

图 1 给出了一个 RDF 数据实体示例，用来表示现实中一个著名欧洲哲学家亚里士多德（Aristotle）。在 RDF 数据模型中，亚里士多德就能通过亚里士多德头像上方所示的 IRI 来进行唯一标识。客观世界的概念、实体和事件很多都是有属性的。图 1 中亚里士多德头像下方给出的属性和属性值描述了亚里士多德这个资源所对应的人的名字是“亚里士多德”。此外，客观世界中不同概念、实体和事件相互之间可能会有各种关系，所以 RDF 模型中不同资源之间也是会存在关系。比如，图 1 给出了亚里士多德和另一个表示希腊城市卡尔基斯（Chalcis）所对应的资源通过一个 placeOfDeath 关系连接了起来，描述了亚里士多德死于卡尔基斯这个事实。

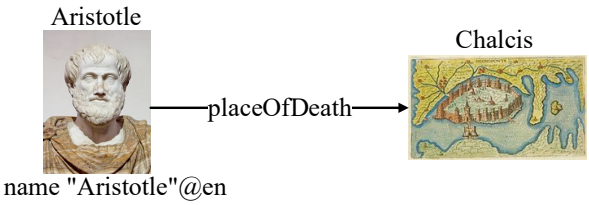


图 1 示例 RDF 资源

[1] https://www.w3.org/RDF/

利用这些属性和关系，很多资源就被连接起来形成了 RDF 数据集。每个资源的一个属性及属性值，或者它与其他资源的一条关系，都被称为一条知识。上述属性以及关系就能表示成三元组。每一条三元组又可被称为一条陈述。一条陈述包含三个部分，通常称之为主体、谓词和客体。其中主体一定是一个被描述的资源。谓词可以表示主体的属性，或者表示主体和客体之间某种关系。当谓词表示属性时，客体就是属性值，通常是一个字面值；否则，客体是另外一个资源。

图 2 的展示了一个著名 RDF 数据集 DBpedia[Jens Lehmann et al, 2015]的片段。这个片段中包括 15 条陈述，描述了欧洲哲学家 Aristotle（亚里士多德）和 Boethius（波伊提乌）所对应的实体及其相关陈述。

主体	谓词	客体
Aristotle	influencedBy	Plato
Aristotle	mainInterest	Ethics
Aristotle	mainInterest	Physics
Aristotle	name	"Aristotle"
Aristotle	placeOfDeath	Chalcis
Boethius	influencedBy	Aristotle
Boethius	mainInterest	Religion
Boethius	name	"Boethius"
Boethius	placeOfDeath	Pavia
Plato	name	"Plato"
Chalcis	country	Greece
Chalcis	postalCode	341 00
Pavia	country	Italy
Pavia	postalCode	27100

图 2 示例 RDF 三元组

面向 RDF 数据集，W3C 提出了一种结构化查询语言 SPARQL[1]，它类似于面向关系数据模型的查询语言 SQL，是一种描述性的结构化查询语言。用户只需要按照 SPARQL 定义的语法规则去描述其想查询的信息即可，而不需要明确指定系统进行查询执行的具体步骤。对于一个 SELECT 语句中，SELECT 子句指定查询应当返回的内容，FROM 子句指定将要使用的数据集，WHERE 子句一组三元模式组成用以指定所返回的 RDF 知识图谱数据片段需要满足的条件。

图 3(a)给出了一个针对哲学家的 SPARQL 查询，目标在于查询出所有“受过亚里士多德影响的伦理学相关的哲学家”。这个查询在图 2 所示 RDF 数据集上所对应的匹配如图 3(b)

[1] http://www.w3.org/TR/rdf-sparql-query/.

所示，即“受过亚里士多德影响的伦理学相关的哲学家”有波伊提乌（Boethius）。

```
SELECT  ?x ?n  WHERE {
?x mainInterest Ethics.
?x influencedBy Aristotle.
?x name ?n.
}
```

?x ⟶ Boethius

?n ⟶ "Boethius"

(a) (b)

图 3 示例 SPARQL 查询

RDF 数据也可以被表示成图的形式。其中，每个实体或者字面值可以被视为图上的点，每个陈述视为连接主体及客体的有向边，而陈述中的谓词就可以视为有向边上的标签。从语义角度上看，RDF 数据本质上就是通过预先定义的语义构成的一个或多个连通图。V.Bönström 等人提出[Valerie Bonstrom et al, 2003]，相比于将 RDF 数据视为 XML 格式数据或三元组的集合，RDF 的图模型可以更好地体现 RDF 数据中涵盖的语义信息。

图 4 展示了图 2 所示 RDF 知识图谱数据集所对应的 RDF 数据图。图 4 中所有的资源都是椭圆，而字面值都是矩形。

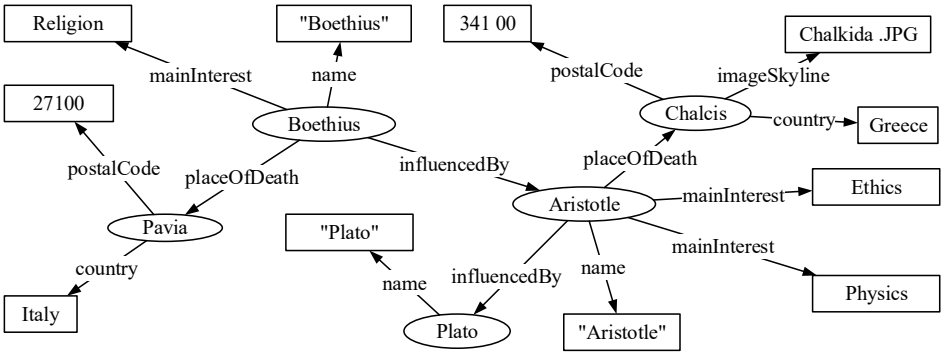


图 4 示例 RDF 数据图

与 RDF 数据的图形式表示类似，一个 SPARQL 查询可以表示为一个查询图。查询中每个变量或者常量对应一个查询图上的点，每个 WHERE 子句中的三元模式对应一条边。图 5 给出了一个图 3(a)所示 SPARQL 查询所对应的查询图。

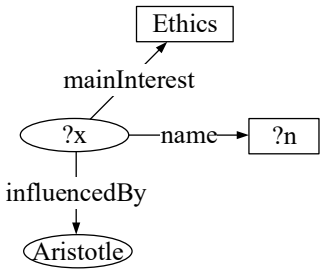


图 5 示例 SPARQL 查询图

现有 RDF 数据存储与查询的基本问题就是：给定一个 RDF 数据集 G 和 SPARQL 查询 Q，找出 Q 在 G 上的匹配。当 RDF 数据和 SPARQL 查询都转化成图的形式，SPARQL 查询语句的查询结果就是其所对应的查询图在 RDF 数据图上的子图匹配[Lei Zou et al, 2011, Lei Zou et al, 2014]。

研究内容和挑战问题

由于 RDF 结构的灵活性，现在 RDF 数据的应用范围日益广阔。越来越多的数据开始表示成 RDF 格式。比如，德国的莱比锡大学和柏林自由大学合作从维基百科上抽取结构化数据形成的知识图谱 DBpedia[Jens Lehmann et al, 2015]已将近有 24 亿条边；另外，德国马克斯普朗克研究所结合维基百科以及 wordnet 所形成的 RDF 知识图谱 YAGO[Fabian M. Suchanek et al,2008 , Johannes Hoffart et al,2013 , Farzaneh Mahdisoltani et al,2015]也已经有将近 2 亿条边；而由用户编辑和维基百科中信息抽取共同形成的 RDF 知识库 Freebase[1]也已经有 19 亿条边。

因此 RDF 数据管理的挑战问题是如何对上述大规模 RDF 知识图谱进行高效的存储和查询。在查询处理过程中，我们需要将 SPARQL 查询图中变量与 RDF 数据图上点进行绑定以得到所有 SPARQL 查询图在 RDF 数据图上的子图匹配。学术界和工业界当前已经构建了不少高效的 RDF 数据管理系统来进行 SPARQL 查询处理。

二、技术方法和研究现状

知识图谱数据管理的挑战问题是如何对大规模 RDF 知识图谱进行高效的存储和查询。总的来说，有两套完全不同的思路。第一种思路是利用已有的成的关系数据库系统来存储和管理 RDF 知识图谱数据，同时将面向 RDF 数据的 SPARQL 查询转换为面向关系数据库的 SQL 查询，利用已有的关系数据库产品或者相关技术来回答查询。这里面最核心的研究问题是如何构建关系表来存储 RDF 数据，并且使得转换的 SQL 查询语句查询性能更高；其二是直接开发面向 RDF 知识图谱数据的 Native 的知识图谱数据存储和查询系统（Native RDF 图数据库系统），考虑到 RDF 数据管理的特性，从数据库系统的底层进行优化。针对以上两个方面的思路，我们分别加以介绍。

1. 基于关系数据模型的 RDF 数据存储和查询

在数据管理方面，关系数据模型自提出以来取得了巨大成功。市面上已经产生了大量成熟的关系数据库。而 RDF 数据的三元组模型可以很容易完成对于关系模型的映射。因此，

[1] http://www.freebase.com/.

不少研究者都尝试使用关系数据模型来设计 RDF 存储方案。下面介绍几种经典的方法。

1）简单三列表

现在已经有不少比较成熟的系统来利用关系数据库进行数据管理，包括 Jena[Kevin Wilkinson et al,2003 , Kevin Wilkinson, 2006]、Oracle[Eugene Inseok Chong et al, 2005]、Sesame[Jeen Broekstra et al, 2003 , Jeen Broekstra et al, 2002]、3store[Stephen Harris & Nicholas Gibbins, 2003]以及 SOR[Jing Lu et al, 2007]。这些系统通过维护一张巨大的三元组表来管理 RDF 数据。这张三元组表包含三列。这三列分别对应存储主体、谓词和客体（或者主体、属性和属性值）。当系统接收到用户输入的 SPARQL 查询时，这些系统将 SPARQL 查询转化为 SQL 查询。然后，根据所得 SQL 查询，这些系统通过对三元组表执行多次自连接操作以得到最终解。

虽然这种方法具有很好的通用性，但最大的问题是查询性能差。首先这张三列表的规模可能非常庞大。而且这种方法可能会产生大量的自连接操作，而在关系数据库系统中自连接操作非常耗时，特别是对于那些数据规模很大的表而言。所以这些方法都有很大的局限。

2）水平存储

水平方法（Horizontal Schema）[Zhengxiang Pan & Jeff Heflin, 2003] 是将知识图谱中的每一个 RDF 主体（subject）表示为数据库表中的一行。表中的列包括该 RDF 数据集合中所有的属性。这种的策略的好处在于设计简单，同时很容易回答面向某单个主体的属性值的查询，即星状查询。然而这种水平存储方法的缺点也是很明显的：其一，因为属性数目很可能超过当前数据库能够承受的数量，所以表中存在大量的列；其二，因为主体并不在所有的属性上有值，所以表中将存在大量空值；。其三，因为主体在属性上可能有多个值，所以水平存储存在多值性的问题；其四，数据的变化可能带来很大的更新成本。在实际应用中，数据的更新可能导致增加属性或删除属性等改变，但是这就涉及到整个表结构的变化，水平结构很难处理类似的问题。

3）属性表

为了降低自连接操作次数，Jena[Kevin Wilkinson et al,2003 , Kevin Wilkinson, 2006]和 Oracle[Eugene Inseok Chong et al, 2005]在单张大三元组表之外还支持利用属性表进行 RDF 数据管理。具体而言，Jena 通过聚类的方式将一些类似的三元组聚类到一起，然后将每一个聚类的三元组统一到一张属性表中进行管理，这种方式下的属性表也被称之为聚类属性表；而 Oracle 利用 RDF 资源的类型信息将三元组进行分类，相同类的三元组放到同一张表中，这种方式下的属性表也被称之为分类属性表。对于上述两种情况，由于 RDF 数据表示的灵活性，会存在部分三元组无法放入任何一个属性表示。此时，Jena 和 Oracle 将这个

三元组另起一张表来进行管理。

属性表也有着一些先天性的缺陷。其一，虽然属性表对于某些查询能够提高查询性能，但是大部分的查询都会涉及多个表的连接或合并操作。其二，RDF 数据由于来源庞杂，其结构性可能较差，从而属性和主体间的关联性可能并不强，类似的主体可能并不包含相同的属性，于是空值的问题就出现了；其三，在现实中，一个主体在一个属性上可能存在多值，然后用 RDBMS 管理这些数据时就带来麻烦。

4）垂直划分策略

针对属性表的问题，SW-Store[Daniel J. Abadi et al,2009]提出了对 RDF 数据按照谓词（或属性）分割成若干表的方法。具体而言，SW-Store 将 RDF 三元组按照谓词（或属性）的不同分成不同的表，每张表能保存在谓词（或属性）上相同的三元组。SW-Store 称这种方法为垂直分割。这种方法的优势在于能避免大量的自连接操作，而变成不同表之间的连接。因为在现有的关系数据库中不同表之间的连接操作要快于自连接操作，所以 SW-Store 能一定程度提高效率。但是，垂直分割缺点在于无法很好地支持 SPARQL 查询中某个三元组模式在谓词（或属性）上是变量的情况。

5）全索引策略

如前所述，简单的三列表存储的缺点在于自连接次数较多。为了提高简单三列表存储的查询效率，目前一种普遍被认可的方法是“全索引（exhaustive indexing）”策略。如 Hexastore [Cathrin Weiss et al,2008]和 RDF-3x[Thomas Neumann & Gerhard Weikum, 2008 , Thomas Neumann & Gerhard Weikum, 2010a , Thomas Neumann & Gerhard Weikum, 2010b]。它们为了加速 RDF 三元组在 SPARQL 查询处理过程中的连接操作速度都将三元组在主体、谓词、客体之间各种排列下能形成各种形态构建都枚举出来，然后为它们构建索引。这样建立的索引恰好是六重索引。

虽然用全索引策略可以弥补一些简单垂直存储的缺点，但三元存储方式难以解决的问题还有很多。其一，不同的三元组其主体/属性/属性值可能重复，这样的重复出现会浪费存储空间。其二，复杂的查询需要进行大量表连接操作，即使精心设计的索引可以将连接操作都转化为合并连接，当 SPARQL 查询复杂时，其连接操作的查询代价依然不可忽略。其三，随着数据量增长，表的规模会不断膨胀，系统的性能下降严重；而且目前此类系统都无法支持分布式的存储和查询，这限制了其系统的可扩展性。其四，由于数据类型多样，无法根据特定数据类型进行存储的优化，可能会造成存储空间的浪费(例如，客体的值可能多种多样，如 IRI、一般字符串或数值。客体一栏的存储空间必须满足所有的取值，而无法进行存储优化)。为了解决这个问题，目前的全索引方法都是利用字典方式将所有的

字符串和数值映射成一个独立的整数 ID。但是这种字典映射的方法很难支持带有数值范围约束和字符串中的子串约束的 SPARQL 查询。

2. 基于图模型的 RDF 数据存储和查询

如前文所述，通过将 RDF 三元组看作带标签的边，RDF 数据自然地符合图模型结构。因此，很多研究者从 RDF 图模型结构的角度看待 RDF 数据。RDF 数据的图模型可以最大限度地保持 RDF 数据的语义信息，也有利于对语义信息的查询。在这种情况下，SPARQL 查询就可以视为在 RDF 数据图上进行子图匹配运算。子图匹配运算是图数据库中一个比较经典的问题，其问题定义在于给定一个数据图和一个查询图，找出数据上所有与查询图子图同态的位置。这个问题已被证明是一个 NP 难问题。

针对 RDF 数据的 SPARQL 查询已经有一些基于图模型的查询处理系统，如 gStore[Lei Zou et al, 2011 , Lei Zou et al, 2014]、dipLODocus$_{[RDF]}$[Marcin Wylot et al, 2011] 和 Turbo$_{HOM++}$[Jinha Kim et al, 2015]。它们都是利用 RDF 数据图的特点来构建索引。

gStore[Lei Zou et al, 2011 , Lei Zou et al, 2014]是由北京大学计算机科学技术研究所数据管理实验室实现并维护的一个基于图的 RDF 知识图谱数据管理系统[1]。gStore 根据每个资源的所有属性和属性值映射到一个二进制位串上。图 6 显示 gStore 对一个 RDF 数据图进行二进制编码的示例。然后，gStore 将所有位串按照 RDF 背后对应的图结构组织成一棵签章树——VS*-tree。VS*-tree 被分为若干层，每一层都是整张 RDF 数据图的摘要。基于 VS*-tree，gStore 可以完成高效的数据存储、更新与查询操作。当 SPARQL 查询进入时，将每个查询点在这个 VSTree 上进行检索，找到相应候选解，然后再将这些候选解通过连接操作拼接起来。

图 7 展示了在 3-5 亿规模的三元组的国际标准测试集（LUBM 和 WatDiv）上，gStore 系统和目前使用最为广泛的 RDF 知识图谱存储查询系统 Virtuoso 和 Apache Jena 之间的查询性能对比情况。由于基于图结构方法的索引可以考虑到查询图整体信息，因此总的来说查询图越复杂（例如查询图的边越多），gStore 相对于对比系统的性能会更好，有的可以达到一个数量级以上的性能优势。gStore 的分布式版本的在 10 台机器组成的 Cluster 上可以进行 50-100 亿规模的 RDF 知识图谱管理的任务。

[1] gStore 项目主页：http://www.gstore.cn/pcsite/index.html

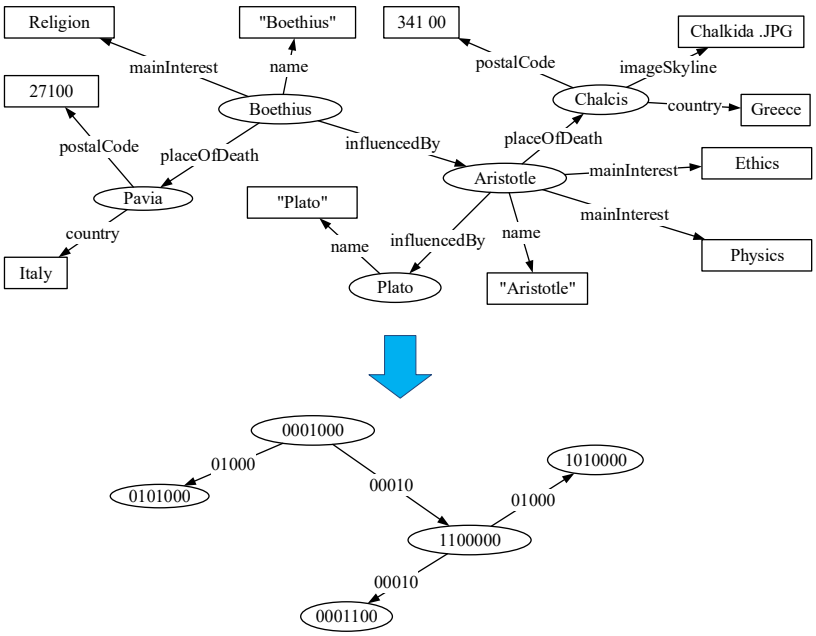
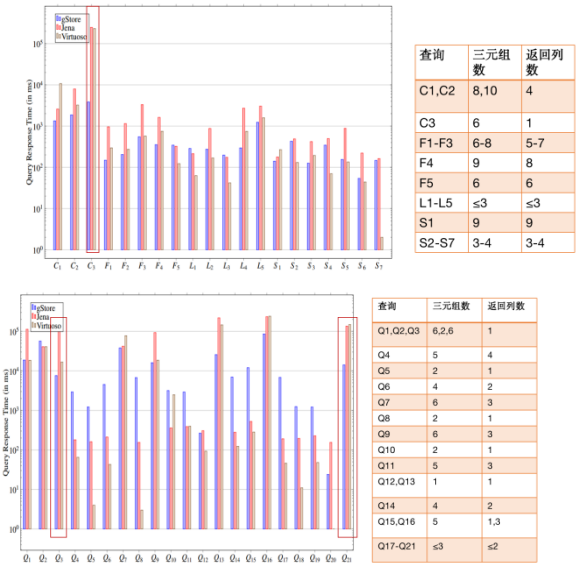


图 6 gStore 进行二进制编码的示例



查询	三元组数	返回列数
C1,C2	8,10	4
C3	6	1
F1-F3	6-8	5-7
F4	9	8
F5	6	6
L1-L5	≤3	≤3
S1	9	9
S2-S7	3-4	3-4

查询	三元组数	返回列数
Q1,Q2,Q3	6,2,6	1
Q4	5	4
Q5	2	1
Q6	4	2
Q7	6	3
Q8	2	1
Q9	6	3
Q10	2	1
Q11	5	3
Q12,Q13	1	1
Q14	4	2
Q15,Q16	5	1,3
Q17-Q21	≤3	≤2

(a) 在 WatDiv 3 亿边规模数据上的评测结果 (b) 在 LUBM 5 亿边规模数据上的评测结果

图 7 在国际通用 RDF 测评数据集上的比较结果

dipLODocus$_{[RDF]}$ [Marcin Wylot et al, 2011]提出一个同时利用 RDF 图结构与考虑数据分析需求的混合存储模式。所谓利用 RDF 数据图结构，就是挖掘出在 RDF 图中若干存储模式，然后将 RDF 数据图中满足这些存储模式的结构存在一起。所谓考虑数据分析需求，就是利用列存储技术存储数值型数据，即将满足某个存储模式的所有结构中特点位置的数值按列存储组织在一起以方便聚集性查询处理。

Turbo$_{HOM++}$[Jinha Kim et al, 2015]将子图匹配的技术应用到了 SPARQL 查询处理上。具体而言，Turbo$_{HOM++}$首先将 RDF 数据图基于每个资源的类信息转化为一般的普通数据图。

然后，Turbo$_{HOM++}$在 SPARQL 查询图上从一个选定查询点出发做宽度优先搜索，得到一颗树宽度优先搜索树。同时，Turbo$_{HOM++}$在数据图上从选定查询点的候选出发结合宽度优先搜索树深度做深度优先搜索，得到选定查询点候选的候选区域，并在这个候选区域中结合一定匹配顺序找到最终 SPARQL 查询的解。

3. 基于新硬件的 RDF 数据存储和查询

随着 GPU、RDMA 等新型硬件的广泛使用，基于这些新型硬件的 RDF 数据查询处理方法也被提了出来。典型的包括 TripleID-Q [Chantana Chantrapornchai & Chidchanok Choksuchat, 2018]、Wukong [Jiaxin Shi et al, 2016 , Zihang Yao et al, 2022]等。

针对 GPU 上的 SPARQL 查询处理，TripleID-Q[Chantana Chantrapornchai & Chidchanok Choksuchat, 2018]提出了一个压缩的 RDF 数据表示方式——TripleID。TripleID 本质是一个基于字典编码的三元组表。整个三元组表都是可以载入到 GPU 中去的。在 SPARQL 查询处理阶段，每个三元组模式通过 GPU 扫描 TripleID 来得到结果。然后，其他的 Union、Join、Filter 等操作在三元组模式的 GPU 扫描结果基础上进行实现。

Wukong [Jiaxin Shi et al, 2016 , Zihang Yao et al, 2022]是一个上海交通大学陈海波教授团队开发的一个基于 RDMA 的 RDF 数据管理系统。Wukong 基于一个 RDMA 上已有的分布式键值数据库 DrTM-KV 来存储 RDF 数据图的邻接表并实现 SPARQL 查询处理。之后，针对 GPU 读取数据带宽大但 GPU 本地内存容量小的特点，该团队进一步开发来 Wukong+G 来支持 GPU 对 Wukong 的优化。

4. 基于最坏情况下最优连接的查询优化

如前文中所介绍，知识图谱结构化查询处理可以转化为查询图在知识图谱上进行子图匹配，所以知识图谱上的查询处理任务往往包含大量的连接操作。因此，如何优化这些连接操作的执行方式以及顺序对于查询任务处理性能优化至关重要。近十年来，关于连接操作执行方式以及顺序的研究，数据库领域出现了一个最新的研究成果——最坏情况下最优连接技术（Worst-case Optimal Join）。该技术最早提出并应用在关系数据库领域，最近也开始应用到 RDF 数据管理方法中。

所谓最坏情况下最优连接技术是针对关系数据库上多表连接操作来提出。Albert Atserias、Martin Grohe 和 Dániel Marx 三位研究者在他们的论文[Atserias A et al, 2008 , Atserias A et al, 2013]针对多表连接最坏情况下的结果数量给出了一个上界，也就是 AGM 界（AGM 为三位作者的姓氏首字母）。之后，所有保证执行过程中满足 AGM 界的多表连接技术称为最坏情况下最优连接技术。最著名的满足 AGM 界的最坏情况下最优连接技术就是通用连接

（Generic Join）技术[Ngo H. Q. et al, 2014 , Ngo H. Q. et al, 2018]。所谓通用连接技术，它会首先确定一个关系边中属性的顺序，然后按照这个顺序依次对每个属性通过多路交集操作来实现多路连接。Jena-LTJ[Aidan Hogan et al, 2019]开始将最坏情况下最优连接技术整合到单机知识图谱查询处理引擎 Jena 中。

三、技术展望与发展趋势

因为 RDF 模型的灵活性，越来越多的知识图谱数据提供方将自身的知识图谱数据表示成 RDF 格式并发布到互联网上。这些发布在互联网上的 RDF 数据之间通过 IRI 相互链接起来，共同构成了一个庞大的覆盖整个互联网的知识图谱。这个庞大的覆盖整个互联网的知识图谱描述了整个互联网上的知识。这样，互联网就由一个文档的网络转化成一个数据的网络，而且是一个计算机可以理解的数据网络。为了让这个数据的网络更加的丰富和完善，W3C 在积极推进 LOD （Linked Open Data）项目[1]。这个项目目的就是将网络上的 RDF 数据集相互链接起来以增强数据可用性。当前，LOD 已成功令上千个 RDF 数据集相互链接在一起。图 8 展示了 LOD 项目中相互连接的数据集。

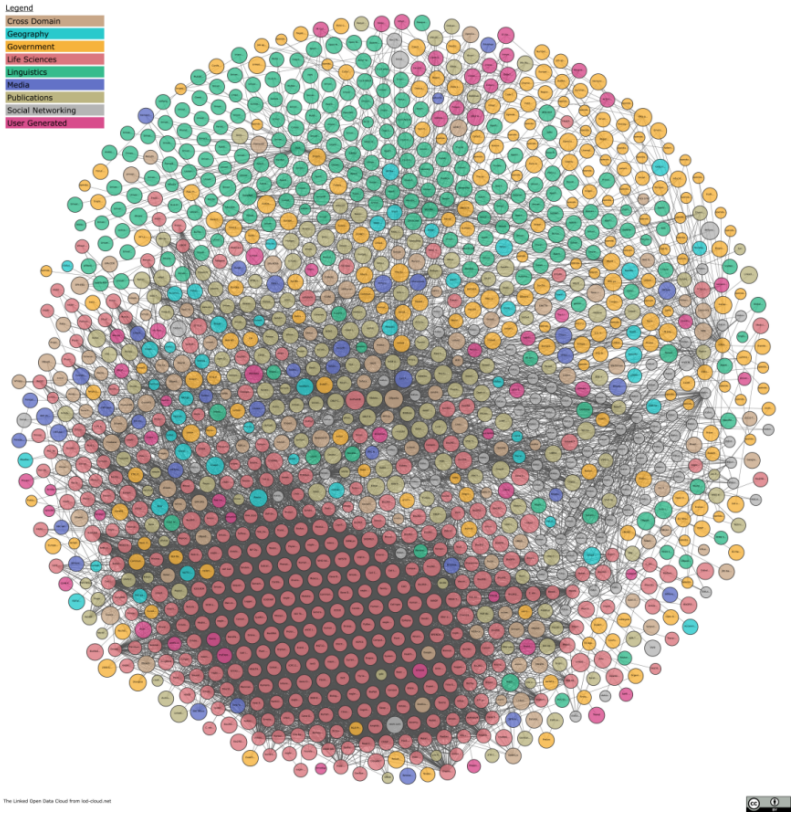


图 8 关联数据

[1] http://linkeddata.org/home.

随着互联网上 RDF 知识图谱数据集的数量与规模日益增长，互联网上的 RDF 知识图谱数据集已经远超了现有单机系统能力。于是，利用分布式数据库系统相关技术来进行 RDF 数据上的查询处理成为了未来研究的趋势。

现阶段，部分研究人员也已经有设计并实现了一些针对 RDF 数据的分布式查询处理方法。这些方法可以被分为三类：一类是基于已有云平台的分布式查询处理方法；一类是基于数据划分的分布式查询处理方法；还有一类是联邦型分布式 RDF 数据查询处理方法。

1. 基于已有云平台的分布式 RDF 数据查询处理方法

所谓基于已有云平台的分布式 RDF 数据查询处理方法都是利用已有云平台存储管理系统进行关联数据的存储，并利用这些已有云平台上成熟的任务处理模式进行查询处理。现有被利用来进行查询处理的云平台系统包括 Hadoop[1]、Spark[2]、Trinity[Bin Shao et al, 2013]等等。

因为 Hadoop 是目前最受欢迎的云计算平台，所以很多研究人员在研究如何利用 Hadoop 进行 RDF 数据上的查询处理。在进行数据预处理的时候，现有基于 Hadoop 的 RDF 数据上的分布式查询处理方法将 RDF 数据转化为平面文件存储在 HDFS 上。在进行查询处理的时候，这些方法将查询分解成若干子查询。每个子查询通过在 HDFS 上的扫描得到候选解，然后用 MapReduce 将候选解连接起来以得到最终解。不同方法之间主要区别就是不同的 RDF 数据转化 HDFS 平面文件的方式。

SHARD[Kurt Rohloff et al, 2010]以 RDF 数据中的主体为核心进行数据划分。SHARD 把一个主体相关的所有三元组聚集在一起并存储成 HDFS 文件中的一行。HadoopRDF[Mohammad Farhan Husain et al, 2011]和 P-Partition[Xiaofei Zhang et al, 2012]都是以 RDF 数据中的属性为核心进行存储。它们把有相同属性的所有三元组聚集一起并存储于一个 HDFS 文件中。

除了基于 Hadoop 的方法之外，现阶段还有部分研究工作是基于其他的云平台系统的。比如基于 Trinity[Bin Shao et al, 2013]系统的 Trinity.RDF[Kai Zeng et al, 2013]和 Stylus[Liang He et al, 2017]、基于 Parquet[3]的 Sempala[Alexander Schätzle et al, 2014]、基于 Spark[Matei Zaharia et al, 2010]的 S2RDF[Schtzle A. et al, 2015]、Sparklify[Claus Stadler et al, 2019]、WORQ[Amgad Madkour et al, 2018]、S2X[Alexander Schätzle et al, 2015]。

[1] https://hadoop.apache.org/

[2] https://spark.apache.org/

[3] http://parquet.apache.org/.

Trinity.RDF[Kai Zeng et al, 2013]提出了利用 Trinity[Bin Shao et al, 2013]进行 RDF 数据管理的方法。Trinity 是微软研发的一个基于内存的分布式图数据管理系统。Trinity.RDF 将 RDF 数据图的邻接表载入 Trinity 的内存云中。当用户提交查询之后，Trinity.RDF 依次对查询中每个变量 v 的候选点 u 进行图搜索直到得到解。为了进一步提升基于 Trinity 进行 RDF 数据管理的性能，Stylus[Liang He et al, 2017]针对 RDF 数据图中的同类实体通常有着相近的谓词组合的特点，将每个实体的邻接表按照谓词组织到一个叫 xUDT 的数据结构中。具体而言，每个 xUDT 对应一类实体，这类实体的邻接表里包含同样的谓词集合，而且这些谓词按序组织成一个序列。当利用 xUDT 对每个点（实体）的邻接表进行组织时，每个实体的邻居信息按照 xUDT 的谓词序列顺序排放。

Sempala[Alexander Schätzle et al, 2014]利用一个基于列存储的云存储系统 Parquet 进行 RDF 数据管理。Sempala 将 RDF 数据转化成基于属性的关系数据表存储在 Parquet 上。在查询处理阶段，Sempala 将查询改写成能在 Parquet 上 SQL 语句以执行得到结果。

目前已经有一系列研究工作基于 Spark 开发了分布式 RDF 图数据管理系统，比如 S2RDF[Schtzle A. et al, 2015]、Sparklify[Claus Stadler et al, 2019]、WORQ[Amgad Madkour et al, 2018]、S2X[Alexander Schätzle et al, 2015]。其中，S2RDF[Schtzle A. et al, 2015]、Sparklify[Claus Stadler et al, 2019]、WORQ[Amgad Madkour et al, 2018]都是利用 Spark 的关系数据库接口 Spark SQL 接口来进行 RDF 数据存储。它们将大规模 RDF 数据按照垂直划分的方式进行划分，然后每个垂直表利用 Spark SQL 接口来进行存储。在此基础上，S2RDF 物化了部分垂直划分数据表之间的连接结果并存储在关系数据表中。而 WORQ 主要是用 Bloom Filter 来加速 Semi-join。WORQ 将 Semi-join 过程中传输连接变量投影改成传输连接变量候选值的 Bloom Filter，然后利用 Bloom Filter 过滤一些无用候选值后再连接。此外，S2X[Alexander Schätzle et al, 2015]利用 Spark 的图计算接口 GraphX 来进行 RDF 图数据存储和查询处理。

总的来说，因为基于已有云平台的查询处理方法利用了现有的云计算框架，所以这些方法都有很好的可扩展性与容错性。但是，由于之前云计算框架很多并未针对 RDF 数据管理进行特殊的优化，所以这些方法进行查询处理的效率不高。

2. 基于数据划分的分布式 RDF 数据查询处理方法

基于数据划分的分布式 RDF 数据查询处理方法首先将 RDF 数据划分成若干子数据集，然后将这些子数据集分配到不同计算节点上。各个计算节点安装单机的 RDF 数据管理系统以管理被分配来的子数据集。当查询输入这些系统中后，这些方法首先将查询也划分成若干子查询，然后这些方法将这些子查询分配到各个计算节点上执行得到部分解，最后这些方法

收集所有部分解通过连接得到最终解。不同基于数据划分的查询处理方法的主要区别在于数据划分时采用的策略不一样。现有划分方法按照是否采用查询日志的情况可以分为两类：数据驱动的方法和查询日志的方法。

1）数据驱动的方法

Jiewen Huang 等人提出的方法[Jiewen Huang et al, 2011]使用现有成熟工具METIS[George Karypis & George Karypis, 1998]来对 RDF 数据划分，划分出来每个子图对应一个数据分片，进而对应一个系统中的工作节点。在每个工作节点内部，使用已有的单机 RDF 数据管理系统对数据分片进行管理。SemStore[Buwen Wu et al, 2014]则是提出了一种叫有根子图的特殊结构来作为划分基本单元对 RDF 知识图谱数据进行划分。所谓 RDF 数据图上点 v 的有根子图就是从 v 出发做遍历得到的所有点构成的子图。SemStore 首先找出能覆盖整个 RDF 数据图的一个有根子图集合，然后将这些有根子图聚成若干类。每一个类里面所有的作为有根子图一个分块被分配到一个对应的机器。

华中科技大学袁平鹏老师研究组还提出了一种基于 RDF 数据图上路径的划分方法[Buwen Wu et al, 2015]。这个方法[Buwen Wu et al, 2015]首先在 RDF 数据图上定义出“源点”和“沉入点”，其中源点指 RDF 数据图上没有入度的点，沉入点指 RDF 数据图上没有出度的点。然后在源点和沉入点基础上定义出“末端到末端路径”，即从源点或者图上环中没有进入环的边的点到沉入点或者末端到末端路径已经路过点的路径。该方法首先找出覆盖全图的末端到末端路径集合，然后将覆盖全图的末端到末端路径集合分成 k 份，每份作为一个分块存储到一台机器上。

MPC[Peng Peng et al, 2022]提出一种基于最小谓词割的图数据划分方法。如果在 RDF 图数据划分的时候可以保证一些谓词不可能是跨分片谓词，那么只包含这些非跨分片谓词的查询就是可以不涉及跨计算节点连接。由于真实查询里面三元组模式中谓词常常是常量，判定一个查询是否只包含非跨分片谓词往往可以在查询解析阶段完成。因此，MPC 研究一个可以使得跨分片谓词数量最小化的点划分策略。这样的点划分策略可以使得不涉及跨计算节点连接的查询的数量最大化。

2）查询日志驱动的方法

随着 RDF 数据被越来越多地应用在各个领域，针对这些 RDF 数据集的 SPARQL 查询日志也越来越多地被发布出来。针对上述查询日志，很多研究讨论了如何从这些查询日志中发现频繁的查询模式并基于这些模式进行 RDF 图数据划分。

WARP[Katja Hose et al, 2013]首先利用 METIS[George Karypis & George Karypis, 1998]来对 RDF 数据的划分。然后，WARP 从查询日志中挖掘并选取出频繁被查询的模式并将它们

称为频繁查询模式。之后，WARP找出这些频繁查询模式的匹配，并把这些匹配复制到METIS划分结果中。

北京大学王选计算机研究所数据管理研究室与湖南大学合作提出了一种基于查询日志的分布式RDF数据图查询处理方法[Peng Peng et al, 2016b, Peng Peng et al, 2019a]以提高分布式 SPARQL 查询处理的效率。该方法从查询日志中挖掘并选取出频繁被查询的模式并将它们称为频繁查询模式。基于该方法所选取的频繁查询模式，该方法将 RDF 数据图划分成若干分片。该方法提出了三种数据划分方式：垂直划分、水平划分和混合划分。三种的数据划分方式针对于三种不同的查询处理目标。垂直划分就是将所有满足相同频繁查询模式的结构划分到相同分片以利用查询局部性提高系统整体吞吐量；而水平划分就是通过扩展关系数据库中“小项谓词” 的概念来将满足相同频繁查询模式的结构尽可能划分到不同分片以利用系统并行性提高查询处理效率；混合划分就是将垂直划分和水平划分进行融合。该方法还提出了一个数据分配方法将划分出的分片分配到不同机器上。查询处理阶段，用户输入的SPARQL 查询也基于频繁查询模式被分解成若干子查询。然后，每个子查询根据其所对应的频繁查询模式被传输到不同机器上去执行以找出匹配。所有子查询的匹配最终通过连接操作合并成最终匹配。

总的来说，基于数据划分的 RDF 数据上的分布式查询处理方法要求按照自身的算法设计进行 RDF 数据的划分与分配，以减少查询处理阶段的通信代价。但是，这些方法的系统受制于数据划分方法的性能。

3. 联邦型分布式 RDF 数据查询处理方法

随着 LOD 的发展，现在越来越多的数据发布者都愿意将数据表示成 RDF 数据格式并链接入关联数据上。其中很多数据发布者在将数据表示成 RDF 数据格式之外还提供SPARQL 查询接口来让别人使用它的数据。这些 SPARQL 查询接口都属于“自治”的系统，即能各自独立地接收 SPARQL 查询并计算出匹配。每一个包含一定 RDF 数据和 SPARQL 查询接口的机器被称为一个 RDF 数据源。这些“自治”的 RDF 数据源被集成到一个系统平台下就形成了所谓的联邦型分布式 RDF 数据管理系统。

针对联邦型分布式 RDF 数据管理系统，现阶段也有一些研究在讨论来其上的查询处理技术。在联邦型分布式 RDF 数据管理系统中，因为各个 RDF 数据源之间相互独立地自治，所以系统在查询处理阶段无法中断各个 RDF 数据源的处理进程。因此，在联邦型分布式RDF 数据管理系统中，系统需要提前将 SPARQL 查询分解成若干子查询并传送到它们对应的 RDF 数据源，以让这些对应的 RDF 数据源对子查询独立地进行处理并得到部分解。之后，系统将这些部分解收集起来并通过连接操作得到最终解。在这个过程中，不同方法之间

的主要区别在于如何进行查询分解并确定每个子查询对应的 RDF 数据源。

DARQ[Bastian Quilitz & Ulf Leser, 2008]是最早讨论如何在联邦型分布式 RDF 数据管理系统上的进行 SPARQL 查询处理。当 SPARQL 查询输入以后，DARQ 根据一个叫服务描述的索引进行查询分解并确定出相关的 RDF 数据源。所谓服务描述，其中包含若干所谓的性能值。每个性能值对应一个数据源，其中包含若干元组 t = (p, r)，其中 p 表示该数据源有 p 这个属性，r 对应于当属性为 p 时主体或者客体若干限制。此外，在查询处理过程中，DARQ 还讨论了两个子查询结果连接方式：一是嵌套循环连接，就是一般的自然连接；二是绑定式连接，就是一个子查询先找出解，然后将解传输到另一个子查询那里，然后将解绑定到第二个子查询进行过滤。

在 DARQ 的服务描述之外，还有 SPLENDID[Olaf Görlitz & Steffen Staab, 2011] 、HiBISCuS[Muhammad Saleem et al, 2014]等方法。其中，SPLENDID 根据每个数据源的 VOID 信息建立一个倒排索引。这个索引将每个属性和类型信息映射到一个数据对（d, c），其中 d 表示属性或类型信息所在的 RDF 数据源，c 表示在 d 这个数据源上属性或类型信息的数量。HiBISCuS[Muhammad Saleem et al, 2014]也构建了与 DARQ 类似的索引。只是，在确定各个子查询的相关 RDF 数据源阶段，HiBISCuS 将查询图建模成一个有向带标签的超图，并利用这个有向带标签的超图进一步减少每个子查询的候选 RDF 数据源。

不同于上述利用索引来确定相关 RDF 数据源的方法，FedX[Andreas Schwarte et al, 2011]可以在查询处理阶段实时确定相关数据源。当查询输入以后，FedX[Andreas Schwarte et al, 2011] 首先将查询中每个三元组模式都传到所有 RDF 数据源上并通过 SPARQL 查询语义中的 ASK 来确定相关数据源。之后，以三元组模式为单位进行查询优化，进而将若干三元组模式聚集在一起并得到连接操作顺序。FedX 所使用的连接方式也是和 DARQ 相似的绑定式连接，但是 FedX 在传输中间结果的时候不再是一个一个传，而是若干个中间结果合在一起传。

北京大学王选计算机研究所数据管理研究室与湖南大学合作讨论了在联邦型 RDF 数据库上的多 SPARQL 查询优化问题[Peng Peng et al, 2019a , Peng Peng et al, 2018]。该方法提出了一个基于查询重写的多查询优化技术。这个多查询优化技术利用 OPTIONAL 操作符、UNION 操作符、 FILTER 操作符和 VALUES 操作符将每个 RDF 数据源上所涉及的查询进行重写以降低了查询执行过程中的远程调用次数并提高了系统性能。

此外，北京大学王选计算机研究所数据管理研究室与湖南大学合作针对关联数据上被预先划分好的 RDF 数据还曾提出过一个基于“局部计算-再合并”的分布式 RDF 数据管理方法[Peng Peng et al, 2016a]——gStoreD 。gStoreD 也是不干预 RDF 数据图预先定义的划分，

即假设数据已经被划分并分布在不同的计算节点上。系统中每台机器根据自身上所存储的 RDF 数据计算出整个 SPARQL 查询的局部匹配。所找出的局部匹配被定义为本地局部匹配。然后，所有被找出的本地局部匹配被归并起来并通过连接操作合并成最终匹配。之后，针对本地局部匹配的特点，gStoreD 还研究并定义了本地局部匹配等价类（LEC），并基于本地局部匹配等价类过滤掉若干不必要掉本地局部匹配[Peng Peng el at, 2019b]。

参考文献

[Jens Lehmann et al, 2015] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, Christian Bizer. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6 (2), 167–195 (2015).

[Valerie Bonstrom et al, 2003] Valerie Bonstrom, Annika Hinze, Heinz Schweppe. Storing RDF as a Graph.In Proceedings of LA-WEB'2003. pp.27-36

[Lei Zou et al, 2011] Lei Zou, Jinghui Mo, Lei Chen, M. Tamer Özsu, Dongyan Zhao. gStore: Answering SPARQL Queries via Subgraph Matching. PVLDB 4(8): 482-493 (2011)

[Lei Zou et al, 2014] Lei Zou, M. Tamer Özsu, Lei Chen, Xuchuan Shen, Ruizhe Huang, Dongyan Zhao. gStore: A Graph-based SPARQL Query Engine. VLDB J. 23(4): 565-590 (2014)

[Fabian M. Suchanek et al, 2008] Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum. YAGO: A Large Ontology from Wikipedia and WordNet. J. Web Sem. 6 (3), 203–217 (2008).

[Johannes Hoffart et al, 2013] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Gerhard Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Artif. Intell. 194, 28–61 (2013).

[Farzaneh Mahdisoltani et al, 2015] Farzaneh Mahdisoltani, Joanna Biega, Fabian M. Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. in CIDR (2015).

[Kevin Wilkinson et al, 2003] Kevin Wilkinson, Craig Sayers, Harumi A. Kuno, Dave Reynolds. Efficient RDF Storage and Retrieval in Jena2. SWDB 2003: 131-150

[Kevin Wilkinson, 2006] Kevin Wilkinson. Jena Property Table Implementation. in SSWS, Athens, Georgia, USA (2006), pp. 35–46.

[Eugene Inseok Chong et al, 2005] Eugene Inseok Chong, Souripriya Das, George Eadon, Jagannathan Srinivasan. An Efficient SQL-based RDF Querying Scheme. VLDB 2005: 1216-1227

[Jeen Broekstra et al, 2003] Jeen Broekstra, Arjohn Kampman, Frank van Harmelen. Sesame: An

Architecture for Storing and Querying RDF Data and Schema Information, in Spinning the Semantic Web (2003), pp. 197–222.

[Jeen Broekstra et al, 2002] Jeen Broekstra, Arjohn Kampman, Frank van Harmelen. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. International Semantic Web Conference 2002: 54-68

[Stephen Harris & Nicholas Gibbins, 2003] Stephen Harris, Nicholas Gibbins. 3store: Efficient Bulk RDF Storage. PSSS 2003.

[Jing Lu et al, 2007] Jing Lu, Li Ma, Lei Zhang, Jean-Sébastien Brunner, Chen Wang, Yue Pan, Yong Yu. SOR: A Practical System for Ontology Storage, Reasoning and Search. VLDB 2007: 1402-1405.

[Zhengxiang Pan & Jeff Heflin, 2003] Zhengxiang Pan, Jeff Heflin. DLDB: Extending Relational Databases to Support Semantic Web Queries. In Proceedings of PSSS'2003.

[Daniel J. Abadi et al,2009] Daniel J. Abadi, Adam Marcus, Samuel Madden, Kate Hollenbach. SW-Store: a vertically partitioned DBMS for Semantic Web data management. VLDB J. 18(2): 385-406 (2009)

[Cathrin Weiss et al,2008] Cathrin Weiss, Panagiotis Karras, Abraham Bernstein. Hexastore: sextuple indexing for semantic web data management. PVLDB 1(1): 1008-1019 (2008)

[Thomas Neumann & Gerhard Weikum, 2008] Thomas Neumann, Gerhard Weikum. RDF-3X: A RISC-style Engine for RDF. PVLDB 1(1): 647-659 (2008)

[Thomas Neumann & Gerhard Weikum, 2010a] Thomas Neumann, Gerhard Weikum. The RDF-3X Engine for Scalable Management of RDF Data. VLDB J. 19(1): 91-113 (2010)

[Thomas Neumann & Gerhard Weikum, 2010b] Thomas Neumann, Gerhard Weikum. x-RDF-3X: Fast Querying, High Update Rates, and Consistency for RDF Databases. PVLDB 3(1): 256-263 (2010)

[Marcin Wylot et al, 2011] Marcin Wylot, Jigé Pont, Mariusz Wisniewski, Philippe Cudré-Mauroux. dipLODocus[RDF] - Short and Long-Tail RDF Analytics for Massive Webs of Data. International Semantic Web Conference (1) 2011: 778-793.

[Jinha Kim et al, 2015] Jinha Kim, Hyungyu Shin, Wook-Shin Han, Sungpack Hong, Hassan Chafi. Taming Subgraph Isomorphism for RDF Query Processing. PVLDB 8(11): 1238-1249 (2015).

[Bin Shao et al, 2013] Bin Shao, Haixun Wang, Yatao Li. Trinity: A Distributed Graph Engine on a Memory Cloud. SIGMOD Conference 2013: 505-516.

[Kurt Rohloff et al, 2010] Kurt Rohloff, Richard E. Schantz. High-performance, Massively Scalable Distributed Systems using the MapReduce Software Framework: the SHARD Triple-store. PSI EtA 2010: 4.

[Mohammad Farhan Husain et al, 2011] Mohammad Farhan Husain, James P. McGlothlin, Mohammad M. Masud, Latifur R. Khan, Bhavani M. Thuraisingham. Heuristics-Based Query Processing for Large RDF Graphs Using Cloud Computing. IEEE Trans. Knowl. Data Eng. 23(9): 1312-1327 (2011)

[Xiaofei Zhang et al, 2012] Xiaofei Zhang, Lei Chen, Min Wang. Towards Efficient Join Processing over Large RDF Graph Using MapReduce. SSDBM 2012: 250-259.

[Xiaofei Zhang et al, 2013] Xiaofei Zhang, Lei Chen, Yongxin Tong, Min Wang. EAGRE: Towards Scalable I/O Efficient SPARQL Query Evaluation on the Cloud. ICDE 2013: 565-576.

[George Karypis & George Karypis, 1998] George Karypis, Vipin Kumar. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. SIAM J. Scientific Computing 20(1): 359-392 (1998).

[Kai Zeng et al, 2013] Kai Zeng, Jiacheng Yang, Haixun Wang, Bin Shao, Zhongyuan Wang. A Distributed Graph Engine for Web Scale RDF Data. PVLDB 6(4): 265-276 (2013)

[Alexander Schätzle et al, 2014] Alexander Schätzle, Martin Przyjaciel-Zablocki, Antony Neu, Georg Lausen. Sempala: Interactive SPARQL Query Processing on Hadoop. Semantic Web Conference (1) 2014: 164-179.

[Jiewen Huang et al, 2011] Jiewen Huang, Daniel J. Abadi, Kun Ren. Scalable SPARQL Querying of Large RDF Graphs. PVLDB 4(11): 1123-1134 (2011).

[Buwen Wu et al, 2014] Buwen Wu, Yongluan Zhou, Pingpeng Yuan, Hai Jin, Ling Liu. SemStore: A Semantic-Preserving Distributed RDF Triple Store. CIKM 2014: 509-518

[Buwen Wu et al, 2015] Buwen Wu, Yongluan Zhou, Pingpeng Yuan, Ling Liu, Hai Jin. Scalable SPARQL Querying using Path Partitioning. ICDE 2015: 795-806.

[Bastian Quilitz & Ulf Leser, 2008] Bastian Quilitz, Ulf Leser. Querying Distributed RDF Data Sources with SPARQL. ESWC 2008: 524-538.

[Olaf Görlitz & Steffen Staab, 2011] Olaf Görlitz, Steffen Staab. SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions. COLD 2011.

[Muhammad Saleem et al, 2014] Muhammad Saleem, Axel-Cyrille Ngonga Ngomo. HiBISCuS: Hypergraph-Based Source Selection for SPARQL Endpoint Federation. ESWC 2014: 176-191.

[Andreas Schwarte et al, 2011] Andreas Schwarte, Peter Haase, Katja Hose, Ralf Schenkel, Michael Schmidt. FedX: Optimization Techniques for Federated Query Processing on Linked Data, International Semantic Web Conference 2011: 601-616.

[Peng Peng et al, 2016a] Peng Peng, Lei Zou, M. Tamer Özsu, Lei Chen, Dongyan Zhao. Processing SPARQL queries over distributed RDF graphs. VLDB J. 25(2): 243-268 (2016).

[Chantana Chantrapornchai & Chidchanok Choksuchat, 2018] Chantana Chantrapornchai, Chidchanok Choksuchat. TripleID-Q: RDF Query Processing Framework Using GPU. IEEE Trans. Parallel Distributed Syst. 29(9): 2121-2135 (2018)

[Jiaxin Shi et al, 2016] Jiaxin Shi, Youyang Yao, Rong Chen, Haibo Chen, Feifei Li. Fast and Concurrent RDF Queries with RDMA-Based Distributed Graph Exploration. OSDI 2016: 317-332

[Zihang Yao et al, 2022] Zihang Yao, Rong Chen, Binyu Zang, Haibo Chen. Wukong+G: Fast and Concurrent RDF Query Processing Using RDMA-Assisted GPU Graph Exploration. IEEE Trans. Parallel Distributed Syst. 33(7): 1619-1635 (2022)

[Atserias A et al, 2008] Atserias A., Grohe M., Marx D. Size Bounds and Query Plans for Relational Joins[C]. In Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS’08), Washington, USA, 2008: 739－748.

[Atserias A et al, 2013] Atserias A., Grohe M., Marx D. Size Bounds and Query Plans for Relational Joins[J]. SIAM Journal on Computing, 2013, 42(4): 1737-1767.

[Ngo H. Q. et al, 2014] Ngo H. Q., Christopher R., Rudra A. Skew Strikes Back: New Developments in the Theory of Join Algorithms[J]. ACM SIGMOD Record, 2014, 42(4): 5-16.

[Ngo H. Q. et al, 2018] Ngo H. Q., Porat E., Christopher R., Rudra A. Worst-case Optimal Join Algorithms[J]. Journal of the ACM, 2018, 65(3): 1-40.

[Aidan Hogan et al, 2019] Aidan Hogan, Cristian Riveros, Carlos Rojas, Adrián Soto. A Worst-Case Optimal Join Algorithm for SPARQL. ISWC (1) 2019: 258-275

[Schtzle A. et al, 2015] Schtzle A., Przyjaciel-Zablocki M., Skilevic S., Lausen G. S2RDF: RDF Querying with SPARQL on Spark[J]. Proceedings of the VLDB Endowment, 2015, 9(10): 804-815.

[Liang He et al, 2017] Liang He, Bin Shao, Yatao Li, Huanhuan Xia, Yanghua Xiao, Enhong Chen, Liang Chen. Stylus: A Strongly-Typed Store for Serving Massive RDF Data. Proc. VLDB Endow. 11(2): 203-216 (2017)

[Alexander Schätzle et al, 2015] Alexander Schätzle, Martin Przyjaciel-Zablocki, Thorsten Berberich, Georg Lausen. S2X: Graph-Parallel Querying of RDF with GraphX. Big-

O(Q)/DMAH@VLDB 2015: 155-168

[Peng Peng et al, 2022] Peng Peng, Lei Zou, M. Tamer Özsu, Cen Yan, Chengjun Liu. Minimum Property-Cut RDF Graph Partitioning. ICDE 2022: 192-204.

[Katja Hose et al, 2013] Katja Hose, Ralf Schenkel. WARP: Workload-aware replication and partitioning for RDF. ICDE Workshops 2013: 1-6

[Peng Peng et al, 2019a] Peng Peng, Lei Zou, Lei Chen, Dongyan Zhao. Adaptive Distributed RDF Graph Fragmentation and Allocation based on Query Workload. IEEE Trans. Knowl. Data Eng. 31(4): 670-685 (2019)

[Peng Peng et al, 2016b] Peng Peng, Lei Zou, Lei Chen, Dongyan Zhao. Query Workload-based RDF Graph Fragmentation and Allocation. EDBT 2016: 377-388

[Peng Peng et al, 2018] Peng Peng, Lei Zou, M. Tamer Özsu, Dongyan Zhao. Multi-query Optimization in Federated RDF Systems. DASFAA (1) 2018: 745-765

[Peng Peng et al, 2021] Peng Peng, Qi Ge, Lei Zou, M. Tamer Özsu, Zhiwei Xu, Dongyan Zhao. Optimizing Multi-Query Evaluation in Federated RDF Systems. IEEE Trans. Knowl. Data Eng. 33(4): 1692-1707 (2021)

[Claus Stadler et al, 2019] Claus Stadler, Gezim Sejdiu, Damien Graux, Jens Lehmann. Sparklify: A Scalable Software Component for Efficient Evaluation of SPARQL Queries over Distributed RDF Datasets. ISWC (2) 2019: 293-308

[Amgad Madkour et al, 2018] Amgad Madkour, Ahmed M. Aly, Walid G. Aref. WORQ: Workload-Driven RDF Query Processing. ISWC (1) 2018: 583-599

[Peng Peng el at, 2019b] Peng Peng, Lei Zou, Runyu Guan. Accelerating Partial Evaluation in Distributed SPARQL Query Evaluation. ICDE 2019: 112-123

第九章 通用和领域知识资源

王昊奋[1]，曹征晖[2]，林俊宇[3]

1. 同济大学 设计创意学院，上海 200092

2. 复旦大学 计算机科学技术学院，上海 200438

3. 中国科学院 信息工程研究所，北京 100093

一、通用知识资源与领域知识资源

1. 定义与概述

在知识图谱领域，**知识资源**包括现有完整的知识图谱以及前文所涉及的知识建模、知识获取、知识融合、知识存储、知识计算五大知识图谱生命周期[王 & 胡, 2017]中使用的数据集、评测基准、模型、工具和平台等。**根据知识资源本身的特性与其适用的应用场景可将知识资源分为通用知识资源和领域知识资源**。

通用知识资源以通用（领域无关）知识图谱为主要载体，同时还包括基于通用任务的数据集、模型与工具平台等。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大量的常识性知识，覆盖面极广[Aminer.Org, 2019]。知识通常以静态的、客观的、明确的三元组[肖, 2019]的形式表示，在知识建模、获取以及融合过程中涉及大量的互联网开放数据，数据经过处理和筛选，形成可供知识图谱使用的信息资源，而数据处理的过程也被整理和抽象为图谱构建过程中的实用工具与平台。

领域知识资源则包括特定领域知识图谱和针对特定领域任务的数据集、评测基准、模型等。领域知识图谱又叫行业知识图谱或垂直知识图谱，通常面向某一特定领域，可看成是一个“基于语义技术的行业知识库”。领域知识图谱通常有着严格而丰富的数据模式，对该领域知识的深度、知识准确性有着更高的要求，包含静态知识和动态知识两类知识。

2. 区别与联系

现实世界的知识丰富多样且极其庞杂，不同类型的知识资源存在广泛的区别和联系。由于知识资源中数据集、模型、工具等均是围绕知识图谱的构建和应用，故通用与领域知识资源的区别与联系主要集中在知识图谱的区别与联系之中。从知识图谱全生命周期的具体子过程来看，通用知识图谱和领域知识图谱存在以下区别：

从**知识建模**角度来看，通用知识图谱中知识的主要表示形式是多个互相关联的事实型知识三元组。领域知识图谱则需要对专家经验知识、行业文本的语义信息进行表示，通常额外

包含较为复杂的本体工程和规则型知识等内容。

从**知识抽取**角度来看，通用知识图谱注重知识的广度，覆盖粗粒度的知识。其在实体抽取层面，关注更多的实体，准确度不高；在关系抽取层面，多采用面向开放域的关系抽取。领域知识图谱注重知识的深度，覆盖细粒度的知识。其在实体抽取层面，关注具有特定行业意义的领域数据，准确度高；在关系抽取层面，多采用预定义关系抽取。

从**知识融合**角度来看，由于通用知识图谱对知识抽取的质量有一定容忍度，因此需要通过知识融合来提升数据质量。领域知识图谱从领域内部的结构化数据、半结构化数据、非结构化数据中抽取知识，并且有一定的人工审核校验机制来保证质量，需要通过融合多源的领域知识来扩大数据层的规模。

从**知识计算**角度来看，由于通用知识图谱的知识覆盖范围较宽，深度较浅，从而导致图谱上的推理路径相对较短。而领域知识图谱的知识相对密集，这就导致图谱上的推理路径相对较长。当然，也存在一些特殊情况，例如 DBpedia 具有丰富的推理规则，推理路径比某些只有少量推理规则的领域知识图谱长[Hang et al., 2021]。

从**知识应用**角度来看，通用知识图谱主要应用在信息搜索和自动问答方面。领域知识图谱的主要应用除了上述方面，还包括决策分析、业务管理等。

综上所述，可以看到在**图谱的构建方式**上，通用知识图谱主要强调知识的广度，因此运用百科数据自底向上地进行构建，通过对互联网开放的数据进行关系抽取等步骤的处理，逐步扩大数据规模。领域知识图谱由于面向不同的领域，其数据模式各不相同，应用需求也有所差异，因此没有一套通用的标准和规范来指导构建，需要基于特定行业通过工程师和业务专家的不断交互定制实现。而在实际的工程实践中，往往需要利用通用知识图谱的广度结合领域知识图谱的深度，合理地优化平衡，使两者相互支撑，才能够形成更加完善的知识图谱。

3. 问题与挑战

随着近年来知识图谱技术的蓬勃发展与越来越多新需求的提出，大量新的知识图谱、数据集、模型等知识资源应运而生，也给知识图谱领域的发展带来了新的任务和挑战。如何有的放矢地利用现有知识资源，融合多模态、事件等知识，为具体需求构建高质量知识图谱，实现深度知识推理，以及提高大规模知识图谱计算效率，是当前所面临的挑战。

二、通用知识资源

国内外多个研究机构建立了一些大型通用知识图谱，根据所描述知识对象、内部知识单元的不同可以将其分为传统的语言类知识库、常识类知识库、世界百科类知识库以及当前发展迅速的多模态类知识库和事理类知识库等类型。

1. 传统的通用知识资源

在 2012 年知识图谱概念正式被 Google [Singhal, 2012] 提出之前，就已经涌现了一批高质量的知识资源。本文将按照语言类知识资源、世界百科类知识资源、常识类知识资源以及针对中文的开放知识资源四大类对传统的通用知识资源进行介绍。

1）语言知识资源

语言知识是使用人类语言应当具备的词法、句法、语义或语用等方面的知识，语言类知识库包括作为知识工程开发的 WordNet 词汇知识库、BabelNet 多语言词汇库以及 HowNet 义原知识库等。WordNet [Miller, 1995]主要定义了名词、动词、形容词以及副词之间的语义关系，如典型的名词之间的上下位关系；当前，WordNet 3.0 已经包含了超过 15 万个词以及 20 万个词之间的语义关系，成为了目前处理英文最多的一类词典知识库，主要用于词义消歧等任务。BabelNet [Navigli & Ponzetto, 2012]是一个多语词汇语义网络和本体，与 WordNet 类似，并在此基础上引入了“跨语言”和“百科”的双重特点，成为了当前规模最大的一种多语言百科词典知识库；最新的 Babelnet 5.0 [Navigli et al., 2021] 中包括 500 种语言，2,000 万个同义词词组，以及 3.8 亿个与 Wikipedia 的链接关系，总计超过 19 亿个 RDF 三元组 [Babelnet.Org, 2013]。另外，国外研究者主导构建的语言类知识资源还有英语单词同义词与反义词词库 Roget [Roget, 2020]、框架语义网 FrameNet [Baker et al., 1998]、语义隐喻库 MetaNet [Dodge et al., 2015]、动词词库 VerbNet [Schuler, 2005]等等。由中国学者构建的“知网”HowNet [Dong & Dong, 2003]是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的大型语言知识库。通过多年的发展逐渐构建出一套以义原为核心的知识体系，基于 HowNet 的义原语言知识，清华大学发布了开源版本 OpenHowNet [Qi et al., 2019]首次将 HowNet 的核心数据开源，并且开发了丰富的调用接口，实现义原查询、基于义原的词相似度计算等功能，近年来也发表了知识表示、知识获取、知识计算等一系列算法、工具，在国内外具有较大的影响力。然而当前针对中文的语言知识库或语言知识图谱还是相对较少，除了 HowNet，还未有大型的语言知识库出现。

2）世界百科知识资源

世界百科知识是指现实世界中各实体间关系的事实知识，对于世界百科类知识资源，国外的 DBpedia [Auer et al., 2007]最早于 2007 年被提出，是一种基于本体映射的知识图谱，为近 10 年来知识图谱研究领域的核心数据集，目前社区知识共享依然充满活力[王 et al., 2022; Lehmann et al., 2015]；DBpedia 使用固定的模式从维基百科中抽取信息实体，当前拥有 127 种语言的超过 2.28 亿实体以及数十亿个 RDF 三元组，涵盖地理信息、人、基因、药物、图

书、科技出版社等多个领域，用户可方便地下载其核心数据集与本体模式[Dbpedia.Org, 2014]，并基于此获取定制知识库以支持增强搜索、推荐和智能问答等应用；Yago [Suchanek et al., 2007]是与 DBpedia 同时期的另一个大型百科类知识图谱，它整合 Wikipedia 与 WordNet 的大规模本体[Rebele et al., 2016]，如 Schema.org [Guha et al., 2016]，拥有关于人、城市、国家、电影和组织的一般性知识；最新的 Yago 4 [Pellissier Tanon et al., 2020]拥有约 5,000 万个实体，20 亿个事实，并且该知识库也是完全开源的。Wikidata [Vrandečić & Krötzsch, 2014]是一个可以众包协作编辑的多语言百科知识库。截至 2022 年，Wikidata 能够支持近 350 种语言、9,000 万个实体及一亿两千万条声明[Wikidata.Org, 2013]，数据规模仍在持续增长中，支持数据集的完全下载。国内的 Zhishi.me [Niu et al., 2011]从开放的百科数据中抽取结构化数据，当前已融合了包括百度百科、互动百科、中文维基三大百科的数据，拥有 1,000 万个实体数据、一亿两千万个 RDF 三元组。以通用百科为主线，结合垂直领域的 CN-DBPedia [Xu et al., 2017]，则从百科类网站的纯文本页面中提取信息，经过滤、融合、推断等操作后形成高质量的结构化数据，目前的 CN-DBpedia 2 [Xu et al., 2019]包含超过 900 万的百科实体以及 6,700 万的三元组关系，在问答机器人、智能玩具、智慧医疗、智慧软件等领域均有应用；XLore [Wang et al., 2013]则是基于中文维基百科、英文维基百科、百度百科、互动百科构建的大规模中英文知识平衡知识图谱，其包括 200 万的概念，2,600 万的实体以及 50 万的关系，并应用于知识图谱自动化构建平台“柏拉图”[Xlore.Cn]、知识图谱表示工具 OpenKE [Han et al., 2018]中。

3）常识知识资源

常识知识泛指普通人应当具备的基本知识信息，常识知识的缺失被认为是制约未来人工智能发展的重要瓶颈问题[唐, 2021] ，在过去的几十年中，国内外研究人员设计和构建了许多包含常识知识的资源。传统的常识知识库以 Cyc、ConceptNet、NELL 为典型代表。Cyc [Lenat, 1995]的主要特点是基于形式化语言表示方法来刻画知识，支持复杂推理，但是也导致扩展性和灵活性不够。Cyc 提供开放版本 OpenCyc，最新的 Cyc 知识库包含有 50 万条术语和 700 万条断言组成。ConceptNet [Liu & Singh, 2004]是一个利用互联网众包、专家创建和游戏三种方法构建地常识知识图谱，其本质为一个描述人类常识的大型语义网络，ConceptNet 5 [Speer et al., 2017]版本已经包含有 2,800 万关系描述。与 Cyc 相比，ConceptNet 采用了非形式化、更加接近自然语言的描述，而不是像 Cyc 那样采用形式化的谓词逻辑。ConceptNet 完全免费开放，并支持多种语言。国内也有学者整理了 ConcptNet 5 中包含的 60 多万条中文数据[Openkg.Org, 2017]。NELL（Never-Ending Language Learner）[Carlson et al., 2010]是卡内基梅隆大学基于互联网数据抽取而开发的三元组知识库。它的基本理念是给定

少量初始样本（少量概念、实体类型、关系），利用机器学习方法自动从互联网学习和抽取新的知识，目前 NELL 已经抽取了 400 多万条高置信度的三元组知识。除上述典型的常识知识库外，还有包括：ATOMIC [Sap et al., 2019]、GLUCOSE [Mostafazadeh et al., 2020]、WebChild [Tandon et al., 2017]、Quasimodo [Romero et al., 2019]、SenticNet [Cambria et al., 2020]、HasPartKB [Bhakthavatsalam et al., 2020]、Probase [Wu et al., 2012]、IsaCore [Lee et al., 2017] 等常识类知识图谱资源。

4）中文开放知识资源

OpenKG [Openkg.Org]是一个面向中文域开放的知识图谱社区项目，其主要目的是促进以中文为核心的知识图谱数据的开放、互联与众包，以及知识图谱工具、算法和平台的开源开放与互联。

OpenKG 聚集了大量开放的中文知识图谱数据、工具及文献。除了前文提到的百科类 Zhishi.me、CN-DBpedia、XLore，还有很多优秀的中文开放知识资源。例如，北京大学中文百科知识图谱 PKU-PIE [Openkg.Org, 2016]，PKU-PIE 从维基百科、DBpedia、百度百科等多个来源自动收集知识形成的知识库，有自己的类别体系和谓词体系，并且和 DBpedia 等常见的数据库进行关联；浙江大学大规模细粒度中文概念图谱 OpenConcepts [Openkg.Org, 2019]，OpenConcepts 拥有 440 万概念核心实体，以及 5 万概念和 1,200 万实体-概念三元组，包括了常见的人物、地点等通用实体。目前数据还在不断更新中，OpenConcepts 实现了为智能推荐、智能问答、人机对话等应用提供数据支持。

OpenKG 对这些主要开放开源数据进行了链接计算和融合工作，并通过 OpenKG 提供开放的 Dump 或开放访问 API，完成的链接数据库也免费向公众开放。此外，OpenKG 还对一些重要的知识图谱开源工具进行了收集和整理，包括知识建模工具 Protégé、知识融合工具 Limes、知识问答工具 YodaQA、知识抽取工具 DeepDive 等。

2. 多模态知识资源

在大数据环境和新基建背景下，数据对象日益丰富，交互方式也在发生变化。然而，大多数传统知识图谱是基于纯符号表示知识，这类早期的大规模通用知识图谱仅仅停留在对文本实体表示的层面，这无疑限制了机器理解现实世界的能力[Zhu et al., 2022]。多模态知识图谱凭借其非结构化多模态的数据组织形式，很好的满足了对异构型数据建模的新需求，目前较为有代表性的多模态知识资源包括 IMGpedia、MMKG、OpenRichpedia 等。

IMGpedia [Ferrada et al., 2017]是一个 2017 年 5 月向公众发布的关联数据集，它整合了来自维基共享资源数据集的图像的视觉信息，汇集了 1,500 万张图像的视觉内容的描述符，这些图像之间的 4.5 亿个视觉相似性关系，来自 DBpedia Commons 的图像元数据链接，以

及与单个图像相关的 DBpedia 资源链接。IMGpedia 的具体用例很多：例如，可以直接查询两张图像的视觉相似关系，根据最近邻计算找到颜色、边缘或者强度相似的图像；可以使用 SPARQL 查询执行图像的视觉语义检索，通过链接到 DBpedia 将图像的视觉相似性与语义元数据结合起来，为了获得更准确的结果，可以使用 SPARQL 属性路径来包括层次分类。同时，IMGpedia 还链接了 DBpedia 和 DBpedia Commons，通过语义信息来提供语义上下文和进一步的元数据。

MMKG [Liu et al., 2019]主要用于联合不同知识图谱中的不同实体和图像执行关系推理，MMKG 是一个包含所有实体的数字特征和（链接到）图像的三个知识图谱的集合，以及对知识图谱之间的实体对齐，MMKG 的提出有利于多关系链接预测和实体匹配的发展。MMKG 选择在知识图谱补全相关工作中广泛使用的数据集 FREEBASE-15K（FB15K）作为创建多模态知识图谱的起点。MMKG 同时也创建了基于 DBpedia 和 YAGO 的版本，称为 DBpedia-15K（DB15K）和 YAGO15K，通过将 FB15K 中的实体与其他知识图谱中的实体对齐。其中基于 DBpedia 的版本主要构建了 sameAs 关系，为了创建 DB15K，提取了 FB15K 和 DBpedia 实体之间的对齐，通过 sameAs 关系链接 FB15K 和 DBpedia 中的对齐实体；构建关系图谱，来自 FB15K 的很大比例的实体可以与 DBpedia 中的实体对齐。MMKG 有潜力促进知识图谱的新型多模态学习方法的发展。

上述的两个多模态知识图谱仍存在一些问题。例如，在 IMGPedia 中关系类型稀疏，关系数量少，图像分类不清晰等，在 MMKG 中图像并没有作为单独的图像实体存在，而是依赖于相应的传统文本实体。这些问题限制了它们在多模态任务中的应用。维基数据和 DBpedia 等大规模知识库是语义搜索和问答的强大资产，大多数知识图谱的构建工作都集中在组织和发现结构化的文本知识上，而很少关注网络上大量的非文本资源，从而损失了大量有价值的信息。为了改善这种情况，东南大学、同济大学等研究人员提出多模态图谱 Richpedia [Wang et al., 2020]，旨在通过在 Wikidata 的文本实体中分配足够多的、多样化的图像来提供一个全面的多模式知识图谱。并发布了首个多模态开放知识图谱 OpenRichpedia [郑 et al., 2021] 从构建过程的数据采集部分看，OpenRichpedia 分析发现文本知识图谱实体对应的图像资源存在大量长尾分布问题，在 Wikipedia 中每一个文本实体只有很少的视觉信息。因此团队研究人员基于现有的传统文本实体，从 Wikipedia、Google、Bing 和 Yahoo 四大搜索引擎中获取文本实体相应的图像，每张图像都作为知识图谱中的一个实体存储于 OpenRichpedia 中。由于 Wikidata 已经为每个文本知识图谱实体定义了唯一的统一资源标识符，因此同时也将这些统一资源标识符添加到 OpenRichpedia 作为文本知识图谱的实体。对于图像实体，可以直观地从 Wikipedia 上收集图像，然后在 OpenRichpedia 中创建相应的统

一资源标识符。对于图谱构建过程中比较关键的关系抽取部分，研究团队主要利用基于规则的关系抽取模板，借助 Wikipedia 图像描述中的超链接信息生成图像的多模态语义关系。

多模态实体链接是多模态数据处理的基础任务之一，旨在将多模态数据中的实体链接到知识图谱中，在多模态数据理解、多模态知识图谱、多模态问答中具有广泛应用意义。东南大学团队发布的多模态实体链接数据集 MELBench [汪 et al., 2021]包含 3 个任务：Weibo-MEL、Wikidata-MEL 和 Richpedia-MEL 数据集，数据源分别包含来自社交媒体、百科知识和多模态知识图谱等领域，分别包含 25,602、18,880 和 17,806 条多模态实体链接数据，每条数据均为人工标注，包含与目标实体相关的文本信息和视觉信息。该数据集能够为多模态实体链接（MEL）任务提供基准数据支持。

3. 事件知识资源

自 2017 年首次被提出[刘 , 2017]，事理图谱的研究与应用在近年来已经有了一定的发展，主要体现在事理图谱基本轮廓的确定与传播、领域事理图谱雏形 Demo 的研制与应用探索、事理图谱在领域的复制与延伸三个方面。

事件的演变和发展有其自身的基本原则，这些原则使事件按顺序发生。发现事件中的这种演化规律对事件预测、决策和发展系统的情景设计具有重要价值。传统的知识图谱主要关注实体和它们之间的关系，忽略了现实世界中的事件逻辑联系和时间上的先后顺序。因此，哈尔滨工业大学刘挺教授团队提出一种新型的知识库——事件逻辑图（Event Logic Graph，ELG）[刘 & 薛, 2018; Ding et al., 2019]，它可以揭示现实世界事件的演化规律和发展逻辑。具体来说，ELG 是一个有向循环图，其节点是事件，边代表事件之间的顺序、因果、条件或超名词（is-a）关系。目前该团队构建了两个 ELG：金融 ELG，由 150 多万个事件节点和 180 多万条有向边组成；旅游 ELG，由大约 3 万个事件节点和 234 万条有向边组成。实验结果表明，ELG 对脚本事件预测的任务是有效的。事理图谱技术对机器人传动系统中高效、有序的故障诊断具有重要的指导意义，也具有实际应用价值。Jianfeng Deng 以机器人传动系统的历史维修日志为研究对象，提出了一种自顶向下的故障诊断事件逻辑知识图谱构建方法[Deng et al., 2022]，实验表明，该方法可以提高事件论证实体和关系联合提取的效果，最后构建了机器人传动系统故障诊断的事件逻辑知识图，为机器人传动系统的自主故障诊断提供决策支持。中文突发事件语料库是由上海大学（语义智能实验室）所构建。根据国务院颁布的《国家突发公共事件总体应急预案》的分类体系，从互联网上收集了 5 类（地震、火灾、交通事故、恐怖袭击和食物中毒）突发事件的新闻报道作为生语料，对其进行文本预处理、文本分析、事件标注以及一致性检查等处理，最后将标注结果保存到语料库中，CEC 合计 332 篇，以支持事件抽取任务。大规模实时（事件逻辑与概念）事理知识库“学迹”

[Datahorizon.Cn, 2020]包括事件概念抽取、事件因果逻辑抽取、事件数据关联推荐与推理。截至目前，“学迹”已经积累事件概念描述三元组 500 余万，因果事件三元组两千余万，概念上下位三元组一百余万。“学迹”为三元组提供了一个搜索入口，围绕事件，提供事件的前序原因、后续结果，事件的关联概念，事件关联产业链的搜索。

三、领域知识资源

和强调知识广度的通用知识图谱不同，领域知识图谱面向特定领域，突出专业深度，通过数据和知识双重驱动，从问题建模分析为出发点，对特定行业场景的多源异构信息进行整合和分析。行业数字化的推进促使越来越多的行业知识图谱落地。但行业数据来源相对狭窄，且部分脱离互联网，主要依赖专业领域的书籍、专利、模型和专家经验，不同领域的构建方案与应用形式也有所不同，面临知识稀疏、知识分散以及构建成本较高的挑战。因此如何利用现有成熟的知识图谱以及高质量开源的数据集、模型、工具和平台等，助力垂直行业落地和应用知识图谱技术也是研究者们所重点关注的。本文将以特殊的社会公共卫生事件新冠疫情相关领域、以及传统的金融、电商、医疗这四个领域为代表介绍领域知识资源的发展情况。

1. 新冠图谱

自 2019 年新型冠状病毒（COVID-19，下文简称新冠）爆发以来，人们的生活乃至生命健康都受到了极大的影响。随着近年来关于新冠疫情的不断发展，有关疫情的各类信息也在不断更新。为更好地助力抗“疫”，分析疫情发展始末，OpenKG 紧随疫情发展，联合国内数所顶级高校和科研机构、企业，构建出针对多种应用场景的知识图谱资源。

新冠百科图谱[Openkg.Org, 2020a]以病毒、疾病、细菌为主体，以百度百科、中文维基百科、互动百科等作为主要数据源，并扩展治疗，疾病等相关内容，构建出实例规模 5,000 余个，三元组 35,000 余组，总计包含 250 余个类别的知识图谱，可用于面向新冠相关术语的语义检索、智能问答，以及相关文档的智能搜索和推荐。新冠科研图谱[Openkg.Org, 2020b]搜集、整理并关联与新冠病毒相关的科研基础数据和科研文献，以 NCBI（美国国家生物技术信息中心网站）中的 Taxonomy 板块为基础构建知识图谱，实例超 20 万，三元组数量接近 200 万的规模，可用于病毒的生物学分类、病毒变异性、潜在治疗药物，病毒传播途径与种类等的预测任务，具有很强的潜在应用价值。目前仍在不断进行数据补充和完善，扩展出病毒分类图谱，新冠基本信息图谱，抗病毒药物图谱，病毒亲缘关系图谱等具体细分图谱。新冠临床图谱[Openkg.Org, 2020c]从目前已有的规范文件入手，基于诊疗规范、研究进展、发病统计，以新冠肺炎为核心延展至病毒、治疗方案、症状、方剂等各类概念，形成新冠临床知识图谱，实现可基于该图谱进行知识问答的应用落地。新冠英雄图谱[Openkg.Org, 2020d]

包括了医疗专家组、因公殉职英雄、武汉当地和全国各地的意见领袖等，涉及生平事迹和基本属性，并与新冠百科、新冠科研、新冠临床等图谱中的一些概念或实体关联。图谱以新冠病毒专家为核心延展至履历、成果、事件、战役等各类概念，形成新冠英雄知识图谱，可基于图谱进行英雄人物动态展示。除此之外，OpenKG 还开源了新冠热点事件图谱[Openkg.Org, 2020e]，新冠健康图谱，新冠物资图谱，新冠流行病学图谱等。上述提及的新冠领域知识图谱的快速构建和上线，极大程度便利了学界与一线人员的抗疫工作，助力社会早日战胜疫情。

2. 金融领域

金融领域积累了丰富的数据，且行业直接需求与知识图谱的核心价值十分契合，是数据智能最早落地并产生价值的行业，在目前行业知识图谱市场的份额占比也是最大的。金融领域的知识资源主要包括企业图谱、专利图谱、产业链图谱等，通过构建图谱以实现多源异构数据的知识整合。

文因互联用深度语义分析技术，将非结构化金融文档转为知识图谱，并基于推理机和知识库管理系统技术，实现大规模金融知识建模和流程机器人，在上交所、北交所、投行、评级、资管多个场景成功落地。明略科技的金融知识图谱以某行近十年的全量数据构建了包括“企业、个人、机构、账户、交易以及行为数据”在内，规模达十亿节点百亿边的知识图谱数据库，通过知识图谱平台建设来帮助该银行风控体系建立了完整的客户关系网及资金流转全貌，支持该行非现场审计、系统性风险管控、精准营销等多项应用。

在 OpenKG 中也开源了数种不同的金融领域知识资源，例如：DuEE-fin [Baidu.Com, 2022]是百度发布的金融领域篇章级事件抽取数据集，包含 13 个事件类型的 1.17 万个篇章，同时存在部分非目标篇章作为负样例。事件类型来源于常见的金融事件，数据集中的篇章来自金融领域的新闻和公告，覆盖了真实应用场景中诸多难以解决的问题。大规模金融研报知识图谱大规模金融研报知识图谱数据集 FR2KG 包含 10 个实体类型，19 个关系类型和 6 种属性，总计 17,799 个实体，26,798 对关系三元组以及 1,328 个属性。金融时序超图（Finanical Temporal Hypergraph Ontology，FTHO）[Ontoweb.Wust.Edu.Cn, 2021]面对金融领域多元关系表示的困境和时序事件表示需求，结合超图概念和事件 5W（When，Where，Why，What，Who）定义，构建了可通用化的金融时序超图模型。除此之外，还有创新投资领域知识图谱[宗, 2021]、基金知识图谱[Openkg.Org, 2021]等。

3. 电商餐娱领域

电商领域的知识图谱，需要以用户需求为核心，建立起商品，用户，购买意愿之间的联系。商品信息天生就拥有知识卡片，但知识生产源头较多，且涉及到较深的商品知识，所以

需要用更标准化的逻辑语言去描述。商品知识图谱对到底要构建哪些知识，缺乏清晰定义，所以在构建电商知识图谱时，一定要明确知识的交付终态。

阿里巴巴的电商知识图谱——新零售电商认知图谱不仅包含了以商品为中心的知识图谱（Product Graph），还包含了以用户需求的显式节点为中心的概念图谱（Concept Net）。形成了以概念、商品、标准产品、标准品牌等为核心，利用实体识别、实体链指和语义分析技术，整合关联了例如舆情、百科、国家行业标准等 9 大类一级本体，包含了百亿级别的三元组，以人、货、场为核心形成了巨大的知识网。此外阿里巴巴另一个开放数字商业知识图谱 AliOpenKG [Alibaba.Com, 2021]包含了超过 18 亿的三元组，多达 67 万的核心概念，2,681 类关系，目前还在持续维护与扩展。基于图谱中的商业要素知识，有利于深度理解零售数据，有利于数智驱动商品运营、商家成长，优化市场供需匹配，产生更多贴近场景需求的智能应用。

美团 NLP 中心构建的大规模餐饮娱乐知识图谱——美团大脑，则是一个场景知识图谱。大众点评作为中国最大的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，连接了数亿用户和数千万商户，积累了宝贵的业务数据，蕴含着丰富的日常生活相关知识。美团依托这一平台，充分挖掘关联各个场景数据，抽取用户评论数据，理解用户在菜品、价格、服务、环境等方面的喜好，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，挖掘出人、店、商品、标签之间的知识关联。美团大脑知识图谱目前有数十类概念，数十亿实体和数千亿三元组。由于美团跨场景推荐的业务较多，所以知识图谱的可解释性极大地助力了这项关键业务。

4. 医疗领域

目前，越来越多的医疗领域知识图谱数据集开源，从细分内容来看，现有的医学数据集可分为电子病历数据集、中医疗法剂方数据集、专科疾病数据集和常见疾病数据集等。

高效规范地书写病历，一直是医务工作者的痛点。Yidu-S4K [Yiducloud.Com, 2019a]和 Yidu-N7K [Yiducloud.Com, 2019b]是由医渡云提出的标准化电子病历数据集，两个数据集根据真实病例分布再经过医学人工编辑而成，前者面向中文电子病历的命名实体识别任务，包含医疗命名实体识别和医疗实体及属性抽取两个子任务，后者针对临床术语标准化子任务，是语义相似度匹配任务的一种。目前国内最大规模的电子病历知识图谱由之江实验室提供，覆盖了 18 大类医学标准术语集、包含 479 万医学概念实例、3,531 万概念相互关系以及 9,600 万篇文献知识关联，临床术语覆盖范围达到国际领先水平。此外，为了推动 CNER 系统在中文临床文本上的表现，CCKS 从 2017 年开始，每年都组织了面向中文电子病历的命名实体识别评测任务，并推出了相应数据集[Ccks, 2017; Ccks, 2018; Ccks, 2020]。

由中医科学院中医药信息研究所搭建的中医药知识服务平台在中医理论的指导下，从中医古籍文献、病案中提取经典名方及其治法，系统收集中医理论和方法，集成八大知识库，包含中医药领域的信息标准、领域本体、术语系统、文献库、知识库等多种知识资源，内容涉及中药、方剂、针灸、临床、养生等领域，提供知识检索、知识问答、知识浏览、知识推荐等多种服务。例如，中成药知识图谱（TCMLS）是一个包含 10 余万个中医概念以及 100 余万个语义关系的大型语义网络，基本覆盖了中医药学科的概念体系，在完整性方面处于中医界的领先地位。它以中成药领域海量文献为基础，构建了以中成药应用为主题的大规模知识库，建立了以病、症、证为核心的囊括组成、适应症、禁忌属性等属性的中成药知识图谱，设计了面向中成药推荐的全局最优图谱路径算法，研发了中成药知识问答系统，同时嵌入中医临床辅助系统进行探索应用，有助于实现中成药准确性、有效性、经济性、安全性的应用目标。

疾病数据集的典型代表有 DiaKG（糖尿病知识图谱数据集）、Disease KG（常见疾病信息知识图谱）等。DiaKG 由两位经验丰富的内分泌专家设计标注指南，侧重“实体”和“关系”，定义了 18 类实体类型和 15 类医学关系，从 41 篇公开发表的糖尿病指南中收集了共计 22,050 个实体和 6,890 个关系，涵盖了近年来糖尿病垂直领域的热点研究内容。DiseaseKG 从“寻医问药”医疗网站上爬取原始数据，对爬取的数据进行预处理后筛选适合做知识存储的相关信息，共定义了 8 类实体（4.4 万实体量级），7 类疾病属性和 11 种关系（31 万关系量级），覆盖了常见的疾病。

目前依托医疗知识图谱和其他ＡＩ技术建立的医疗互联网产品包括百度的“灵医智惠”，阿里巴巴的“医知鹿”、“DoctorYou”，腾讯的觅影，中国平安的“平安好医生”和丁香园的丁香医生等。以中国平安为例，平安智慧医疗推出的中文医疗指示图谱集成了 60 万医学概念，530 万医学关系，千万医学数据，覆盖了医学核心概念。基于此医疗知识图谱，平安医疗提供多个智能服务场景，包括疾病预测、智能预诊/分诊、智能影响筛查、智能随访追踪、智能质量控制等。

四、工具与平台

1. 工具

在知识资源的构建和应用过程中，实用的工具包括：知识建模工具、知识获取工具、知识融合工具、知识图谱存储工具、知识推理工具、图挖掘分析工具、语义搜索和智能问答工具。下文将针对每种工具类型进行简单介绍。

知识建模工具，Protégé 和 NeOn Toolkit。Protégé 是一个本体编辑器，其基于 RDF（S）、

OWL 等语义网规范提供 PC 图形化界面和在线 Web 版本——WebProtégé，通常适用于原型场景构建。NeOn Toolkit 是一个适用于本体工程生命周期的工具，其以 Eclipse 插件的方式为用户提供服务。

知识获取工具，从结构化数据中获取知识的目标通常是把关系数据库中的数据转换成 RDF 形式的知识，W3C 为此制定了从关系数据库映射到 RDF 数据集的标准语言 R2RML。典型的开源工具有 D2RMAP 和 D2RQ。D2RQ [Bizer & Seaborne, 2004]是一个将关系数据库转换为虚拟的 RDF 数据库的平台，主要包含 D2R Server [Bizer & Cyganiak, 2006]，D2RQ Engine 和 D2RQ Mapping Language 3 个组件。从半结构化数据中获取知识通常是指使用包装器的方法从网页数据中获取知识，如 Lixtio [Baumgartner et al., 2001]提供了一种用户可视化配置的方式进行半自动化生成网页包装器的工具，WIE 是一个通过网页自动分析从而辅助生成包装器的工具，适用于抽取目标数据中的表格信息。Deepdive 与 Snorkel 提供了一套面向特定关系的、基于远程监督学习的抽取框架，使用现有知识库和规则定义来自动生成语料，框架自动完成模型的训练过程，并使用机器学习算法来减少各种形式的噪音和不确定性，用户可以使用简单的规则来影响（反馈）学习过程以提升结果的质量。DeepKE 是浙江大学开发的基于深度学习方法的开源中文关系抽取工具，使用了包括卷积神经网络、循环神经网络、注意力机制网络、图卷积神经网络、胶囊神经网络以及语言预训练模型等在内的多种深度学习算法，但该工具同样仅用于关系的抽取。上述工具主要针对关系的抽取，未提供针对概念实体、事件等知识的抽取功能。

知识融合工具 DBpedia Mapping Tool 是一个用于把从 Wikipedia 中抽取的信息通过映射融到 DBpedia 中的工具，其以可视化的方式让用户进行 DBpedia 中本体（类、实体、数据类型等）和信息模块的映射。Knowledge vault [Dong et al., 2014]是谷歌推出的一个互联网规模的知识库，它融合了海量的从互联网中基于先验知识库抽取的信息，并通过监督学习的方法对这些知识进行融合。这些融合工具通常是针对特定场景设计的，通用性和可配置程度比较低，难以实现复杂多变场景下的知识整合。

知识图谱中最主要的数据结构为基于图的结构，因此，**知识图谱存储方式**也即图结构数据的存储主要有 RDF 存储和图数据库两种方式。Neo4j 是第一代图数据库的代表，它使用了原生图存储结构，但不使用 schema（即 schema free）是一种自由的图数据管理方式，同时它还支持 ACID 事务的处理，并提供 Cypher 查询语言。Janus Graph 是在 Titan 的基础上发展起来的第二代图数据库的代表，设计原理是在现有的成熟存储（如 NOSQL）上实现对图的存储逻辑，底层存储的分布式能力使其天然具备分布式能力。在数据量大规模增长与实时查询分析要求不断提高的背景下，基于原生、并行图设计的图数据库逐渐成为新兴发展方向，

也被称为第三代图数据库。其中的代表产品为商业数据库 Tiger Graph 与 Plantgraph，它们能够有效地支持 OLTP 和 OLAP 等多种应用场景，解决大规模图数据场景下的多步连接问题。目前，第三代图数据库还只在一些拥有大数据量与高性能要求的商业场景下得到使用，尚未有开源的产品出现。

RDFox 是一个**本体知识推理工具**，支持共享内存并行 OWL2RL 推理。RDFox 支持 Java、Python 多语言 APIs 访问，还支持一种简单的脚本语言与系统的命令行交互，但 RDFOX 完全基于内存对硬件要求较高，在超大规模数据场景下难以使用。Dros 是一个使用 Java 语言开发的基于 RETE 算法（一种前向推理算法）业务规则推理弓擎，其使用“If-Then" 句式和事实的定义，使引擎的使用非常直观，同时还支持将 Java 代码直接嵌入到规则文件中。Link prediction Tool 是一个在大规模网络中自动发现缺失的链接的工具，主要用于社交网络中的链接预测。SNAP（Stanford Network Analysis Platform）是斯坦福大学研发的个通用高性能大规模网络分析与操作平台，能够高效地实现大规模网络中的链接预测。

上文提到的多数图数据相关工具只支持 OLTP 模式的图查询功能以及一些简单的图算法，对于大规模的图挖掘分析支持较少。基于图数据库实现图挖掘分析的模式需要集成第三方的图挖掘分析工具，如 Spark graphx、Graphlab 和 Giraph 等。最常用的为 Spark graphx，它是在实时计算引擎 Spark 上为图计算设计与实现的一套计算框架，方便用户通过统一的模式进行图算法编程，但由于其基于通用的计算框架来实现图计算，因此性能较图分析的专用系统要低。Plato 是腾讯开源的一个支持十亿级别节点的超大规模图计算框架，其基于自适应图计算引擎，能够根据不同类型的图算法，提供自适应计算模式、共享内存计算模式和流水线计算模式等多种计算模式。但它是一个重量级的图计算框架，集成成本相对较高，并且开发者需要基于其独特的底层 AP 编程，定制化开发成本也较高。Euler 是阿里开源的大规模分布式图表示学习框架，内置 Deep Walk、Node2Vec 等业界常见的图嵌入算法。

知识链接是支持语义搜索的重要方法，**知识实体链接工具**有 Wikipedia miner 和 DBPedia Spotlight 等。这些早期的工具通常是以开放的知识图谱（Wikipedia、DBpedia 等）为知识链接的目标知识库使用字符串匹配、向量相似度等算法进行计算；当前，基于深度学习、知识图谱表示学习的方法已经成为知识链接的最新发展方向智能问答方向知名的开源工具有 Active QA 和 gAnswer 等。Active QA 是谷歌开源的一款使用强化学习来训练 AI 智能体进行问答的研究项目，在强化学习框架的推动下，智能体逐步学会提出更具针对性的具体问题并理解、问答问题，从而得到所寻求的结果。gAnswer 是一个基于知识图谱的自然语言问答系统，能够将自然语言问题转化成包含语义信息的查询图，并将查询图转化成标准的 SPARQL 查询，将这些查询在图数据库中执行，最终得到用户的答案上述问答工具只适用于特定的场

景（如 gAnswer 用于 KBQA），而在复杂企业级的场景中通常需要支持所有类型的问答任务。

2. 平台

在现阶段的工业级应用场景中，国内外越来越多的企业和研究机构开始引入平台化方案，即围绕生命周期构建相应的行业知识图谱服务平台，在平台的基础上进行应用的构建，实现一个功能完整的知识图谱信息系统来支撑知识图谱的应用落地。

知识图谱服务平台主要负责构建知识图谱和提供具体场景应用服务，将来自上游数据提供方的初步结构化数据进行信息抽取、知识融合、知识加工，逐步构建起知识图谱，再为下游最终用户提供具体场景下基于知识图谱的数据智能化应用服务，可显著提高各行业中知识图谱的落地效率和效果，应用领域包括金融、客服、工业、科研、医疗等。

目前，国外主流的知识图谱平台有：Palantir 可拓展大数据分析平台、IBM Watson Discovery 服务及其相关产品所使用的知识图谱框架 Knowledge Studio、Oracle 知识图谱平台、Metaphactory 知识图谱信息系统解决方案平台，以及开源知识图谱项目 LOD2。

表 1 国外知识图谱平台

公司	平台名称	平台简介	主要特点（服务）
Palantir	Palantir 平台	Palantir 是用于知识图谱创建、管理、搜索、发现、挖掘和积累的可扩展的大数据分析平台	数据集成、搜索发现、知识管理、算法引擎、算法引擎
IBM	IBM Watson Discovery 知识图谱框架	Watson Discovery Services 使用该框架并提供相关服务，这些服务已经部署在 IBM 以外的许多行业配置中。	该框架直接支持 Watson Discovery，它关注于使用结构化和非结构化的知识来发现新的、不明显的信息，以及发现之上的相关垂直产品；该框架允许其他人以预先构建的知识图谱为核心构建自己的知识图谱。
Oracle	Oracle 知识图谱平台	Oracle知识图谱平台基于其自身多年的存储经验,在具有明显优势的存储层上进行构建,上层通过W3C标准的RDF和OWL组织和表示图谱,使用SPARQL对数据统一查询服务。	对数据存储与访问的支持性比较好,可以实现基于内存的并行图计算,提供许多工具完成从各种大数据平台、关系数据库到知识图谱的映射与转换。
Metaphacts	Metaphactory 平台	Metaphactory 提供了一套从知识存储、知识管理到知识查询与应用开发的端到端的知识图谱平台解决方案。	Metaphactory 主要针对结构化数据进行查询和管理，且兼容常见的知识图谱存储形式，实现不同数据源、不同格式的知识图谱混合查询，提供了搜索、可视化和知识编辑管理的接口,可用于知识图谱资产管理，快速应用程序构建和面向最终用户的交互。
Stardog	Stardog 平台	Stardog是一个企业级知识图谱平台,通过将数据转换成知识,使用知识图谱进行组织,对外提供查询、检索和分析等服务。	Stardog能够把关系数据库映射成虚拟图,并且支持OWL2的推理和 Gremlin,但其仅对结构化数据（ RDBMS Excel等)的处理,没有针对非结构化数据的知识抽取,也不具有知识融合功能。
/	LOD2 开源知识图谱项目	LOD2 是构建结构化链接数据的企业级管理工具和方法	提供一个搜索、浏览和生成链接数据的平台,其侧重于链接数据的生命周期管理,而对于其他类型的数据需要首先转换成链接数据。

同时，传统解决方案商旗下知识平台和初创型知识服务平台以其在具体领域中的垂直深耕，并整合了知识图谱的设计、构建、编辑、管理、应用等全生命周期实现，在市场上也具有一定的竞争力。这类典型的知识平台有：明略知识图谱信息系统 SCOPA，其提供了基于知识图谱技术的知识管理和洞察分析平台，实现从客观数据汇聚到抽象知识沉淀的认知跃迁，为组织提供知识驱动的辅助决策；柯基数据的认知智能引擎提供全周期的知识图谱构建和运维管理平台，平台通过动态本体实现多源异构数据的知识获取与融合存储，可构建复杂的多模态知识图谱，提供从基础数据到知识管理、知识应用的全方位智能服务，赋能医药、军工、能源、金融等行业的数智化转型。PlantData 知识图谱管理系统（Knowledge Graph

Management System，KGMS），以行业知识图谱全生命周期为理论指导，结合多行业、数十个项目实战经验，打造全流程一体化的管理平台。星环科技的知识图谱全场景解决方案，内置全套数据组件，使用 3D 空间图实现知识图谱的可视化，并提供了成熟的行业模板；渊亭 DataExa-Sati 认知智能平台，可帮助客户打造行业知识图谱，帮助企业快速生成成熟的解决方案；此外还有包括达观数据、东软、北大医信、鼎富科技等等一批知识图谱平台提供商。企业级的知识图谱信息系统、知识工作自动化平台、知识图谱平台软件服务等方案相继被各厂商提出，正快速成为以知识图谱为核心的新一代信息系统发展的有力支撑。

表 2 国内知识图谱平台

公司	平台名称	平台简介	主要特点（服务）
百度	知识图谱开放平台	基于知识图谱、自然语言、搜索与推荐等核心技术，依托高效生产、灵活组织、便捷获取的智能应用知识的全链条能力，提供企业知识应用全生命周期一站式解决方案，助力企业提升效率、提高决策智能水平	数据引入、服务接入、知识生产与组织、平台化综合管理、知识搜索
腾讯	腾讯知识图谱（Tencent Knowledge Graph，TKG） 腾讯知识图谱一站式平台	腾讯知识图谱是一个集成图数据库、图计算引擎和图可视化分析的一站式平台。腾讯知识图谱用于构建和分析包含千亿级节点关系的知识图谱，并支持在图谱上搭建企业级应用服务	知识图谱自动构建、图谱在线查询、提供多种图计算模型、图数据可视化展现、图查询语言、独立部署
阿里巴巴	藏经阁 阿里巴巴知识图谱服务平台	以多源大规模数据为对象，研究从大数据向通用、领域知识转化的共性关键技术，研发并推出知识建模、知识获取、知识融合、知识推理计算和知识赋能的平台服务	通过实现知识建模、知识获取、知识融合、知识推理计算和知识赋能五个模块，提供从数据、信息、知识到知识服务一整套技术平台化服务，同时，特定领域知识图谱可插拔，特定领域知识图谱加载后，可以提供特定领域的知识服务
华为	华为云 知识图谱 KG	华为知识图谱是一款知识图谱构建工具，提供一站式知识图谱构建平台，提供本体设计、信息抽取、知识映射、多源融合以及增量更新等功能	本体设计，信息抽取，知识映射，知识融合，知识服务（知识图谱问答、智能文案系统、行业知识图谱解决方案、智能知识推荐）
明略科技	知识图谱信息系统 SCOPA	可视化数据分析平台，构建在明略自研知识图谱数据库 NEST 之上，实现知识图谱行业解决方案快速落。目前已应用到公共安全、金融、税务、工业等多个行业几百个项目中	关系网络分析、时空轨迹碰撞、实时多维检索、信息比对碰撞、智能协作系统、实时数据接入
柯基数据	KGDATA 知识图谱平台	柯基数据知识图谱平台通过动态本体实现多源异构数据的知识获取与融合存储，可构建复杂的多模态知识图谱，提供从基础数据到知识管理、知识应用全方位智能服务，已赋能医药、军工、能源、金融等行业多客户多业务部门的数智化转型	多模态知识图谱、动态本体构建、非结构化数据标注与训练、结构化数据增量更新、事件抽取、语义检索、智能问答、智能推荐
海义知科技	PlantData 知识图谱 认知智能中台	KGMS：企业级知识图谱管理平台； KGBuilder：配置式自动化图谱构建工具； KGAssist：插件式知识服务助手； KGRobot：会话式图谱机器人开放平台； KGPro：统一知识图谱分析引擎；	关联分析、路径分析、图数据探索、图谱可视化、推理、自然语言检索、智能BI、语义标注
达观数据	达观智能知识图谱平台	基于客户的多源异构数据整合构建知识中台，为客户量身打造基于知识图谱的数据智能化应用，为制造、政务等行业客户提供业务场景智能升级服务	文本挖掘、智能推荐、垂直搜索、文档智能审阅、企业级搜索引擎、客户意见洞察、光学字符识别、机器人流程自动化、数据挖掘分析、文本审核
渊亭科技	DataExa-Sati 认知智能平台	渊亭Dataexa-Sai认知智能平台能够帮助客户打造行业知识图谱,采用分布式服务架构和自研分布式图计算引擎,实现行业级知识图谱构建和分析,从可视化知识建模、多源异构知识提取和知识融合、万亿级别高性能图存储计算引擎、复杂知识推理等角度,快速、精准地从知识图谱中提取出有价值的信息,帮助企业快速生成成熟的解决方案	聚焦金融、政务、国防、工业互联网四大行业，为客户提供认知中台、AI中台、数据中台三大中台产品及AI+行业解决方案，打通“数据-AI-认知”的闭环服务。
海致星图	海致星图金融知识图谱平台	海致星图金融知识图谱平台从零散数据中发现知识，帮助组织机构实现业务智能化	银行智能营运分析：自动化分析财务报表、外源文档、行内文档，提高银行运营决策、产品设计、营销推广、风险管理效率。
星环科技	知识图谱平台 Sophon KG	星环知识图谱平台（Sophon KG）是一款集知识的获取、融合、存储、计算以及应用为一体的自 研知识图谱产品。支持拖拽式图谱构建、知识抽取、知识存储、分布式图谱计算、知识推理以及图谱查询分析	零代码图谱构建、交互式图谱分析、文本标注、图算法数据挖掘、智能语义检索

五、应用

知识图谱于 2012 年被谷歌正式提出的初衷是为了改善搜索，基于谷歌知识图谱的搜索不是简单地返回网页的超链接，而是真正理解用户请求并将其链接到现实世界认知概念的指代，然后返回指代的相关结果，可大幅度提升用户的搜索体验。截至目前，谷歌的知识图谱涵盖了广泛的主题，包括超过 10 亿个实体和 700 亿条事实。与之同时期的，微软必应（Bing）知识图谱也针对搜索场景，它包含了物理世界的知识，如人物、地点、事物、组织、位置等类型的实体，以及用户可能采用的行为。当用户输入搜索文本时，如果知识图谱中存在相关的知识时，必应搜索引擎将显示来自必应知识图谱的知识面板，可充分展示用户感兴趣的内容。领英图谱（LinkedIn graph）也是微软公司旗下的知识图谱应用，其中包括人员、工作、技能、公司、位置等实体，可实现更加有效的职场社交。脸书（Facebook）公司拥有全球最大的社交知识图谱，该图谱以用户为中心，同时包括用户关心的其他信息如兴趣爱好、从事行业等信息，基于图谱的知识资源可增加用户对脸书产品的体验，包括内容搜索和兴趣推荐等。在搜索及社交应用场景中，国内与国外相同，有相应的大型互联网厂商提出的知识图谱，例如百度、搜狗的面向搜索的知识图谱，以及面向社交场景的微博图谱。

经过近 10 年的发展，当前知识图谱的应用俨然远超其最初的搜索场景，由相对通用的搜索、问答、推荐等场景向核心业务决策过程转变。在行业应用方面，随着面向不同行业的知识图谱落地应用，以信息系统为载体的知识图谱典型应用（包括智能问答、推荐系统、个人助手等）也逐渐走进各个行业领域。

知识图谱在国外有着较为成熟的行业应用积累，如 IBM Watson 最早被研发应用于医疗领域，随着产品的不断延伸也逐步应用于金融等其他领域中。Palantir 相关产品已经分别应用于国防安全与金融领域，形成包括反欺诈、网络安全、国防安全、危机应对，保险分析、疾病控制、智能化决策等解决方案。国内人工智能及知识图谱在产业中地落地也呈现井喷得态势，知识图谱在国内的行业应用落地已经处于世界领先水平，在金融、情报分析、能源电力、医疗、工业、教育、政务、公安、营销和客服等场景均得到了广泛应用。

表 3 知识图谱行业应用

		知识图谱平台	知识图谱应用											
			通用	领域										
				公安领域	金融领域	能源领域	客服领域	医疗领域	教育领域	司法领域	营销领域	舆情领域	政务领域	工业领域
大数据智能公司	明略科技	√		√	√	√					√	√	√	√
	国双	√		√		√				√	√	√	√	√
	海致	√		√	√						√			
	百分点	√		√	√							√	√	√
	一览群智	√		√	√							√		
	海乂知科技	√				√								√
	柯基数据	√			√	√	√	√	√	√			√	√
	蓝凌	√												
	文因互联	√			√									
互联网公司	阿里巴巴	√	√	√	√	√	√				√	√		√
	腾讯	√	√	√	√	√	√	√			√	√		
	百度	√	√		√	√	√	√						√
	京东数科	√					√				√		√	
	Google	√	√			√	√		√		√	√		
	amazon	√	√				√			√		√		
	美团	√	√		√		√				√	√		
	今日头条	√	√											
	搜狗	√	√											
AI 公司	科大讯飞	√						√	√	√				
	第四范式	√												
	松鼠AI	√							√					
	追一科技	√												

六、总结与展望

知识对于大数据智能具有重要意义，通用知识资源和领域知识资源已经在多个场景展现了其研究和应用价值，但是知识图谱的作用主要还是体现在增强搜索、推荐和智能问答的效果。另外，大规模知识图谱在深度问答（特别是基于语义分析和推理的问答系统）、演化分析、对话理解等方面的应用还处于初级阶段。如何快速构建高质量知识图谱，融合多模态、事件等知识，利用知识图谱为主的知识资源实现深度知识推理，以及提高大规模知识图谱计算效率，是当前知识图谱发展所面临的挑战。借助知识图谱强大知识库的深度知识推理能力和逐步扩展的认知能力，相关行业从业者将能够对特定的问题进行分析、推理、获得决策支持，从而在各行各业中解放生产力，助力业务转型。

参考文献

[王 & 胡, 2017] 王 昊奋, 胡 芳槐. 行业知识图谱构建与应用. CCKS, 2017

[Aminer.Org, 2019] AMiner.org. Research Report of Knowledge Graph. (2019)[2022-05-06]. https://www.aminer.cn/research_report/5c3d5a8709e961951592a49d

[肖, 2019] 肖 仰华. 知识图谱：概念与技术. 第一版. 电子工业出版社, 2019

[Hang et al., 2021] Hang Ting-Ting, Feng Jun, Lu Jia-Min. Knowledge Graph Construction Techniques: Taxonomy, Survey and Future Directions. Computer Science, 48(2):175-189, 2021

[Singhal, 2012] Introducing the knowledge graph: Things, not strings. (2012-03-16)[2022-04-30]. https://blog.google/products/search/introducing-knowledge-graph-things-not/

[Miller, 1995] Miller George A. WordNet: a lexical database for English. Communications of the ACM, 38(11):39-41, 1995

[Navigli & Ponzetto, 2012] Navigli Roberto, Ponzetto Simone Paolo. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial intelligence, 193:217-250, 2012

[Navigli et al., 2021] Navigli Roberto, Bevilacqua Michele, Conia Simone, et al. Ten Years of BabelNet: A Survey. IJCAI, 4559-4567, 2021

[Babelnet.Org, 2013] BabelNet Statistics. (2013-01)[2022-04]. https://babelnet.org/statistics

[Roget, 2020] Roget Peter Mark. Roget's thesaurus. Good Press, 2020

[Baker et al., 1998] Baker Collin F, Fillmore Charles J, Lowe John B. The berkeley framenet project. COLING, 1998

[Dodge et al., 2015] Dodge Ellen K, Hong Jisup, Stickles Elise. MetaNet: Deep semantic automatic metaphor analysis. Workshop on Metaphor in NLP, 40-49, 2015

[Schuler, 2005] Schuler Karin Kipper. VerbNet: A broad-coverage, comprehensive verb lexicon. University of Pennsylvania, 2005

[Dong & Dong, 2003] Dong Zhendong, Dong Qiang. HowNet-a hybrid language and knowledge resource. NLPKE, 820-824, 2003

[Qi et al., 2019] Qi Fanchao, Yang Chenghao, Liu Zhiyuan, et al. Openhownet: An open sememe-based lexical knowledge base. arXiv, 1901.09957, 2019

[Auer et al., 2007] Auer Sören, Bizer Christian, Kobilarov Georgi, et al.: Dbpedia: A nucleus for a web of open data. ISWC, 722-735, 2007

[王 et al., 2022] 王 萌, 王 昊奋, 李 博涵, et al. 新一代知识图谱关键技术综述. 计算机研究与发展, 2022

[Lehmann et al., 2015] Lehmann Jens, Isele Robert, Jakob Max, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic web, 6(2):167-195, 2015

[Dbpedia.Org, 2014] DBpedia Association.DBpedia. (2014)[2022-04]. https://www.dbpedia.org/

[Suchanek et al., 2007] Suchanek Fabian M, Kasneci Gjergji, Weikum Gerhard. Yago: a core of semantic knowledge. WWW, 697-706, 2007

[Rebele et al., 2016] Rebele Thomas, Suchanek Fabian, Hoffart Johannes, et al. YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. ISWC, 177-185, 2016

[Guha et al., 2016] Guha Ramanathan V, Brickley Dan, Macbeth Steve. Schema. org: evolution of

structured data on the web. Communications of the ACM, 59(2):44-51, 2016

[Pellissier Tanon et al., 2020] Pellissier Tanon Thomas, Weikum Gerhard, Suchanek Fabian. Yago 4: A reason-able knowledge base. ESWC, 583-596, 2020

[Vrandečić & Krötzsch, 2014] Vrandečić Denny, Krötzsch Markus. Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10):78-85, 2014

[Wikidata.Org, 2013] wikidata.org.Wikidata stats. (2013)[2022-04]. https://wikidata-todo.toolforge.org/stats.php

[Niu et al., 2011] Niu Xing, Sun Xinruo, Wang Haofen, et al. Zhishi. me-weaving chinese linking open data. ISWC, 205-220, 2011

[Xu et al., 2017] Xu Bo, Xu Yong, Liang Jiaqing, et al. CN-DBpedia: A never-ending Chinese knowledge extraction system. IEA/AIE, 428-438, 2017

[Xu et al., 2019] Xu Bo, Liang Jiaqing, Xie Chenhao, et al. CN-DBpedia2: An Extraction and Verification Framework for Enriching Chinese Encyclopedia Knowledge Base. Data Intelligence, 1(3):271-288, 2019

[Wang et al., 2013] Wang Zhigang, Li Juanzi, Wang Zhichun, et al. XLore: A Large-scale English-Chinese Bilingual Knowledge Graph. ISWC (Posters & Demos), 121-124, 2013

[Xlore.Cn] xlore.cn.bolatu. [2022-04-30]. https://bolatu.xlore.cn/Index

[Han et al., 2018] Han Xu, Cao Shulin, Lv Xin, et al. Openke: An open toolkit for knowledge embedding. EMNLP, 139-144, 2018

[唐, 2021] 唐 杰. 认知图谱——人工智能的下一个瑰宝. MEET 2021 智能未来大会, 2021

[Lenat, 1995] Lenat Douglas B. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11):33-38, 1995

[Liu & Singh, 2004] Liu Hugo, Singh Push. ConceptNet—a practical commonsense reasoning tool-kit. BT technology journal, 22(4):211-226, 2004

[Speer et al., 2017] Speer Robyn, Chin Joshua, Havasi Catherine. Conceptnet 5.5: An open multilingual graph of general knowledge. AAAI, 4444-4451, 2017

[Openkg.Org, 2017] ConceptNet5 中 文 数 据 集 . (2017)[2022-04]. http://openkg.cn/dataset/conceptnet5-chinese

[Carlson et al., 2010] Carlson Andrew, Betteridge Justin, Kisiel Bryan, et al. Toward an architecture for never-ending language learning. AAAI, 2010

[Sap et al., 2019] Sap Maarten, Le Bras Ronan, Allaway Emily, et al. Atomic: An atlas of machine

commonsense for if-then reasoning. AAAI, 3027-3035, 2019

[Mostafazadeh et al., 2020] Mostafazadeh Nasrin, Kalyanpur Aditya, Moon Lori, et al. Glucose: Generalized and contextualized story explanations. arXiv preprint arXiv:2009.07758, 2020

[Tandon et al., 2017] Tandon Niket, De Melo Gerard, Weikum Gerhard. Webchild 2.0: Fine-grained commonsense knowledge distillation. ACL, 115-120, 2017

[Romero et al., 2019] Romero Julien, Razniewski Simon, Pal Koninika, et al. Commonsense properties from query logs and question answering forums. CIKM, 1411-1420, 2019

[Cambria et al., 2020] Cambria Erik, Li Yang, Xing Frank Z, et al. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. CIKM, 105-114, 2020

[Bhakthavatsalam et al., 2020] Bhakthavatsalam Sumithra, Richardson Kyle, Tandon Niket, et al. Do dogs have whiskers? a new knowledge base of haspart relations. arXiv preprint arXiv:2006.07510, 2020

[Wu et al., 2012] Wu Wentao, Li Hongsong, Wang Haixun, et al. Probase: A probabilistic taxonomy for text understanding. SIGMOD, 481-492, 2012

[Lee et al., 2017] Lee Wooseok, Sunwoo Dam, Emmons Christopher D, et al. Exploring heterogeneous-isa core architectures for high-performance and energy-efficient mobile socs. GLSVLSI, 419-422, 2017

[Openkg.Org] http://www.openkg.cn/

[Openkg.Org, 2016] OpenKG.org.北京大学中文百科知识图谱-PKU-PIE 知识库. (2016-10-10)[2022-04-30]. http://openkg.cn/dataset/pku-pie

[Openkg.Org, 2019] Open Concepts. (2019)[2022-04-30]. http://openconcepts.openkg.cn/

[Zhu et al., 2022] Zhu Xiangru, Li Zhixu, Wang Xiaodan, et al. Multi-Modal Knowledge Graph Construction and Application: A Survey. arXiv preprint arXiv:2202.05786, 2022

[Ferrada et al., 2017] Ferrada Sebastián, Bustos Benjamin, Hogan Aidan. IMGpedia: a linked dataset with content-based analysis of Wikimedia images. ISWC, 84-93, 2017

[Liu et al., 2019] Liu Ye, Li Hui, Garcia-Duran Alberto, et al. MMKG: multi-modal knowledge graphs. ESWC, 459-474, 2019

[Wang et al., 2020] Wang Meng, Wang Haofen, Qi Guilin, et al. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. Big Data Research, 22:100159, 2020

[郑 et al., 2021] 郑 秋硕, 郑 健雄, 漆 桂林, 王 萌.多模态知识图谱 OpenRichpedia. (2021)[2022-04-30]. http://www.richpedia.cn/

[汪 et al., 2021] 汪 鹏, 周 星辰, 邓 臻凯, 李 国正, 谢 佳锋, 吴 江恒.多模态实体链接数据集 MELBench. (2021)[2022-04-30]. http://www.openkg.cn/dataset/melbench

[刘, 2017] 刘 挺. 从知识图谱到事理图谱. CNCC 2017 中国计算机大会, 2017

[刘 & 薛, 2018] 刘 焕勇, 薛 云志.事理图谱，下一代知识图谱. (2018)[2022-04-30]

[Ding et al., 2019] Ding Xiao, Li Zhongyang, Liu Ting, et al. ELG: an event logic graph. arXiv preprint arXiv:1907.08015, 2019

[Deng et al., 2022] Deng Jianfeng, Wang Tao, Wang Zhuowei, et al. Research on Event Logic Knowledge Graph Construction Method of Robot Transmission System Fault Diagnosis. IEEE Access, 10:17656-17673, 2022

[Datahorizon.Cn, 2020] OpenKG.org.学迹:大规模实时(事件逻辑与概念)事理知识库. (2020-03-23)[2022-04-30]. http://openkg.cn/dataset/event-concept-graph-xueji

[Openkg.Org, 2020a] 新冠开放知识图谱.百科. (2020-02-10)[2022-04-30]. http://www.openkg.cn/dataset/covid-19-baike

[Openkg.Org, 2020b] 新冠开放知识图谱.科研. (2020)[2022-04-30]. http://www.openkg.cn/dataset/2019-ncov-research

[Openkg.Org, 2020c] 新冠开放知识图谱.临床. (2020)[2022-04-30]. http://www.openkg.cn/dataset/2019-ncov-clinic

[Openkg.Org, 2020d] 新冠开放知识图谱.英雄. (2020)[2022-04-30]. http://www.openkg.cn/dataset/2019-ncov-hero

[Openkg.Org, 2020e] 新冠开放知识图谱.事件. (2020)[2022-04-30]. http://openkg.cn/dataset/covid-19-event

[Baidu.Com, 2022] Baidu.com.Baidu. DuEE-fin 金融领域篇章级事件抽取数据集. (2022-03-01)[2022-04-30]. http://www.openkg.cn/dataset/duee-fin

[Ontoweb.Wust.Edu.Cn, 2021] 金融时序超图. (2021)[2022-04-30]. http://www.openkg.cn/dataset/ftho

[宗, 2021] 创新投资领域知识图谱. (2021-01-30)[2022-04-30]. http://www.openkg.cn/dataset/invest-on-invent

[Openkg.Org, 2021] 基金知识图谱. (2021)[2022-04-30]. http://www.openkg.cn/dataset/fundkg

[Alibaba.Com, 2021] Alibaba.com.开放的数字商业知识图谱. (2021-12-22)[2022-04-30]. http://www.openkg.cn/dataset/aliopenkg

[Yiducloud.Com, 2019a] OpenKG.org.Yidu-S4K：医渡云结构化 4K 数据集. (2020-11-09)[2022-

04-30]. http://openkg.cn/dataset/yidu-s4k

[Yiducloud.Com, 2019b] OpenKG.org.Yidu-N7K：医渡云标准化 7K 数据集. (2020-11-09)[2022-04-30]. http://openkg.cn/dataset/yidu-n7k

[Ccks, 2017] CCKS 2017 评测二. https://www.biendata.xyz/competition/CCKS2017_2

[Ccks, 2018] CCKS 2018 面向中文电子病历的命名实体识别. https://www.biendata.xyz/competition/CCKS2018_1

[Ccks, 2020] CCKS 2020 中文医学文本命名实体识别. https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/2020_ccks_ner

[Bizer & Seaborne, 2004] Bizer Christian, Seaborne Andy. D2RQ-treating non-RDF databases as virtual RDF graphs. ISWC (Posters), 2004

[Bizer & Cyganiak, 2006] Bizer Christian, Cyganiak Richard. D2r server-publishing relational databases on the semantic web. ISWC (Posters), 2006

[Baumgartner et al., 2001] Baumgartner Robert, Flesca Sergio, Gottlob Georg. Visual web information extraction with lixto. Semannot, 2001

[Dong et al., 2014] Dong Xin, Gabrilovich Evgeniy, Heitz Geremy, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. KDD, 601-610, 2014

第十章 知识图谱质量评估与管理

李直旭[1]， 王萌[2]， 漆桂林[2]， 阮彤[3]

1. 复旦大学 计算机科学技术学院，上海 200438

2. 东南大学 计算机科学与工程学院，江苏 南京 215556

3. 华东理工大学 信息学院，上海 200237

一、任务背景与定义

1. 任务背景

在当前的大数据时代背景下，各领域的数据量与知识量呈爆炸式增长，知识图谱的构建都在追求尽量自动化。因此,**对知识图谱的质量评估与管理成为知识图谱构建与应用中必不可少的一环**[肖仰华 et al, 2020]。知识图谱的质量评估与管理工作是知识图谱应用落地效果的重要保障。在知识图谱的构建层面，高质量的知识图谱是图谱构建的目标；在知识图谱的应用层面，知识图谱的质量也将直接影响其在具体应用场景中的效用[肖仰华 et al, 2020]。

2. 任务定义

知识图谱中的质量的考察对象主要是概念、实体、属性这三类个体知识对象，以及概念之间的关系、概念与实体之间的关系、实体之间的关系等三类关系知识对象。为了对知识图谱进行质量评估与管理，本着“尽早发现，尽早维护，尽早解决”的原则，我们需要在图谱构建的每个阶段做出努力[肖仰华 et al, 2020]。

具体来说，按照[肖仰华 et al, 2020]中的划分方法，知识图谱的质量管理任务可以分为知识图谱构建前、中、后三个阶段实施。如图 1 所示：构建前的质量管理主要在于对数据来源的质量管理，即对于获取知识的数据源头做质量评估与管理；构建中的质量管理主要是知识获取手段和知识融合阶段的质量管理；构建后的质量管理指的是在知识图谱完成初步构建后，对知识图谱的质量进行进一步的完善与常规维护。例如，补全缺失的知识，发现并纠正错误知识，发现并更新过期知识等。

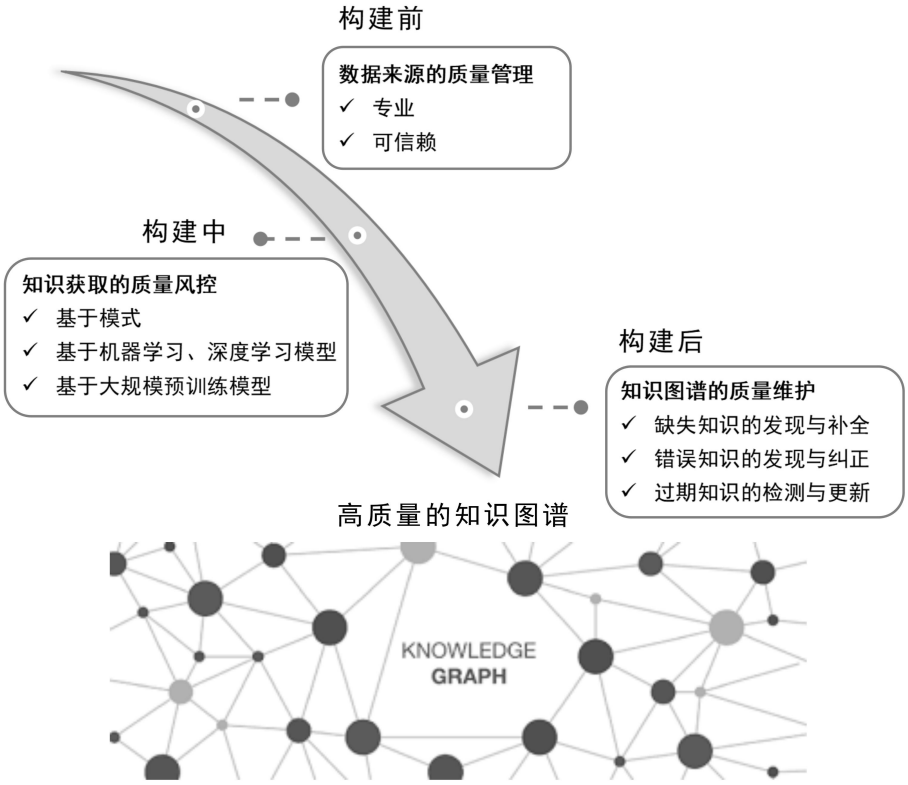


图 1 知识图谱质量管理全周期概览

1）图谱构建前：数据来源的质量管理

知识图谱构建前的质量控制主要关注数据来源的质量。在新闻和传播领域很早就开始关注信息来源（信源）的可信度评估问题，并提出衡量信源可信度最关键的两个因素专业（Expertise）和可信赖（Trustworthiness）[肖仰华 et al, 2020]。“专业”衡量的是信源在某领域的专业性，而“可信赖”衡量信源所提供内容的可靠性。知识图谱中的知识来自各种各样的知识源（或数据源），最常见的如新闻媒体、知识库、数据库、互联网上的各类网站以及用户贡献的知识等。在知识图谱构建之初，如何对数据来源进行可信度评估，不仅是知识图谱构建的起点，更是重点所在[肖仰华 et al, 2020]。

2）图谱构建中：知识获取的质量风控

知识图谱的构建过程需要不断从各种知识源获取知识。知识源不同，获取知识的手段不同，所需的质量控制方法也不同[肖仰华 et al, 2020]。当前，面向自然语言文本的自动化知识获取技术仍然是知识获取技术的主要阵地，主流方法包括基于模式、基于机器学习和基于深度学习模型的知识获取技术，另外今年来随着大规模预训练模型的兴起，基于大模型的知识获取技术也受到广泛关注。具体分类来说：对于基于模式的知识获取技术，重点要关注的是所获取模式本身的质量以及其可能引入的噪音，尤其是在自举式迭代抽取过程中可能发生的“语义漂移”问题[Paşca et al, 2006]。而对于基于机器学习、深度学习模型的知识获取技术，

则需要关注于相关模型方方面面的数据和性能问题，包括：训练数据的质量问题、小样本或者零样本问题、样本不均衡问题、过拟合问题等等。近年来，随着大规模预训练模型的快速发展，很多知识获取技术都依托于预训练模型来进行，包括抽取式模型、生成式模型等等。虽然模型的性能得到了大幅提升，但依然有引入噪音知识的风险，仍然需要尽力做好质量的风险控制[肖仰华 et al, 2020]。

3）图谱构建后：知识图谱的质量维护

虽然在前期阶段做出了必要的质量管理工作，但自动构建而成的知识图谱不可避免地还是会存在质量问题，具体包括知识缺失、知识错误和知识过期等[肖仰华 et al, 2020]。因此，知识图谱构建后的质量维护仍然十分必要。比较核心的工作主要包括：1）缺失知识的发现与补全； 2）错误知识的发现与纠正；3）过期知识的检测与更新[肖仰华 et al, 2020]。

二、知识图谱质量评估体系与机制

1. 评估度量维度

表 1 知识图谱评估度量维度

维度名称	描述	常见度量
精确性[Wang et al, 1996] [Naumann et al, 2002]	数据的正确、可靠程度	RDF 文档的语法有效性[Hogan et al, 2010] 文字、单词的句法有效性[Guns et al, 2013] 三元组语义有效性[Zaveri et al, 2015]
可信度[Wang et al, 1996] [Zaveri et al, 2015]	数据的正确、真实和可信程度	图谱级可信度[Zaveri et al, 2015] 语句级可信度[Zaveri et al, 2015] 空值与未知值设置[Zaveri et al, 2015]
一致性[Mecella et al, 2002] [Paulheim et al, 2017]	数据中的两个或多个值不会相互冲突矛盾	在插入新数据时检查图谱模式层限制约束[Zaveri et al, 2015] 类一致性约束[Bechhofer et al, 2016] 关系一致性约束[Bechhofer et al, 2016]
相关性[Wang et al, 1996] [Bizer et al, 2007]	数据对当前任务的适用和帮助程度	1. 语句级排序[Zaveri et al, 2015]
完整性[Wang et al, 1996][Mendes et al, 2012] [Luggen et al, 2019]	数据具有足够的广度和深度	模式层完整性[Pipino et al, 2002] 属性完整性[Pipino et al, 2002] 整体完整性[Hogan et al, 2020]

		可链接完整性[Darari et al, 2018]
时效性[Wang et al, 1996] [F¨arber et al, 2018] [Zaveri et al, 2015]	数据的时效	图谱的更新频率[Zaveri et al, 2015] 语句的有效期[Zaveri et al, 2015] 语句的修改日期[Zaveri et al, 2015]
易于理解性[Wang et al, 1996]	数据清晰、无歧义和容易理解的程度	资源描述符[Heath et al, 2011] [Hogan et al, 2012] 多语言标签[Zaveri et al, 2015] 可解释的 RDF 序列化[Zaveri et al, 2015] 自描述 URLs[Zaveri et al, 2015]
可访问性[Wang et al, 1996] [Naumann et al, 2002] [Zaveri et al, 2015]	数据可用或快速检索的程度	资源的可引用性[Jain et al, 2010] 公共 SPARQL 节点配置[Zaveri et al, 2015] RDF 导出配置[Zaveri et al, 2015] 图谱的元数据[Heath et al, 2011]
许可性[Zaveri et al, 2015]	授予使用者重用数据的权限	机器可读的许可信息[Hogan et al, 2012] [Heath et al, 2011]
互联性[Zaveri et al, 2015]	同一概念的实体相互链接的程度	通过 owl:sameAs 关系互联[Hogan et al, 2012] 外部 URIs 的有效性[Zaveri et al, 2015]
唯一性[Zaveri et al, 2015]	数据在广度、深度和范围上无冗余的程度	冗余类的比例[BEHKAMAL et al, 2014] 相似属性的比例[BEHKAMAL et al, 2014] 冗余实例的比例[BEHKAMAL et al, 2014] 属性值冗余的比例[BEHKAMAL et al, 2014]
安全性[Flemming et al, 2011]	数据不受更改和误用的保护程度	使用数字签名[Flemming et al, 2011] 数据的真实性[Flemming et al, 2011]
性能[Flemming et al, 2011]	性能是一个影响信息系统或搜索引擎质量的维度，而不是影响数据集本身	低延时[Flemming et al, 2011] 高吞吐量[Flemming et al, 2011] 数据源的可扩展性[Flemming et al, 2011]

通用性[Zaveri et al, 2015] [Flemming et al, 2011]	以不同的表示和国际化的方式提升数据的可用性	以不同的序列化格式提供数据[Flemming et al, 2011] 以不同种语言提供数据[Flemming et al, 2011] [Auer et al, 2010] [Gayo et al, 2012]
代表性[Ricardo et al, 2018]	知识图谱中是否包含高层次偏差	数据偏差[Hogan et al, 2020] 模式层偏差[Hogan et al, 2020]
数据丰富度[Ruan et al, 2018]	数据内容丰富程度	类与实例数目[Ruan et al, 2018] 网络链接度[Ruan et al, 2018]
使用质量[Ruan et al, 2016]	基于使用上下文的数据质量度量	用户构造查询的难易程度[Ruan et al, 2016] 基于特定上下文用户可获得的信息[Ruan et al, 2016]

文章[F¨arber et al, 2018] 将度量分成了内在的（Intrinsic）、上下文相关（Contextual）、表示（Representation）和可访问性（accessibility）等 4 个类别，又进一步分成了 11 个维度。其中，内在类别（Intrinsic category）包含了精确性、可信度与一致性三个维度，而上下文相关列别（Contextual ）的包括了相关性、完整性和时效性三个维度。而表示类别分为易理解和互操作两个维度，可访问性包含了可访问，许可性与互联性三个维度。每个维度都包含了多个度量。表 1 全面阐述了知识图谱的各评估维度，并给予了定义以及解释。

1）度量的类别与维度

首先，数据质量被称为”适用性 fitness for use“，因此，数据是否满足需求，是和应用需求有关的。而在上下文相关类别中，相关维度和度量隐含的定义了需求，比如说需要哪些类别，需要哪些属性，而 completeness 维度的类别完整性与属性完整性隐含了这样的需求。其次，外部人员使用数据的便利程度，是通过可访问性（accessibility）类别来度量的。这个特性大量的和开放链接数据的标准有关，比如说，是否有 SPARKQL 节点配置，或者运行节点是否可以导出数据等等。再者，开放链接数据的表示，一般基于 RDF 语言，而对 RDF 特定语言设施的使用，会对理解和互操作造成影响。比如说，描述字段和多语言字段描述，会增加理解。而使用空白节点，可能会降低互操作性。最后，常见的数据的正确性，数据间的一致性等等，可以看作是数据的本质特征，放在内在（Intrinsic）类别中。

也有文献定义了其他维度，如安全性，性能，通用性、代表性、丰富性、可使用性等等。安全性和性能可以分到可访问性类别，性能与使用过程的用户体验有关，而安全性代表了对数据访问过程的某种保护。通用性指的是主要是可以用不同方式输出和表达数据，可以放在表示类别里面。代表性指的是总样本很大，数据集合只是整体的一个采样的情况下，该数据

集合是否代表了整个分布，这个可以放在上下文维度，可以看作是完整性的一个侧面。丰富性代表的是数据大小，可以看作是内在特征，而可使用性的两个维度，参考了软件工程领域有关外部质量、内部质量与使用质量（quality in use）的划分， 提出了可查询性和信息性，前者指的是基于数据集合构造一个查询的难易，而后者指的是在这个查询下的信息量，前者可以放在可访问性（accessibility）类别里面，而后者可以作为上下文相关的数据完整性度量，可以放在完整性里面。

2）常见维度与主要度量

由于维度和度量诸多，此处选择最为常见的或是本质的一些度量，如准确性，一致性，完整性，唯一性。下一部分描述与开放链接数据知识表示相关性密切的一些维度和度量。

语法准确性关注知识图谱中每一条三元组在语法层面上正确与否。我们说一条三元组在语法上是正确的，则其实体所对应的概念是符合此关系或属性的特性的。比如说，<大海，颜色，黄色>这样一个三元组，虽然与事实不符，但却并没有违背属性/关系的规则，这是一个语法上正确的三元组。又如属性<始于>需要跟一个时间类型<xsd:dateTime >，而图谱数据中错误地将其用字符串类型< xsd:string >表达。

语义准确性强调三元组所表述的语义信息与真实世界的事实的接近程度。一般来说，我们将三元组的语义准确性分数限制在 0-1 之间，语义准确性分数越接近 0 代表三元组表述越有违事实，越接近 1 则代表三元组所表述事实在语义层面越接近真实情况。基于语义准确性分数，三元组可以在某个划分阈值下被分为正确的或者是错误的三元组。举例来说，<大海，颜色，黄色>从语义角度评估一定是一个低准确性得分的三元组，相应的，<大海，颜色，蓝色>则能正确的表述客观事实。

完整性关注的是是否所有领域所需知识都被知识图谱所表达出来。Wang 等人[Wang et al, 1996]将完整性描述为数据对于目标任务具有足够广度、深度以及范围的程度。Pipino 等人[Pipino et al, 2002]指出完整性可以分为本体完整性、属性完整性以及数量完整性。本体完整性指的是本体中的类目与属性在图谱中的呈现出来的程度。属性完整性指的是图谱中某个类别下某个属性的缺失程度。数量完整性关注本体中某个概念的实体在图谱中出现的情况与真实世界实体个数的比例。

代表性又可以称为图谱的偏向性，其在一个较高的视角关注图谱的知识偏向问题。图 2-1 展示了完整性与代表性的区别。代表性在默认知识图谱的不完整性的基础上，认为知识图谱是所谓“完美”知识图谱的一个子集，并讨论这个子集偏向哪方面的知识。如图所示，完整性与代表性关注的侧重点不同，代表性总是更加关心知识图谱中的数据更偏向哪一个层面。具体来讲，由于数据本身具有偏向性，并且人们的思想本身也存在偏见，那么无论是由原始

数据中抽取的知识，又或是人工编写的知识都无法避开偏向性的问题。既然在知识图谱中本身的知识是不完整的，那么去衡量其在各个角度上的知识偏向也就成为了一个重要的问题。

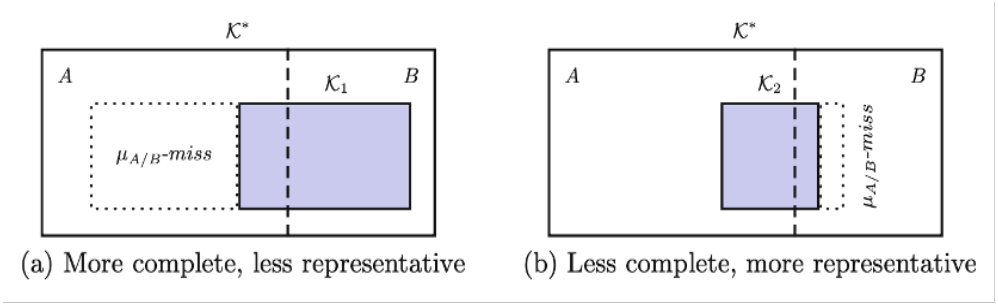


图 2 完整性 vs 代表性[Soulet et al, 2018]

可信度被定义为评估知识正确、真实和可信的程度。比起准确性，该指标更强调主观的一些判断。例如，图谱级可信度度量，主要对图谱的数据来源进行可信评估。来源可分为三类：知识是否来自于领域专家；知识是否来自于社区贡献者；知识是否从数据源自动提取。再如，语句级可信度度量，可以通过 dcterms:provenance、dcterms:source 等信息标明数据来源，提升数据的可信度。一个比较有意思的度量是空值可信度度量，即如果图谱的空值或未知值，也代表了某种逻辑含义，则可信度增高。比如说，空值表示某人没有后代。空值赋予含义，说明图谱做得非常细致。

一致性被定义为数据中的两个或多个值不会相互冲突的程度。准确性侧重于评估数据的语义正确性，而一致性通过逻辑的相关性，找到数据噪音或是数据质量问题。文章中定义的几个一致性度量还是比较简单的，本质上是通过图谱模式约束检查，减少图谱中数据不一致的错误。如数据类型是否符合预期、插入实体的类型是否有效、实体与其所属概念是否矛盾等，实例数据与指定类的一致性程度，实例数据与指定关系的一致性程度等等。

时效性被定义为数据的时效。时效性是一个相对的概念，取决于实际应用场景。对于描述现实世界中事物的常识性知识图谱，无需频繁刚更新迭代便能满足时效性要求，而对于一些基于公共热点事件建模的领域知识图谱，则需要定时更新维护才能满足时效性要求。该指标可通过图谱的更新频率、语句的有效期和修改日期进行度量。

唯一性衡量数据在广度、深度和范围上无冗余的程度。可以从类、属性、实例和属性值多个角度衡量冗余程度。对于大规模知识图谱，由于数据集合可能是通过自动抽取以及多源数据融合获得，因此，会有相同类不同名称，相同属性不同名称，或是同一个实例多次出现等各种情况。

3）与 RDF 规范密切相关的维度与度量

易理解性被定义为数据清晰、无歧义和容易理解的程度。该维度用于评估使用者对数据

源的可理解程度。基于 RDF 规范，数据源的可理解性能通过资源描述符、多语言标签、可解释性的RDF序列化和自描述URLs等多种方式来提升。例如，引入rdfs:label和rdfs:comment增强数据的可理解性；插入诸如 rdfs:label、skos:prefLabel 等标签信息，以人可读的方式描述数据资源，或是多语言标签可以让不同国家的使用者理解数据资源。

可访问性被定义为数据可用或快速检索的程度。该维度也包含了很多度量，如公共 SPARQL 节点配置，允许用户在图谱上执行复杂查询任务。或是 RDF 导出配置，当 SPARQL 节点使用不便时，用户可导出转存 RDF 数据。

互联性概念体现了知识图谱的特点，指不同数据集合之间的实体可以互相链接的程度。这些互联通常是在实例层上通过 owl:sameAs 链接建立。因此，可以计算链接到外部知识图谱的 owl:sameAs 关系比率来度量实例级的互联率。另外，图谱中可能包含引用 RDF 资源或 Web 文档链接，通常由 owl:sameAs、owl:equivalentProperty 和 owl:equivalentClass 关系完成。可以通过评估外部 URLs 是否存在超时相应、客户端错误和服务器端错误来衡量其有效性。

2. 评估流程与方法

不失一般性，知识图谱数据评估流程一般分为一下几个步骤：

- 确定应用需求。即前文所说的应用场景或是应用上下文。 不同的应用上下文需要的数据质量是不一样的。
- 基于应用上下文，确定需要进行评估的维度和度量，确定度量计算方法。
- 导入需要评估和比对的一到多个数据集合。
- 度量计算
  - 如果是机器评估，则编写程序进行度量。
  - 如果是人工评估，基于全量数据，或是对数据进行采样、分组，分发到多个评估人员。
- 执行度量，获得度量结果。
- 基于结果进行统计分析。

这个过程比较复杂的在于基于应用上下文确定度量，以及度量的具体的计算方法。大多数度量的计算方法还是比较简单和直接的，通过检查系统的某些状态信息，匹配一些既定义的规则，或是数据的统计即可获得，在论文[F¨arber et al, 2018] 中，几乎所有的度量都可以用上述方法获得。相比而言，比较复杂的在于语义层面上的度量，即数据本身的准确性、一致性的探测以及数据对于应用需求的完整性的覆盖。因此，下文重点阐述这些维度上的度量计算方法。

1）语法准确性评估：

语法准确性的评估需要使知识图谱中图谱的三元组表达符合其规则的定义，正如前文所说可以基于一些即定义的规则进行简单的评估。例如对于<foo:gender>这样的属性，其值只能被限定在如 mail, femail 这样的词语之间。现有的语法纠错工具如 w3c 提供的 rdf validation service 提供了丰富的语法验证服务。如图 3 所示，基于语法验证工具，可以轻松的对三元组进行语法解析与验证。

图 3 RDF 验证服务

除此以外，基于手工设计的规则，Fürber 等人[Fürber et al, 2011] 提出 SWIQA 框架，以此来对数据质量进行量化评估。SWIQA 利用一些手工设计的规则来判断三元组的正确与否，基于在某个维度上语法正确的三元组个数，可以计算其在此规则上的知识图谱语法准确性得分。表 2 中列出了框架中的一些语法规则，我们可以运用或设计这样的规则来进一步评估知识图谱的语法准确性。除了语法层面的规则外，也涵盖了部分语义准确性、完整性的规则以在多个维度衡量图谱质量。由于规则设计的不全面性，SWIQA 这样的框架可以在语法准确性评估中发挥一定作用，但却不能很好的在其他维度评估知识图谱。可以看出，语法准确性的度量是可以轻松依据语法规则实现的。我们在评估相关知识图谱语法准确性时，最有效的方式也是定义当前知识图谱需要关注的语法规则。

表 2 SWIQA 框架语法准确性规则

语法规则名	定义	举例
句法规则	Syntactic Rules 定义了字符的类型或文字值的模式。	属性<foo:country-name>的值必须只包含字母。
合法值规则	Legal Value Rules 是对某一属性的允许值的明确定义。	属性<foo:gender>只能包含值″male″，″female″，″m″或″f″。
合法值域规则	Legal Value Range Rules 是对数值属性允许的值范围的明确定义。	属性<foo:population>必须只包含大于 0 的值。

2）语义准确性评估

基于知识图谱表示学习的方法： 知识图谱表示学习基于特定视角建模，学习实体与关系的向量表示，有着较高的学习效率，可以应对大型的知识图谱。在评估三元组的语义准确性方面，可以利用其得分，衡量三元组的置信程度。经典的表示学习方法可以为三元组的语义准确性给出打分，但由于其得分函数的不一致，我们可以在计算语义准确性时修改最后得分函数，如使用一些非线形函数将值限制在[0-1]之间或修改损失函数重新优化模型，以量化其语义准确性。也可以基于原先的得分与判定阈值相差的多少来判定三元组语义层面的正确与否。传统的表示学习方法有基于翻译模型的方式[Bordes et al, 2013]，基于张量分解的方式[Trouillon et al, 2016]等。除了单一的表示学习建模方式外，也有方法在知识图谱表示学习的基础上，融入其他信息，增强模型性能。Xie 等人[Xie et al, 2018] 提出的 CKRL 是一种基于知识图谱表示学习进行可信度判断的算法。该方法定义了局部三元组可信度和全局路径可信度来检测知识图谱中的噪声三元组，结合了三元组的内部语义信息和知识图谱的全局推断信息。 PRGE[Bougiatiotis et al, 2020]将路径信息添加到模型中，并通过引导损失函数，将三元组置信度注入 TransE 的损失函数中引导模型关注可能正确的三元组。KGTtm[Jia et al, 2019]利用交叉神经网络结构检测知识图谱中的错误，从知识图谱语义信息、整体结构信息及路径信息来评价三元组的可信度。如图 4 所示，其融合了各方面信息来评估三元组准确性。

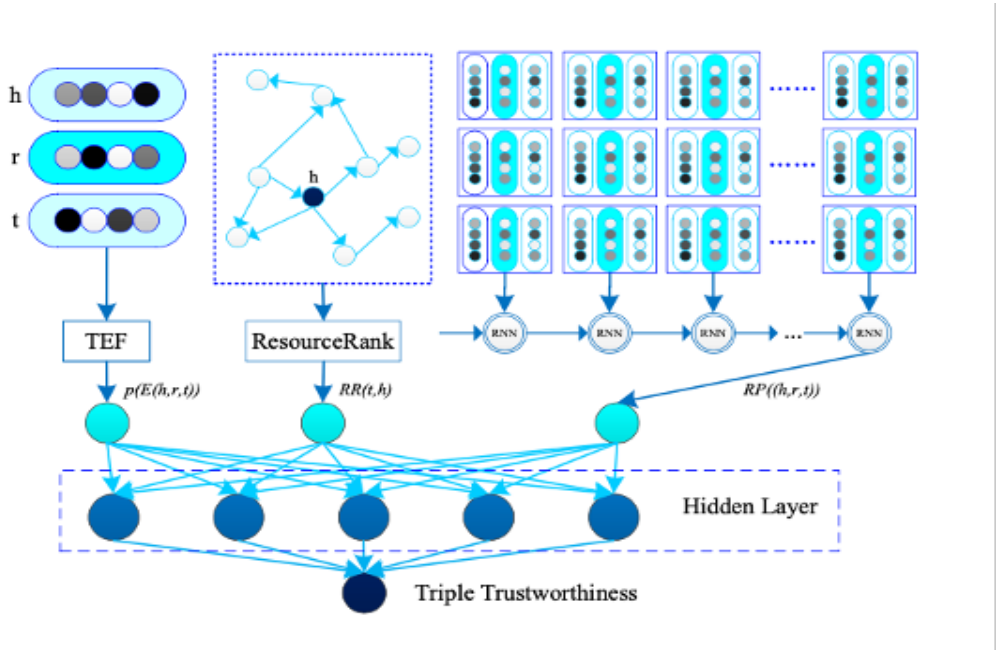


图 4 多信息融合模型 KGTtm[Jia et al, 2019]

基于证据搜索的方法：这类方法在文本中搜索支持三元组成立的证据，但这会花费很多时间，所以它们通常处理的是像 FactBench[Gerber et al, 2015]这样的小规模数据集。这些方法的缺点是难以适应较大的数据集，且过于依赖外部语料库和搜索引擎。例如，

Defacto[Esteves et al, 2017]使用像谷歌这样的搜索引擎，并根据网页找到证据。它利用 BOA 自然语言的模式[Gerber et al, 2012]来获取三元组对应的自然语言语句，并以语句作为输入。然后，它试图通过在 Web 中搜索文本信息来寻找支持三元组有效性的证据[Gerber et al, 2015]，使用互联网上的信息的效率并不高，并且依赖于搜索引擎 API，其中许多开放的 API 如 Google 现在已经关闭，这也导致这样的方法显露出更多不足。FactCheck[Syed et al, 2018]首先将三元组语料转化为自然语言，然后在静态语料库中搜索这些自然语句。 如图 5 所示，其从最初搜索的文档中提取特征，并将这些特征输入到像 Naïve Bayes 这样的机器学习模型中，以获得三元组的语义准确性分数。

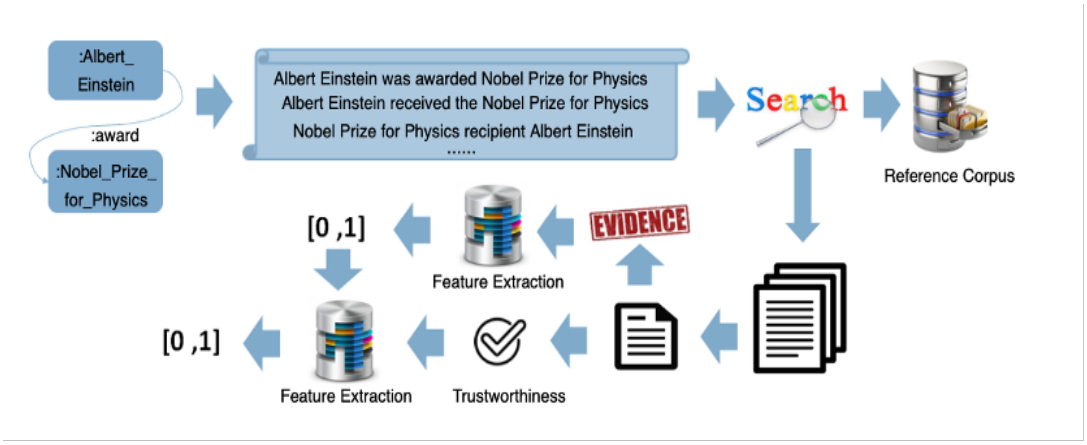


图 5 证据搜索模型 FactCheck[Syed et al, 2018]

3）完整性评估

如表 3 中所示的完整性规则，基于规则的 SWIQA 框架也在完整性评估方面定义了相关概念，通过面相具体的任务，制定此任务下某些类所必要的属性，来评估知识图谱的属性完整性。除此以外，本体完整性也可以通过筛查本体中类与关系在图谱的缺失程度，来计算其完整性。这些都可以依靠我们对数据指定的完整性规则，来进一步的评估其完整性。数量完整性在很多时候比较难以进行评估，我们很难界定具有怎样规模的知识图谱才是一个“完美”的知识图谱。那么在这样的情况下，完整性也很难被量化出来。也有一些工作进行了这方面的研究，Fariz 等人[Darari et al, 2018]研究了如何去描述完整性，并且提出了一种理论框架，基于此能够以三元组的形式去表达知识图谱中的完整性。Luis[Galárraga et al, 2016]等人解释了什么样的数据是完整的，提出了许多对完整性的假设，并在这些定义之下计算知识图谱的完整性。对于解决完整性问题，描述完整性的金标准还是需要去定义的，在此基础上才能更好的讨论知识图谱的完整性。

表 3 SWIQA 框架完整性规则

语法规则名	定义	举例
强制属性规则	如果目标任务需要，则一些属性将成为强制性的。	为了能够导航到每个位置，所有< foo:Location>类的地理坐标属性都必须被明确

4）代表性评估

代表性或者偏向性的评估通常也需要给出一定的评价标准，以一定的视角来看待图谱是否具有明显的偏向。典型的工作如 Arnaud 等人[Soulet et al, 2018] 提出的基于数学统计定律的方法，其以 Benford's Law 为理论数据分布，将其扩展到知识图谱的数据上，进行代表性的评估。除此以外，不少工作从不同角度来指出图谱的偏向性。Ricardo[Ricardo et al, 2018] 指出了偏见的问题，并对现有数据偏见做了分类，使我们意识到网络数据中的偏见问题。Janowicz 等人[Janowicz et al, 2018]用已知分布如地理，人口的分布来评估偏向性。 用语言分布来分析多语言知识库中的偏向性。Kaffee 等人[Kaffee et al, 2017]用社会中敏感问题如性别问题，用现有社会男女分布来分析知识库男女分布。由此可见，代表性可以针对某一个问题，在一定的角度上，借助现有的真实数据分布以及知识图谱中的知识分布来进一步分析知识图谱的代表性或者是偏向性。

三、知识图谱质量维护方法

如本章任务定义中所示，知识图谱构建的各阶段都存在引入质量问题的风险。对于知识图谱的质量维护工作应该本着"应早尽早"的原则：尽量从"源头"上预先发现和解决质量问题，把各阶段可能产生的质量风险控制在萌芽阶段，以免该质量问题在后续阶段变得更加难以识别，且产生级联放大的后果[肖仰华 et al, 2020]。

对于知识图谱构建所需数据来源的质量管理，以及各类知识获取技术的质量风控，由于更多涉及到其他领域的研究工作，在此不做赘述。我们接下来主要介绍一下关于知识图谱构建完成后的质量维护方法。特别说明：本小节主要框架与内容部分摘录于电子工业出版社出版、肖仰华老师主编的《知识图谱：概念与技术》[肖仰华 et al, 2020]中由李直旭老师主笔撰写的知识图谱质量控制章节。

1. 缺失知识的发现与补全

1）实体类型补全

实体类型补全（或实体判型）是对知识图谱中某些尚未挂载到对应上位概念的实体找到其所对应的上位概念。早期的实体判型方法主要依赖于启发式先验概率模型。该类方法主要

通过构建一些启发式规则或概率模型来实现实体类型补全。例如基于三元组谓词的启发式实体判型方法 SDType[Paulheim et al, 2013]首先统计出所有谓词的头尾实体的类型分布，从而再根据每个实体所搭配的谓词的三元组情况，来具体估算出每个实体的类型分布概率；还有一些工作如 Probase+[Liang et al, 2017]则考虑利用协同过滤的思想补全上下位关系，基本思想是认为相似语义的元素倾向于在概念图谱中共享上位词/下位词。也取得了不错的效果。

当前较为主流的实体判型工作是深度实体分类模型，尤其是对于细粒度的实体判型任务，其挑战来自于其大规模和细粒度的实体标签集。现有的很多方法使用预定义的标签层次[Ren et al, 2016] [Shimaoka et al, 2017] [Abhishek et al, 2073] [Chen et al, 2020] [Ren et al, 2020]或来自训练数据的标签共现统计[Rabinovich et al, 2017][Xiong et al, 2019][Lin et al, 2019]作为外部约束。然而，这些方法需要预定义的标签结构或来自训练数据的统计数据，因此很难扩展到新的实体类型或领域。超细粒度的标签集也导致了长尾问题。为了解决相互依赖的和长尾的实体难以细粒度分类的问题，Liu 等人[Liu et al, 2021]认为标签之间隐含的外在和内在的依赖关系可以提供关键的知识，并基于此提出了与上述方法不同的标签推理网络（LRN），通过发现和利用数据中隐含的标签依赖性知识来顺序推理细粒度实体标签。

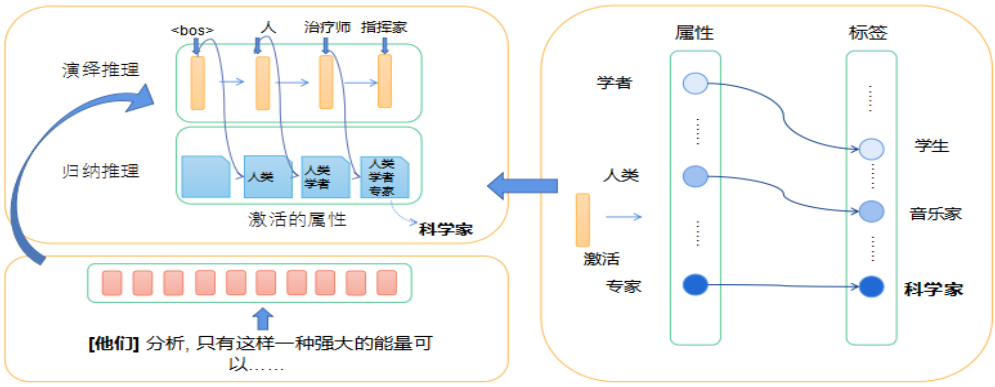


图 6 Liu 等人提出的 LRN 框架[Liu et al, 2021]

具体来说，LRN 首先通过一个上下文敏感的编码器对实体提及进行编码，然后通过两种标签推理机制，即利用外在依赖关系的演绎推理和利用内在依赖关系的归纳推理，依次生成实体标签。在 Seq2Set 框架中，标签依赖性知识可以有效地在 LRN 的参数中建模，从训练数据中自动学习，并在连续的标签解码过程中自然利用。如图 6 所示，将 mention 编码以后，首先通过外在依赖的演绎推理，得出标签“person”，此时“person”的属性“human”被激活，接着通过归纳推理，得出标签“scientist”。这种方法能够有效地以端到端的方式对复杂的标签依赖关系进行建模、学习和推理。

除了利用实体类型之间的隐含依赖关系，也有工作通过改变嵌入方式捕捉类型之间的层次。Onoe 等人[Onoe et al, 2021]就利用了盒式嵌入，在高维空间中表示实体类型。如图 7 所

示，实体的提及和上下文都通过 BERT 被嵌入到盒子空间中，然后每一种类型的概率都使用软体积计算的方法来计算。这种方法可以自然捕捉实体类型之间的复杂依赖关系，从而很好地完成实体的细粒度分类。

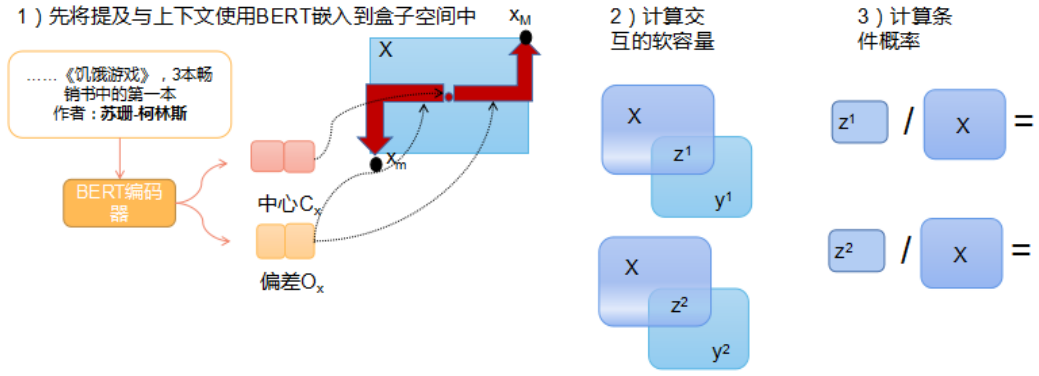


图 7 Onoe 等人提出的基于盒子嵌入的框架[Onoe et al, 2021]

上述方法基于分类范式，往往将实体分类到一组预定义的类型。然而预定义的类型集合是有限的，因此这些方法无法将实体分配给预定义集之外的类型。此外少样本和零样本的问题也没能得到充分解决。针对这些问题，Yuan 等人[Yuan et al, 2023]提出了一种新颖的生成式实体类别标记任务，即给定一个带有实体提及的文本，实体在文本中扮演的角色的多种类型是使用预训练语言模型生成的。具体来说，如图 8 所示，首先构建输入，去指导基于预训练语言模型的模型学习。接着课程指导模块负责衡量训练数据中每个样本的难度，然后为模型训练过程设计合适的课程。最后根据设计好的课程，模型被进一步训练，在这一步中，课程通过自定进度的学习（SPL）进行自我调整，使模型能够动态地调整其学习进度。

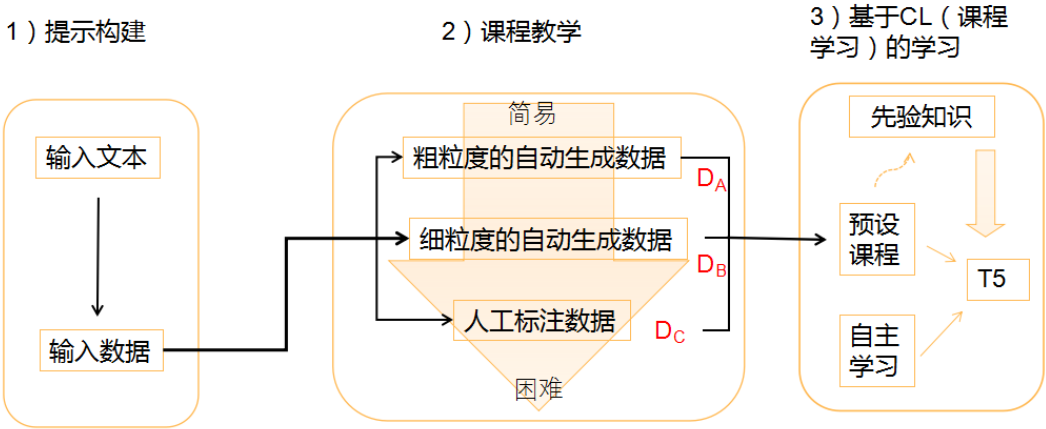


图 8 Yuan 等人提出的框架[Yuan et al, 2023]

与基于分类的实体类型方法相比，基于预训练语言模型的生成式实体类别标记可以为实体生成超出预定义集合的类型；此外，由于预训练语言模型在预训练期间编码了大量知识，因此该方法能够进行概念推理并处理少样本和零样本的困境。作者还使用了课程学习的训练

策略，将类型长度和数据类别作为先验知识诱导预训练模型可以生成高质量细粒度的类型。

2）实体关系补全

实体关系补全是对知识图谱中不完整的关系三元组（头实体, 关系, 尾实体）进行补全。相关方法大体可以分为三类：1）概率图模型：大致做法是为知识图谱上的每一条候选知识附上一定的概率用以此衡量该候选知识的合理性，通过概率推理发现缺失关系[Richardson et al, 2006][Chen et al, 2014]；2）路径排序算法；基本思想是用连接两个实体的路径作为特征，来预测两个实体间的关系[Lao et al, 2010][Gardner et al, 2015]。然而这些方法泛化能力都不够，难以达到较好的补全结果。

当前主流的实体关系补全模型是基于表示学习的补全模型。首先要在低维向量空间中对知识图谱中的实体和关系进行表示，表示形式包括向量、矩阵或张量形式[Paulheim et al, 2015][Wienand et al, 2014]。之后，在每个知识条目上定义打分函数，来判断三元组成立的可能性。推理阶段也会根据打分函数等进行推理。最初知识图谱中的知识表示用到的信息源主要是三元组信息，之后实体的类别、关系路径、实体的描述文本，以及简单的逻辑规则等也被融入表示学习过程中，这样可以学习到更为准确和全面的表示，从而提升推理补全任务的准确度[肖仰华 et al, 2020]。

考虑到三元组重要的邻居信息，为更好地进行关系补全，Nathani 等人[Nathani et al, 2019]提出使用编码器-解码器模型，并在编码时采用多头图注意力网络，从而获得不同实体的多跳邻居信息。由于实体在与不同关系关联时角色并不相同，因此 Nathani 等人扩展图注意力机制，在注意力机制中考虑多跳邻居的实体和关系特征，通过学习与实体相关联的所有邻居三元组表示来更新实体的表示。最后通过解码器对缺失的关系进行补全。

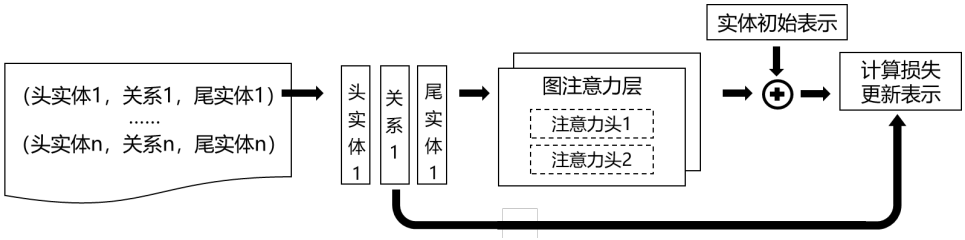


图 9 Nathani 等人提出的模型框架图[Nathani et al, 2019]

另一方面，知识图谱的层次结构和逻辑关系也受到了人们的关注，ATTH[Chami et al, 2020] 提出利用双曲空间对于知识图谱层次结构的建模优势，在双曲空间中采用基于注意力的几何操作学习知识图谱，提升补全效果。在双曲空间中，ATTH 采用双曲平移变换建模层次结构。在建模逻辑关系时，ATTH 对于对称关系采用反射几何操作，对于非对称关系或合成关系采用旋转几何操作，考虑到知识图谱中存在各类复杂的关系，ATTH 采用注意力机制平衡反射和旋转这两个操作。在实体补全例如尾实体补全时，首先将头实体与关系进行逻辑

模式编码和层次结构建模，再通过计算其与尾实体之间的双曲距离进行尾实体补全。

为获得更好的补全效果，除了使用三元组作为信息源进行知识表示，人们也将其他重要信息融入表示学习中，例如 CyGNet[Zhu et al, 2021]将三元组的有效时间信息融入表示学习中，采用复制-生成的模式利用历史上已发生过的事实提升补全效果。对于实体补全任务，CyGNet 采用复制模式基于其相关的历史词表计算历史词表中各实体的概率，并采用生成模式基于所有实体词表计算各实体概率，最后综合复制模式与生成模式的预测概率进行实体补全。

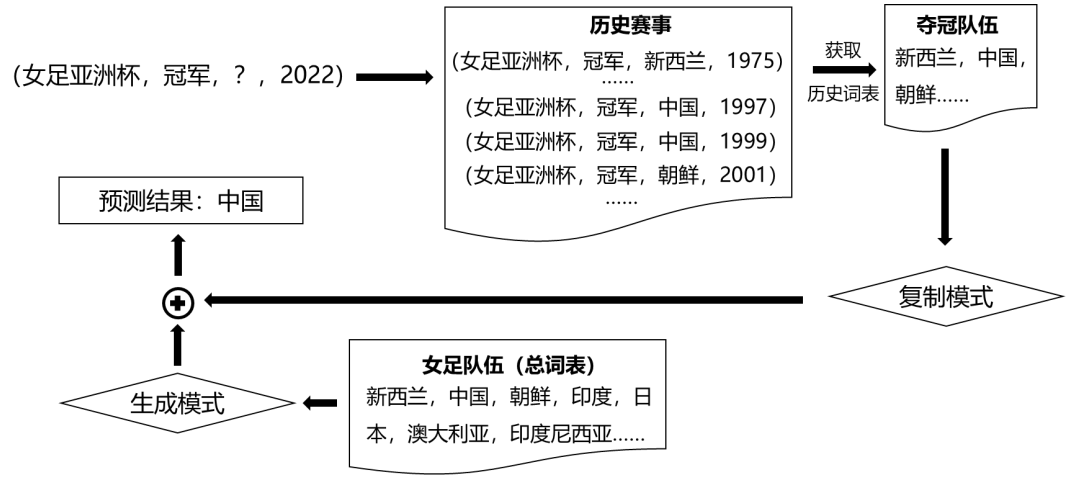


图 10 CyGNet 模型推理机制[Zhu et al, 2021]

3）实体属性补全

实体缺失属性的补全具体分为俩个子任务。其一是实体的缺失属性发现，该子任务往往被转化为某种实体类型下的实体的必有属性判定问题。即如果已知一个概念的必有属性有哪些，而这一概念下的某一个实体并没有这些属性和对应的属性值，则可以判定实体缺失这些属性；其二是对实体所缺失属性的属性值进行补全[肖仰华 et al, 2020]。

必有属性判定：早期的概念必有属性发现方法往往通过统计此概念下已有实体的属性分布情况来判定[肖仰华 et al, 2020]。比如，如某种实体类型的所有实体拥有某属性的比例超过给定阈值（如 70%），则认为该属性为这类实体的必有属性[Moser et al, 1992]。此外，还有一些方法会基于一些特定假设来建立检查实体属性或属性值完整程度的判定机制。常见的判定规则包括[Razniewski et al, 2018]：考察属性的重要程度；参考同一概念下的其他实体；参考相似实体；以及做一些实体与属性的模式挖掘与规则匹配等。

近年来，基于深度学习和表示学习的必有属性判定模型成为主流的技术。有学者认为实体的属性可以通过它的概念类别来获得，实体可以作为它的概念类别的实例并继承它们的属性。相比于用一个单词来描述概念，他们使用概念路径来表示实体概念的层次结构，使得概念得到细化，并且更加准确。同时，实体往往具有多个概念，概念路径可决策实体的不同概

念。例如“苹果”和“橘子”均具有颜色的属性，是由其二者均具有“水果/植物/生物/物”这一概念路径决定的，并且，对于实体“苹果”来说，不同的概念路径区分其不同概念，例如概念“水果”、“公司”、”电影“等等。受到知识图谱表示学习的启发，他们将图谱所具有的层次化概念体系和属性集合中的属性映射到连续的向量空间，从而将属性获取问题转化为链接预测任务，即从属性集合中为图谱中的实体预测合适的属性。

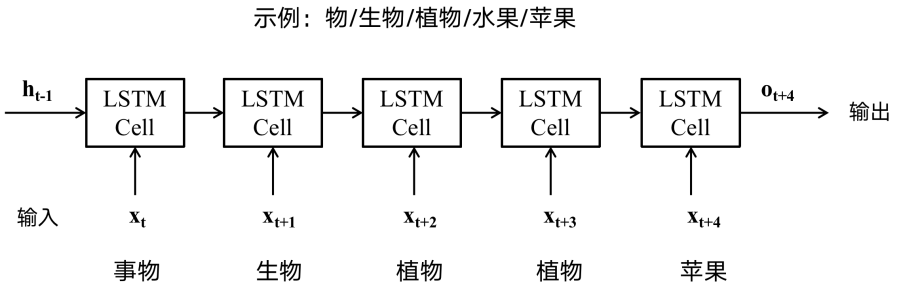


图 11 LSTM 对概念路径编码

首先对于概念路径，如图 11 所示，他们通过 LSTM 来实现获取概念路径的表示。由于实体的属性由其概念路径决定的，他们基于注意力来学习实体的概念体系和属性之间的映射关系，如图 12 所示。最后基于翻译模型的思想来预测实体的属性，即当某概念路径 p 拥有属性 a 时，他们之间的距离较短，反之，距离较长。

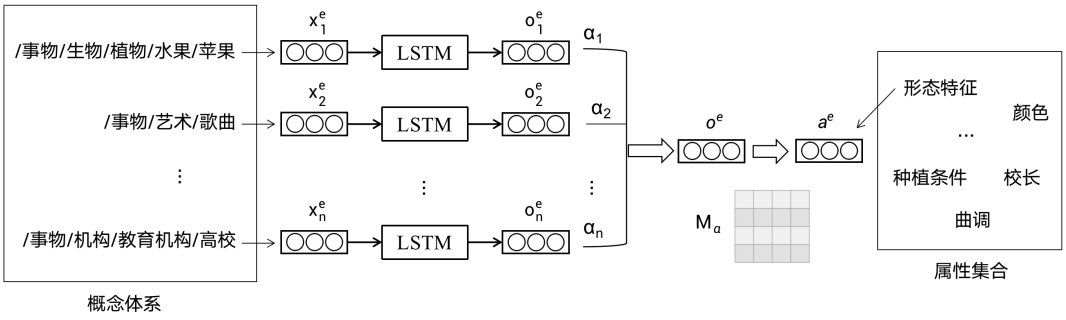


图 12 属性预测整体框架

另外，Fu 等人[Fu et al, 2013]设计一种基于属性密度的关联属性聚合模型来生成概念的向量表示，挖掘多属性之间的共现关联特征。他们认为，属性 p 是知识图谱中所有含有属性 p 的实体组成的虚拟概念 cp 的必要属性，当知识图谱中的具体概念 c 与 cp 越相似，就认为 p 越有可能是 c 的必有属性。具体来说，在判定属性 p 是否为概念 ci 的必有属性的过程中，为了充分利用图谱中不同概念下属性间的共现关联性来判定概念的必有属性，他们首先设计一个基于属性密度的关联属性聚合模型来生成虚拟概念 cp 和具体概念 ci 的向量表示，如图 3 所示，对于具体概念 ci 来说，收集概念 ci 下所有实体，进而获取这些实体所拥有的所有属性（除了属性 p）；同样，对于虚拟概念 cp 来说，收集图谱中含有属性 p 的所有实体，进而获取这些实体所拥有的所有属性（除了属性 p），接着按照属性的密度大小定义权值来加

权聚合属性表示，用以获取概念的表示。

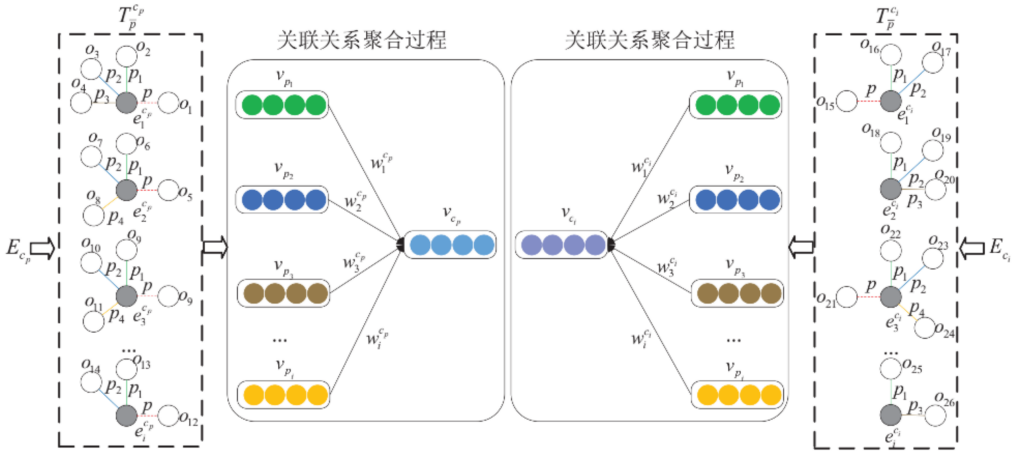


图 13 关联属性聚合框架[Fu et al, 2013]

获取了属性 p 的虚拟概念表示 vcp 和具体概念表示 vci 之后，通过余弦相似度计算两者相似度，vcp 与 vci 越相似，则 p 是 ci 的必有属性概率越大。同时，将图谱中概念的上下层次结构信息融入，即当属性 p 同时以高概率被判定为概念 ci 及其子概念 cs 的必有属性时，则 p 以高概率被判定为 ci 的必有属性。

属性值补全：属性值按照数据类型可以粗分为：可计量类型数据如年龄、收入、年份等；和不可计量类型数据如代表作、国籍、工作单位等[肖仰华 et al, 2020]。对于可计量类的属性值，传统的统计学方法会采用近似补全的方法，如取平均值法、最大似然估计法等，然而此类统计学方法得到的结果并非真实值，并不适用于知识图谱的场景。因此，知识图谱中的属性值补全通常把可计量类型数据也当做不可计量类型数据进行补全。常见的补全方法包括：基于众包的补全法、基于搜索引擎的补全法和基于文本的补全法等，即主要通过人力补全，或者借助于外部资源来进行检索补全。具体方法不再赘述。

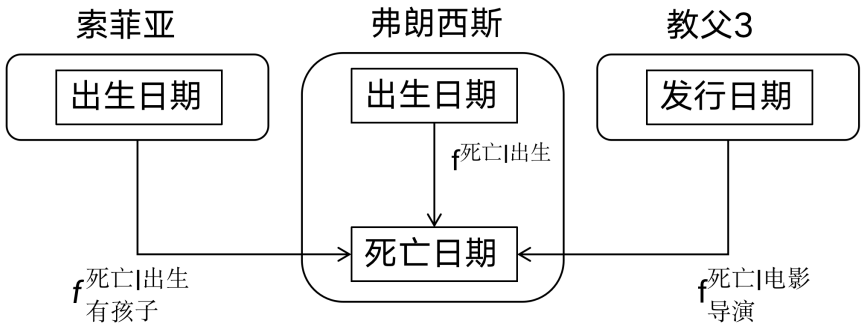


图 14 属性值预测的传播示例[Bayram et al, 2021]

近年来，属性值的预测更加自动化。Bayram 等人[Bayram et al, 2021]提出多关系属性传播算法 MRAP 通过迭代式地在知识图谱的多关系结构中传播信息来预测实体的属性值。具

体来说，它利用多个回归函数，对应根据节点之间的连接关系和属性的类型，从一个节点预测另一个节点的属性。对于中心实体自身来说，它还对中心实体内部的属性对采用另外一组回归函数用于预测。每个回归函数具有不同的权重，用于加权多个预测结果。如图 14 所示，中心实体为 Francis Ford Coppola，有两种信息传播方式：（1）实体外部多个邻居结点的传播：将邻居 Sofia Coppola、The Godfather III 的属性、关系信息传播进来；（2）实体内部属性对之间的传播：将属性 date_of_birth 的信息传播过来。这两种信息传播的方式联合预测中心实体 Francis Ford Coppola 的属性 date_of_death 的属性值。

2. 错误知识的发现与纠正

知识图谱中的错误知识的发现与纠正主要集中在发现与实体相关的错误知识，包括实体的概念、实体间的关系、实体属性值等。下面分别介绍一些相关工作。

1）错误实体类型检测

在知识图谱构建初步完成后，对于知识图谱中错误实体类型的检测，一方面可以采用“概念互斥”原则，即同一个实体不应该出现在俩个互斥的概念之下，他不可能既是一个“城市”，也是一个“演员”；另一方面，则需要依赖知识图谱中的知识来推断可能出错的实体类型[肖仰华 et al, 2020]。比如根据知识图谱计算实体的属性及属性值与实体概念之间的概率关系，从而根据属性来推断其概念。例如：一个实体有“代表作”这一属性多半是书籍，则其概念是“作家”的可能性较大，是“导演”的可能性较小。

2）错误实体关系检测

传统的错误实体间关系的检测研究大致分为两类：基于知识图谱内部数据的检测方法和借助知识图谱外部数据的检测方法[肖仰华 et al., 2020]。基于知识图谱内部数据的检测方法期望从图谱内部数据的关联关系中挖掘和建立错误关系判定规则。如[Dong et al, 2014]将知识图谱建模为图，从任意实体出发进行随机游走，如果能够通过一条路径到达目标实体，就将此路径记为一条可行路径。如果一对待考察实体对间的语义关系能在知识图谱中找到很多条可行路径，那么这对实体间的该语义关系则大概率是正确的[肖仰华 et al, 2020]。另一类代表性方法[Paulheim et al, 2014]使用数据的分布特征来发现错误实体关系：如果一个关系为 p 的三元组，其头实体和尾实体的类型都分别落在关系 p 的已有三元组的先验头/尾实体类型的高频分布区域，则正确概率较大，否则错误风险较大[肖仰华 et al, 2020]。借助外部数据的错误关系检测方法主要依托于互联网或外部本体库资源进行检测[肖仰华 et al, 2020]。比如基于互联网的检测方法[Li et al, 2012]主要利用搜索引擎获得三元组相关的页面，再训练监督模型来基于页面的置信度和上下文信息等来判定三元组是否正确。而利用外部本体库

进行错误实体关系检测的方法则通过可以通过将三元组中的关系和实体上升到本体层来考察其搭配上是否存在问题[Paulheim et al, 2015]。然而这些传统方法都只能处理一些容易发现的错误，很多没有显著类型矛盾的错误则难以发现。

近年来，基于知识图谱表示学习 TransE[Bordes et al, 2013]及其各类变种[Wang et al, 2014][Lin et al, 2015][Ji et al, 2015]的错误关系发现方法成为主流。然而，大部分的知识图谱表示学习方法仅仅集中于三元组的结构信息，忽略了实体中丰富的层次类型结构，针对这一不足，论文[Xie et al, 2016]提出了 TKRL 模型。层次类型结构是指在不同的场景下，实体可能具有不同的角色，这对知识图谱的表示学习有着很重要的作用。为了改进知识图谱表示的创建过程， [Guo et al, 2018]提出了一种新的知识图谱分布式表示学习方法——规则引导嵌入（rule-guided embedding，简记为 RUGE）。RUGE 借助软规则的迭代引导完成知识图谱表示学习。所谓软规则，就是那些不总是成立、带置信度的规则。这类规则可以经由算法从知识图谱中自动抽取。RUGE 模型已被[Hong et al, 2020]用于生成规则增强的噪声嵌入，用于低质量错误检测，重点是检测实体类型错误。[Cheng et al, 2018]提出了一种基于规则的错误检测方法，该方法引入了一组简单图的自动修复语义模式，称为图修复规则（GRR），其关注知识图谱中冲突的知识和冗余错误检测。[Ho et al, 2018]的工作方向相反，使用嵌入来指导规则学习，而[Zhang et al, 2019]的工作是迭代的，因此嵌入和规则可以相互指导。最重要的是， [Belth et al, 2020]提出了一种有趣的统一图形摘要和细化方法，开发了一组图形模式形式的软规则，以识别各种异常类型。由此产生的模型规定了节点标签之间关系的归纳规则，并优于基于知识图谱表示的错误检测方法，如 TransE 等。

除此之外，知识图谱中的实体之间的多步关系路径也蕴含了丰富的语义信息。 比如(小明，出生地，山东)，(山东，位于，中国)隐含了实体小明和实体中国之间的“国籍”关系。为突破现有 TransE 等模型孤立学习每个三元组的局限性，考虑关系路径对知识表示学习的帮助，方法[Lin et al, 2015]以 TransE 作为基础进行扩展，提出 Path-based TransE（PTransE）模型，将知识图谱中的关系路径融入到知识表示学习模型中。考虑到并不是所有关系路径都是可靠并且对知识表示学习是有意义的，该方法提出关系路径的置信度，设计了基于路径约束的资源分配算法来选择关系路径。同样，[Xie et al, 2018] 提出了一个新的基于噪声的置信度感知知识表示学习框架(CKRL)。更具体地说，CKRL 模型遵循了 TransE 提出的基于翻译的框架，以三元组置信度的方式学习知识表示。同时考虑了局部路径信息和全局路径信息，提出了三种三元组置信度。为了使三元组置信度具有更好的通用性和灵活性，该方法只考虑知识图谱内部的结构信息以获得更好的全局一致性。

3）错误属性值检测

属性值的错误问题与关系的错误类似，传统方法采用统计离群值检测技术，发现与相关数据分布不相符的离群值作为可能的错误[肖仰华 et al, 2020]。但这种方法容易受到异常值的影响[Wienand et al, 2014]，可能的解决方案是使用多种相互独立的异常值检测方法来检测错误值[Fleischhacker et al, 2014] 。当然，也可以利用互联网等外部信息源来发现与纠正错误属性值，但为了提升搜索的召回率，需要设计不同的查询词来查找期望得到的属性值[Liu et al, 2015]。

3. 过期知识的检测与更新

知识图谱的更新问题是知识图谱的长期质量维护问题。传统方法可以采用定期全局更新机制，即设定一个时间跨度（如 1 个月）将知识图谱的所有内容进行一次全面的重新构建，然而显然这种方法的更新代价极大[肖仰华 et al, 2020]。之后，DBpedia 提出了基于更新流的改良方案[Hellmann et al, 2009]：每当维基百科产生一个更新，都将产生一条记录并主动推送给 DBpedia 进行更新[肖仰华 et al, 2020]。然而，提供更新流的数据源很少，这种被动更新的机制往往难以实施。近年来，局部更新机制成为主流，即每次只更新知识图谱的局部知识，其关键挑战在于如何识别出发生了变化的知识[肖仰华 et al, 2020]。大量的研究围绕变化知识的识别或者知识更新的预测展开，大体可分为以下几类：基于更新频率预测的更新机制、基于时间标签的更新机制、以及基于热点事件发现的更新机制。

1）基于更新频率预测的更新机制

理想状况下，如果我们知道每一种知识的更新频率，再根据更新频率来设定每一种知识的更新机制，则可以实现最小代价的更新。然而，知识图谱构建方往往无法获取完整的更新日志，因此这种机制需要根据有限的采样观测来准确估计知识的更新频率，从而实现主动局部更新[Cho et al, 2003]。然而，上述方法只有在观测频率不低于知识更新频率时才能做出较为准确的评估。如果知识更新频率比我们的观测频率要高，则观测到的更新次数比实际的更新次数要少，那么该方法得到的估计量也比实际的值小，我们对值的评估也就不再准确。为了应对这一状况，可以通过分析的期望值来近似拟合出一个更为精确的值。此外，如果可以得知每次观测前知识发生更新的最后时间，可以根据这一信息进一步提升更新频率估计的准确度，这些具体改进方法详见文献[Cho et al, 2003]。

2） 基于时间标签的更新机制

此类更新机制[Jiang et al, 2016]主要利用事实间的时序关系来预测即将更新的事实知识。例如，对于一个人，与其相关的事实存在以下时间序列关系：出生→求学→工作→死亡。较早（较晚）发生的关系被称为先验（后续）关系[肖仰华 et al, 2020]。用于发现事实间的时序关系的有两种：一种是基于时间序列信息的时间感知模型，其结合事实发生的时间，将时

序信息在特定向量空间进行表示学习，使得训练得到的向量表示能够自动分离先验关系和后续关系，从而确定事件间的时序关系；另一种则利用时间相关的语义约束作为整数线性规划的约束，构造相应的推理模型[肖仰华 et al, 2020]。此类模型考虑了以下三种约束：（1）时间分离约束，即具有相同头实体和相同函数关系的任意两个事实的时间间隔是不重叠的。比如，一个人在同一时间段内只能是一个人的配偶；（2）时间顺序约束，即对于某些时间顺序关系，一个事实总是先于另一个事实发生。比如，一个人必须在他毕业之前出生；（3）时间跨度约束，即某一事实在知识图谱的时间范围之外的其他时间段内无效。比如，奥巴马在2009—2017 年期间担任美国总统，而从 2017 年起他不再是美国总统[肖仰华 et al, 2020]。以上两种模型可互为补充。此类基于时间标签的更新机制不仅能对时间敏感的数据做出较为准确的更新预测，还广泛应用于知识图谱查错等领域，但其适用范围仅限于时间敏感的数据。

3）基于热点事件发现的更新机制

基于热点事件发现的更新机制[Liang et al, 2017]主要适用于通用知识图谱，即认为知识图谱中需要更新的实体往往会被社会大众关注到，成为热搜，即会伴随一些热点事件或热词出现。因此，该机制提出对互联网上的热词进行实时监控，识别出热门实体并将其百科页面信息同步到知识库中[肖仰华 et al, 2020]。此外，文献[Acosta et al, 2013]也提出了一种根据百科各种语义特征构建的词条更新频率预测器。该机制可以以较高的准确率预测出需要更新的实体，从而以较低的代价对知识图谱进行实时更新。

四、问题与展望

总体而言，知识图谱质量评估与管理任务目前还处于起步阶段，仍有大量需要解决的问题亟待解决。具体而言：

在知识图谱的质量评估指标方面，现有的评估度量维度已经比较丰富，每种度量都刻画了数据质量的一个特点，然而评估度量指标的设计与最终确保图谱质量的目标仍然存在一定差距，其原因在于质量度量与构建过程的分裂，即度量结果与质量原因，无法自动或半自动的对图谱构建过程产生影响。另外，图谱的核心质量属性，如图谱的正确性和一致性，无法通过廉价方式获得。其他方面数据质量的提升，虽然重要但不本质。从长远来说，利用多源数据构建大规模领域语言模型，交叉验证图谱的深层质量属性，再辅以人工校验，可能是可行的一个方向。

在知识图谱的质量评估流程和方法方面，针对不同的评估度量维度存在不同的度量方法。整体上看，基于符号规则及数学统计的方法问题在于效率瓶颈，基于知识图谱表示学习的方法问题在于欠缺考虑除了知识图谱本身结构之外的语义信息，而基于传统证据搜索的方法问

题在于外部证据的覆盖率不能保证每一条细粒度的三元组评估需求。在未来，如何结合神经符号推理、大规模预训练语言模型相关的最新技术，提升质量评估过程中的特征学习效果和效率是值得探索的方向。

特别地，各类垂域智能化应用场景（金融、医疗、工业制造等）对知识图谱的质量要求往往更高。而领域当中实体和关系类型的非规范性以及上下文特性，导致人工都难以评估。因此，需要首先明确定义图谱使用上下文需求，才能给出领域图谱的质量定义和评估，而评估重点应放在依据而不是量化的结果上。对垂域知识图谱的质量评估与管理目前还主要依赖于专家的手动维护，效率低、成本高。未来希望可以有较为完善的、可操作性强的框架、工具及系统来对知识图谱质量控制降本增效[肖仰华 et al, 2020]。

参考文献

[Mendes et al, 2012] Mendes P N, Mühleisen H, Bizer C. Sieve: linked data quality assessment and fusion. Joint EDBT/ICDT Workshops 2012: 116-123.

[Paşca et al, 2006] Paşca, Marius, Lin D, Bigham J, et al. Names and similarities on the web: fact extraction in the fast lane. ACL 2006: 809-816.

[Paulheim et al, 2013] Paulheim H, Bizer C. Type inference on noisy rdf data. ISWC 2013: 510-525.

[Liang et al, 2017] Liang J, Xiao Y, Wang H, et al. Probase+: Inferring Missing Links in Conceptual Taxonomies. IEEE Transactions on Knowledge & Data Engineering. 29(6):1281-1295(2017).

[Suzuki et al, 2018] Suzuki M, Matsuda K, Sekine S, et al. A Joint Neural Model for Fine-Grained Named Entity Classification of Wikipedia Articles. J. Ieice Transactions on Information & Systems. 101(1):73-81(2018).

[Richardson et al, 2006] Richardson M, Domingos P. Markov logic networks. Machine Learning. 62(1-2):107-136(2006).

[Chen et al, 2014] Chen Y, Wang D Z. Knowledge expansion over probabilistic knowledge bases. ACM SIGMOD 2014: 649-660.

[Lao et al, 2010] Lao N, Cohen W W. Relational retrieval using a combination of path-constrained random walks. Machine Learnin. 81(1):53-67(2010).

[Gardner et al, 2015] Gardner M, Mitchell T. Efficient and expressive knowledge base completion using subgraph feature extraction. EMNLP 2015: 1488-1498.

[Bordes et al, 2013] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for

modeling multi-relational data. NeuIPS 2013: 2787-2795.

[Wang n et al, 2014] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes. AAAI 2014: 1112-1119.

[Razniewski et al, 2018] Razniewski S, Suchanek F, Nutt W. But What Do We Actually Know? workshop on AKBC 2018: 40-44.

[Moser et al, 1992] Moser W, Adlassnig K P. Consistency checking of binary categorical relationships in a medical knowledge base. Artificial Intelligence in Medicine. 4(5):389-407 (1992).

[Dong et al, 2014] Dong X, Gabrilovich E, Heitz G et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. (2014).

[Paulheim et al, 2014] Paulheim H, Bizer C. Improving the quality of linked data using statistical distributions. IJSWIS. 10(2): 63-86 (2014).

[Li et al, 2012] Li Z, Sharaf M A, Sitbon L, et al. WebPut: Efficient Web-Based Data Imputation. Lecture Notes in Computer Science. 7651:243-256 (2012).

[Paulheim et al, 2015] Paulheim H, Gangemi A. Serving DBpedia with DOLCE–more than just adding a cherry on top. ISWC 2015: 180-196.

[Wienand et al, 2014] Wienand D, Paulheim H. Detecting incorrect numerical data in dbpedia. ESWC 2014: 504-518.

[Fleischhacker et al, 2014] Fleischhacker D, Paulheim H, Bryl V, et al. Detecting errors in numerical linked data using cross-checked outlier detection. ISWC 2014: 357-372.

[Liu et al, 2015] Liu S, d'Aquin M, Motta E. Towards Linked Data Fact Validation through Measuring Consensus. ESWC 2015.

[Hellmann et al, 2009] Hellmann S, Stadler C, Lehmann J, et al. DBpedia live extraction. OTM 2009: 1209-1223.

[Cho et al, 2003] Cho J, Garcia-Molina H. Estimating frequency of change. TOIT. 3(3): 256-290 (2003).

[Jiang et al, 2016] Jiang T, Liu T, Ge T, et al. Towards time-aware knowledge graph completion. COLING. 2016: 1715-1724.

[Liang et al, 2017] Liang, J.,Zhang, S. & Xiao, Y. How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source. IJCAI. 2017: 3749-3755.

[Acosta et al, 2013] Acosta M, Zaveri A, Simperl E, et al. Crowdsourcing linked data quality assessment. 2013: 260-276.

[Nathani et al, 2019] Nathani D, Chauhan J, Sharma C, et al. Learning Attention-based Embeddings for Relation Prediction in Knowledge Graphs. ACL 2019: 4710-4723.

[Chami et al, 2020] Chami I, Wolf A, Juan D C, et al. Low-Dimensional Hyperbolic Knowledge Graph Embeddings. ACL 2020: 6901-6914.

[Zhu et al, 2021] Zhu C, Chen M, Fan C, et al. Learning from History: Modeling Temporal Knowledge Graphs with Sequential Copy-Generation Networks. AAAI 2021: 4732-4740.

[Fu et al, 2013] Fu C, Li Z, Yang Q, et al. Multiple Interaction Attention Model for Open-World Knowledge Graph Completion. 2019.

[Bayram et al, 2021] Bayram E, García-Durán A, West R. Node attribute completion in knowledge graphs with multi-relational propagation. ICASSP 2021: 3590-3594.

[Ren et al, 2016] Ren X, He W, Qu M, et al. Afet: Automatic fine-grained entity typing by hierarchical partial-label embedding. EMNLP 2016: 1369-1378.

[Shimaoka et al, 2017] Shimaoka S, Stenetorp P, Inui K, et al. Neural Architectures for Fine-grained Entity Type Classification. EACL 2017: 1271-1280.

[Abhishek et al, 2073] Abhishek A, Anand A, Awekar A. Fine-grained entity type classification by jointly learning representations and label embeddings. EACL 2017: 797-807.

[Chen et al, 2020] Chen T, Chen Y, Van Durme B. Hierarchical Entity Typing via Multi-level Learning to Rank. ACL 2020: 8465-8475.

[Ren et al, 2020] Ren Q. Fine-grained entity typing with hierarchical inference. ITNEC 2020: 2552-2558.

[Rabinovich et al, 2017] Rabinovich M, Klein D. Fine-Grained Entity Typing with High-Multiplicity Assignments. ACL 2017: 330-334.

[Xiong et al, 2019] Xiong W, Wu J, Lei D, et al. Imposing Label-Relational Inductive Bias for Extremely Fine-Grained Entity Typing. NAACL 2019: 773-784.

[Lin et al, 2019] Lin Y, Ji H. An attentive fine-grained entity typing model with latent type representation. EMNLP-IJCNLP 2019: 6197-6202.

[Ma et al, 2016] Ma Y, Cambria E, Gao S. Label embedding for zero-shot fine-grained named entity typing. COLING 2016: 171-180.

[Yuan et al, 2018] Yuan Z, Downey D. Otyper: A neural architecture for open named entity typing. AAAI 2018: 260-270.

[Zhang et al, 2020] Zhang T, Xia C, Lu C T, et al. MZET: Memory Augmented Zero-Shot Fine-

grained Named Entity Typing. COLING 2020: 77-87.

[Ren et al, 2020] Ren Y, Lin J, Zhou J. Neural zero-shot fine-grained entity typing. ISWC 2020: 846-847.

[Onoe et al, 2019] Onoe Y, Durrett G. Learning to Denoise Distantly-Labeled Data for Entity Typing. NAACL 2019: 2407-2417.

[Janowicz et al, 2018] Janowicz K, Bo Y, Regalia B, et al. Debiasing Knowledge Graphs: Why Female Presidents are not like Female Popes. ISWC 2018.

[Wang et al, 2014] Zhen Wang, Jianwen Zhang, Jianlin Feng and Zheng Chen. Knowledge Graph Embedding by Translating on Hyperplanes. AAAI 2014: 1112–1119.

[Lin et al, 2015] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu and Xuan Zhu. Learning Entity and Relation Embeddings for Knowledge Graph Completion. AAAI 2015: 2181–2187.

[Ji et al, 2015] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu and Jun Zhao. Knowledge Graph Embedding via Dynamic Mapping Matrix. ACL 2015: 687–696.

[Socher et al, 2013] Richard Socher, Danqi Chen, Christopher D. Manning and Andrew Y. Ng. Reasoning With Neural Tensor Networks for Knowledge Base Completion. NeuIPS 2013: 926–934.

[Xie et al, 2016] Ruobing Xie, Zhiyuan Liu and Maosong Sun. Representation Learning of Knowledge Graphs with Hierarchical Types. IJCAI 2016: 2965–2971.

[Guo et al, 2018] Shu Guo, Quan Wang, Lihong Wang, Bin Wang and Li Guo. Knowledge Graph Embedding With Iterative Guidance From Soft Rules. AAAI 2018: 4816–4823.

[Ali et al, 2020] Ali M A, Sun Y, Li B, et al. Fine-grained named entity typing over distantly supervised data based on refined representations. AAAI 2020: 7391-7398.

[Del, 2015] Del Corro L, Abujabal A, Gemulla R, et al. Finet: Context-aware fine-grained named entity typing. EMNLP 2015: 868-878.

[Dai et al, 2019] Dai H, Du D, Li X, et al. Improving Fine-grained Entity Typing with Entity Linking. EMNLP-IJCNLP 2019: 6210-6215.

[Liu et al, 2021] Liu Q, Lin H, Xiao X, et al. Fine-grained Entity Typing via Label Reasoning. EMNLP 2021: 4611-4622.

[Onoe et al, 2021] Onoe Y, Boratko M, McCallum A, et al. Modeling Fine-Grained Entity Types with Box Embeddings. ACL 2021: 2051-2064.

[Yuan et al, 2023] Yuan S, Yang D, Liang J, Li Z, et al. Generative Entity Typing with Curriculum Learning. EMNLP 2023 (Under Review)

[Hong et al, 2020] Yan Hong, Chenyang Bu and Tingting Jiang. Rule-enhanced Noisy Knowledge Graph Embedding via Low-quality Error Detection. ICKG 2020: 544–551.

[Cheng et al, 2018] Yurong Cheng, Lei Chen, Ye Yuan and Guoren Wang. Rule-Based Graph Repairing: Semantic and Efficient Repairing Methods. ICDE 2018: 773–784.

[Ho et al, 2018] Vinh Thinh Ho, Daria Stepanova, Mohamed H. Gad-Elrab, Evgeny Kharlamov and Gerhard Weikum. Rule Learning from Knowledge Graphs Guided by Embedding Models. ISWC 2018: 72–90.

[Zhang et al, 2019] Wen Zhang, Bibek Paudel, Liang Wang, Jiaoyan Chen, Hai Zhu, Wei Zhang, Abraham Bernstein and Huajun Chen. Iteratively Learning Embeddings and Rules for Knowledge Graph Reasoning. WWW 2019: 2366–2377.

[Belth et al, 2020] Caleb Belth, Xinyi Zheng, Jilles Vreeken and Danai Koutra. What is Normal, What is Strange, and What is Missing in a Knowledge Graph: Unified Characterization via Inductive Summarization. WWW 2020: 1115–1126.

[Lin et al, 2015] Yankai Lin, Zhiyuan Liu, Huan-Bo Luan, Maosong Sun, Siwei Rao and Song Liu. Modeling Relation Paths for Representation Learning of Knowledge Bases. EMNLP 2015: 705–714.

[Xie et al, 2018] Ruobing Xie, Zhiyuan Liu, Fen Lin and Leyu Lin. Does William Shakespeare REALLY Write Hamlet? Knowledge Representation Learning With Confifidence. AAAI 2018: 4954–4961.

[Wang et al, 1996] Wang R Y, Strong D M. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. 12(4): 5-33 (1996).

[Naumann et al, 2002] Naumann F. Quality-driven query answering for integrated information systems [M]. Springer, 2002.

[Zaveri et al, 2015] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked data: A survey. Semantic Web. 7(1): 63–93 (2015).

[Mecella et al, 2002] M. Mecella, M. Scannapieco, A. Virgillito, R. Baldoni, T. Catarci, and C. Batini. Managing data quality in cooperative information systems. OTM 2002: 486–502.

[Paulheim et al, 2017] Paulheim H, Cimiano P. Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web. 8(3): 489-508 (2017).

[Bizer et al, 2007] Bizer C. Quality-Driven Information Filtering-In the Context of Web-Based Information Systems. Calvin Edu, 2007.

[Kaffee et al, 2017] Kaffee L A , Piscopo A , Vougiouklis P , et al. A Glimpse into Babel: An Analysis of Multilinguality in Wikidata. Opensym 2017: 1-5.

[Luggen et al, 2019] M. Luggen, D. Difallah, C. Sarasua, G. Demartini, P. Cudr´eMauroux, Non-parametric class completeness estimators for collaborative knowledge graphsthe case of wikidata. ISWC 2019: 453– 469.

[F¨arber et al, 2018] M. F¨arber, F. Bartscherer, C. Menne, A. Rettinger, Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. Semantic Web. 9: 77–129 (2018).

[Soulet et al, 2018] Soulet A, Giacometti A, Béatrice Markhoff, et al. Representativeness of Knowledge Bases with the Generalized Benford's Law. ISWC. 2018.

[BEHKAMAL et al, 2014] BEHKAMAL, Behshid, KAHANI, et al. A Metrics-Driven Approach for Quality Assessment of Linked Open Data. Journal of theoretical and applied electronic commerce research, 2014.

[Flemming et al, 2011] Annika Flemming. Qualitätsmerkmale von Linked Data-veröffentlichenden Datenquellen. Diplomarbeit https://cs.uwaterloo.ca/~ohartig/files/ DiplomarbeitAnnikaFlemming.pdf, 2011.

[Ricardo et al, 2018] Ricardo Baeza-Yates. Bias on the Web. Communications of the ACM, 61(6):54–61 (2018).

[Hogan et al, 2010] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the Pedantic Web. WWW Workshop 2010: 628.

[Guns et al, 2013] R. Guns. Tracing the Origins of the Semantic Web. Journal of the American Society for Information Science and Technology. 64(10): 2173–2181 (2013).

[Bechhofer et al, 2016] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, and P. F. Patel-Schneider. OWL Web Ontology Language Reference. https://www.w3.org/TR/2004/REC-owl-ref- 20040210, 2004. [Online; accessed 06-Apr-2016].

[Pipino et al, 2002] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. Communications of the ACM. 45(4): 211–218 (2002).

[Hogan et al, 2020] Hogan A, Blomqvist E, Cochez M, et al. Knowledge Graphs. 2020.

[Darari et al, 2018] Fariz Darari, Werner Nutt, Giuseppe Pirrò, and Simon Razniewski. Completeness Management for RDF Data Sources. ACM Transactions on the Web. 12(3): 1–53 (2018).

[Jain et al, 2010] Jain P, Hitzler P, Yeh P Z, et al. Linked Data Is Merely More Data. AAAI 2010.

[Heath et al, 2011] Heath T, Bizer C. Linked Data: Evolving the Web into a Global Data Space. Molecular Ecology. 11(2): 670–684 (2011).

[Hogan et al, 2012] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of linked data conformance. Web Semantics: Science, Services and Agents on the World Wide Web. 14–44 (2012).

[Auer et al, 2010] Sören Auer, Matthias Weidl, Jens Lehmann, Amrapali Zaveri, and Key-Sun Choi. ISWC 2010: 1-16.

[Gayo et al, 2012] Jose Emilio Labra Gayo, Dimitris Kontokostas, and Sören Auer. Multilingual linked open data patterns. Semantic Web Journal. (2012).

[Jain et al, 2013] P. Jain, P. Hitzler, K. Janowicz, and C. Venkatramani. There's No Money in Linked Data[J]. http://corescholar. libraries.wright.edu/cse/240, 2013.

[Ruan et al, 2018] Tong Ruan, Liang Zhao, Yang Li, Haofen Wang, Xu Dong. On Evaluating Web-Scale Extracted Knowledge Bases in a Comparative Way. Semantic Web Inf. Syst. 14(1): 98-120（2018）.

[Ruan et al, 2016] Ruan, T., Li, Y., Wang, H., Zhao, L. From Queriability to Informativity, Assessing "Quality in Use" of DBpedia and YAGO. ESWC 2016: 52-68.

[Galárraga et al, 2016] L Galárraga, Razniewski S, Amarilli A , et al. Predicting Completeness in Knowledge Bases. WSDM 2016: 375-383.

[Fürber et al, 2011] C Fürber, Hepp M. SWIQA - A Semantic Web information quality assessment framework. ECIS 2011.

[Trouillon et al, 2016] Trouillon T P, Bouchard G M. COMPLEX EMBEDDINGS FOR SIMPLE LINK PREDICTION. ICML 2016: 2071-2080.

[Bougiatiotis et al, 2020] Bougiatiotis K, Fasoulis R, Aisopos F , et al. Guiding Graph Embeddings using Path-Ranking Methods for Error Detection innoisy Knowledge Graphs. 2020.

[Jia et al, 2019] Shengbin Jia, Yang Xiang, Xiaojun Chen, Kun Wang, and Shijia. Triple Trustworthiness Measurement for Knowledge Graph. WWW 2019: 2865–2871.

[Esteves et al, 2017] Diego Esteves, Anisa Rula, Aniketh Janardhan Reddy, and Jens Lehmann. 2018. Toward Veracity Assessment in RDF Knowledge Bases: An Exploratory Analysis[J]. Data and Information Quality. 9(3): 1-26 (2017).

[Gerber et al, 2015] Daniel Gerber, Diego Esteves, Jens Lehmann, Lorenz Bühmann, Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, and René Speck. DeFacto-Temporal and multilingual Deep

Fact Validation. 35(2): 85–101 (2015).

[Syed et al, 2018] Zafar Habeeb Syed, Michael Röder, and Axel-Cyrille Ngonga Ngomo. FactCheck: Validating RDF Triples Using Textual Evidence. ACM CIKM 2018: 1599–1602.

[Gerber et al, 2012] Daniel Gerber and Axel-Cyrille Ngonga Ngomo. Extracting MultilingualNatural-Language Patterns for RDF Predicates. EKAW 2012: 87-96.

[肖仰华 et al, 2020] 肖仰华 et. al. 知识图谱: 概念与技术. 电子工业出版社, 2020.

第十一章 基于知识的问答与对话

何世柱，张元哲，刘康

中国科学院自动化研究所 模式识别国家重点实验室，北京 100190

一、任务定义、目标和研究意义

问答系统（Question Answering）和**对话系统（Dialogue System）**是知识图谱典型应用场景，它们能够接受自然语言形式描述的用户需求，通过问句分析、知识获取、知识推理、答复生成等操作自动得到精确答案，是新一代信息检索系统的重要形式 [Etzioni and Oren 2011]。**问答系统**是指让计算机自动回答用户所提出的问题，是信息服务的一种高级形式。不同于现有的搜索引擎，问答系统返回用户的不再是基于关键词匹配的相关文档排序，而是精准自然语言形式的答案。**对话系统**是能够以自然语言与人类进行对话的计算机系统，它旨在让计算机能够"听懂"人类语言，并对于用户所提的消息返回准确、流利、一致的回复，甚至完成特定的操作 [赵军 et al. 2022]。问答是信息获取的高效方式，对话是人机交互的自然形式，问答可以看作一种需求明确的单轮对话。知识图谱是问答和对话系统的重要基础，其包含大量描述精准的结构化语义内容，有助于问答和对话系统为用户提供精准的知识服务。

问答与系统系统一直伴随的人工智能技术的发展。一方面，问答与对话能够用来评估计算机系统的语言理解能力和智能发展程度，自图灵测试提出以来，它就是人工智能和自然语言处理的长远目标之一；另一方面，相较于传统搜索引擎只能返回一个网页集合，问答与对话系统可以用自然语言直接与用户进行交互，并且在多轮对话中通过答复、确认、反馈等操作完成信息获取，不再需要仔细浏览搜索引擎返回的多个冗余信息。近些年，问答系统更是取得一系列倍受关注的成果。2011 年，IBM Watson 自动问答机器人在美国智力竞赛节目 Jeopardy 中战胜人类选手，在业内引起了巨大的轰动。同年，苹果公司在其智能手机中集成了移动生活助手 Siri，它能够以语音交互的方式帮助人们完成预定闹钟、查询天气等任务，引起了人们对问答系统广泛应用的无线遐想。2014 年开始，微软在中国推出了一个拟人化的智能聊天机器人，能够与人类就任何话题进行情感化对话。随着人工智能技术的突飞猛进，以及问答与对话系统在精准营销、情感陪护、智能教育等方面有着巨大商业价值，各大 IT 巨头更是相继推出以问答系统为核心技术的产品和服务，如移动生活助手（Siri、Google Now、Cortana、小冰等）、智能音箱（Echo、叮咚音箱、公子小白、百度度秘、天猫精灵等）等，这似乎让人们看到了黎明前的阳光，甚至认为现有的问答与对话技术已经十分成熟。特别是近些年，随着人工智能热潮到来，无论是学术界还是产业界，都给予其极大关注和投入。因

此，对其开展研究具有非常重要的学术和实际意义。

尽管 IBM Watson 系统在 Jeopardy 中战胜了人类选手，但是其核心并没有突破传统基于“检索 + 匹配”和“检索 + 抽取”的技术模式，缺乏对于文本语义深层次的分析和处理，难以实现知识的深层逻辑推理，无法达到人工智能的高级目标。Watson、Siri、Echo 等应用的成功也已经被证明仅仅局限于限定领域、特定类型的问题，离语义的深度理解以及智能问答还有很大的距离；微软小冰等开放域对话系统也存在类似的问题，不能够做到举一反三和良好的泛化能力。因此，面对已有问答与对话系统的不足，为了提升信息服务的准确性与智能性，研究者近些年逐步把目光投向知识图谱（Knowledge Graph），鉴于知识图谱以规范化、体系化、规模化方式描述各领域的通用知识内容，研究人员希望知识图谱的引入能够提高问答与系统的智能化水平，能够更好地做到举一反三、可解释等能力。本章把基于知识图谱的问答与对话系统分别称为知识问答（Knowledge Graph based Question Answering）和知识对话（Knowledge Powered Dialogue System）。

二、 研究内容和关键科学问题

知识问答与对话是一类需要综合利用自然语言处理、知识工程、规划、决策等多个领域技术的综合性计算机应用系统，在响应用户需求的时候，系统首先需要正确理解用户所提的自然语言问题，在感知交互环境下抽取其中的关键语义信息，然后在已有知识系统中通过匹配、检索、推理等手段获取相关知识内容，最后生成满足背景知识和应用场景的答复内容返回给用户。其中所涉及的关键技术包括：语义分析、信息检索、知识推理、语言生成等。传统问答与对话系统多集中在限定领域，针对特定类型的简单需求进行答复。然而伴随大规模数据的积累和深度学习技术的快速发展，已有知识图谱的规模在不断增大，建模知识类型不断延伸，所涉及的领域不断增多。现有研究趋于向处理更复杂类型问题、更深层次知识推理、更自然友好交互形式等问答与对话系统构建。总体来讲，主要面临如下三个关键科学问题。

1. 问句语义解析

在利用知识图谱作为知识载体响应用户需求的时候，由于知识图谱与用户需求在知识结构的组织和知识内容的表达上存在较大差异，因此将用户描述的自然语言问句转换为知识图谱可以接受形式是首要科学问题。问句语义解析就是完成上述目标的任务，它把问句映射为确定性的形式化语义表示的过程，这种形式化语义表示可以被计算机理解和识别，从而转化为可执行的形式化查询语句，进而在知识图谱管理系统中执行该查询语句就能够得到答案返回给用户。例如，对于自然语言问句“路遥写的哪本书获得了矛盾文学奖？”，语义解析模型需要转换为“λx. 作者 (x, 路遥 _ 中国当代作家), 获奖 (x, 矛盾文学奖). ”这类形式化

语句。具体过程需要分析问句中的词、短语等语义单元与知识图谱中的实体、概念、关系等语义单元进行链接，通过分析问句语义单元之间的关系，将知识图谱中对应的实体、概念、关系进行组合，形成面向目标知识图谱的形式化语义表示形式。实现上述目标不仅需要利用词法分析、句法分析，还需要实体链接、关系预测等技术。传统问答语义解析面向单一领域的小规模知识图谱，所涉及的语言词汇和表达方式有限，对应的实体、概念、关系规模较小，通常可以采取人工定义映射辞典、组合模板等方式进行语言的理解和语义的组合，或者利用SVM、对数线性模型等传统机器学习方法训练语义解析模型。但是，面对大规模、多领域知识图谱，不仅需要面对更加开放的用户问句表达，并且随着实体、概念、关系规模的继续增大，语义解析算法的复杂度也呈指数增加，因此，传统依赖人工模板规则和统计机器学习方法面临着句法挑战。近年来，随着大规模语义解析标注数据集的构建和深度学习模型的发展，人们开始利用 Seq2Seq、强化学习、Transformer 等模型与技术实现端到端的语义解析，基于编码器-解码器框架，这类神经语义解析方法以黑盒方式直接将用户问句自动转换为目标形式化语句。尽管神经语义解析自动生成形式语言的序号序列，但是他们依然面临多方面挑战。首先，与机器翻译、自动摘要等序列生成任务不同，形式语言具有确定性的语法，其他表达不仅具有层次性，还具有明显的结构；其次，大量形式语言表达复杂且存冗余的符号表达；再者，问句语义的理解非常依赖目标知识图谱，相同表达在不同知识图谱下语义完全不同；最后，神经语义模型作为数据驱动的方法难以处理领域泛化和组合泛化问题。因此，如何解决上述问题实现高质量问句语义解析是知识问答与对话的首要科学问题。

2. 大规模知识推理

知识图谱的不完备性阻碍了知识问答与对话的应用，人们在构建知识资源的时候难免会利用常识、领域等背景知识，因此，现有知识图谱中有大量知识未显式表示，而这些隐含的知识在回答用户问题时候至关重要。例如，知识图谱中描述了一个人的工作“所在地”信息，但是未对其“出生地”进行描述（虽然知识图谱的 Schema 中约定了人物类型实体包含“出生地”属性，但大量实体缺乏“出生地”属性的描述，这种不完备广泛存在于当前知识图谱中），即无法直接回答诸如“某某人出生于哪里?”这样的问题。但是一般情况下，在中国，一个人的“父母”的“所在地”所属的国家就是他/她的“出生地”。这些隐含知识天然存在于人的知识体系中，但在当前知识图谱中，由于原始数据不完备、人工构建时遗漏信息、知识获取技术不足等问题，所构建的知识图谱中存在大量内容缺失的现象。因此，在利用知识图谱进行知识服务过程中，就需要通过知识推理来实现隐含知识的发现以支撑更智能化的问答和对话。长期以来，推理的研究主要关注基于符号匹配和逻辑演算的推理模式，例如，一阶谓词逻辑、描述逻辑等。但是，基于符号表示的知识推理完全依赖于符号间的严格匹配，难以

克服符号间语义鸿沟的影响，不便于知识的泛化和推理的大规模计算。近年来，以知识图谱表示学习为核心的推理模式受到越来越多的关注，人们直接利用原始图谱数据基于深度神经网络和自监督学习方式得到符号的数值化表示，符号之间的关联蕴含在数值空间中，通过数值计算能够利用快速计算平台处理大规模数据。尽管当前方法取得了不错的效果，但是在真实问答和对话场景中依然面临不少问题，首先，大规模知识图谱涉及的实体、关系、概念非常多，知识推理中采用的符号逻辑方法计算效率较低和表示学习方法推理精度不足，无法在大规模知识图谱中实现即快速计算又推理精准；其次，当前很多方法仅仅关注独立地运用知识图谱和语言文本，没有充分利用它们各自的优势，以至于在事实分类、链接预测等推理任务中难以取得较好性能；最后，真实应用往往需要处理持续出现的新概念、新事实等新知识，知识推理在增量学习中常常会遗忘原有知识。因此，如何解决上述问题实现大规模高效率的精准知识推理是支撑问答和对话的关键科学问题。

3. 融合知识图谱的文本生成

在利用知识图谱进行知识服务过程中，普通用户乐于接受更友好的交互形式，即问答和对话系统不仅需要接受用户使用自然语言表达的意图，还需要提供自然语言形式的答复内容，该目标的实现需要利用融合知识图谱的文本生成技术。例如，对于问题“洛国富是哪儿人”，系统需要生成“洛国富出生于巴西，他现在是中国人。”这类文本型答复返回给用户。为了回答上述问题并生成文本答复，系统不仅需要识别实体词“洛国富”，还通过实体链接和知识匹配从知识图谱中搜索相关内容，并融合利用所对应的“出生地”和“国籍”等知识内容生成流利的自然语言文本。生成自然语言文本作为答复形式是对话系统的长期目标，早期为了通过图灵测试研发的对话系统和近年来发展迅猛的聊天机器人，都是针对用户提出的消息（包括疑问句、陈述句等多种形式）返回流利自然的回复。早期的系统首先从原始消息中提取关键词，然后利用消息-回复模板，采取信息槽填充的方式最终合成自然语言的回复。近年来，随着机器学习的发展和网络大规模数据的充沛，人们利用深度神经网络模型从大规模自然语言的消息-回复数据集中自动学习对话模式，虽不再需要人工构建显式的模板，但是这类系统仍然只能进行闲聊式交互，难以提供有内容的知识性回复。因此，如何通过融合知识图谱的文本生成，兼顾答复内容的丰富性和语言表达的多样性，是实现问答和对话更广泛应用的另一个科学问题。

三、技术方法和研究现状

近年来，知识问答与对话技术发展迅猛，下面根据技术路线的不同分别介绍知识问答和知识对话的技术方法和研究现状。对于知识问答，主要包括：基于语义解析的知识库问答方

法、基于检索排序的知识库问答方法，以及融合知识图谱的文本问答方法这三类技术路线。对于知识对话，主要包括：基于对话语义理解的任务型对话方法和融合知识图谱的生成式对话方法两类技术路线。

1．基于语义解析的知识库问答方法

语义解析方法则将用户的问句转译为形式化的查询语言。代表性的语义解析技术有归纳逻辑编程、同步上下文无关语法和依存组合语义等。近年来，基于组合范畴语法的方法和基于深度学习的方法较为突出。组合范畴语法使用句法范畴来组合语义，在语言学研究中有很深的历史渊源，这种方式可以有效处理如长距离依赖 [Steedman 1997] 等自然语言处理的难点，被很多语义解析工作 [Zettlemoyer and Collins 2005, Misra and Artzi 2016] 所采用。此外，使用现成的句法解析结果作为语义解析的输入 [Ge and Mooney 2009] 也是研究热点之一，有助于在低资源语言上实现语义解析。

基于深度学习的方法是语义解析的最新方法，它以黑盒方式直接将用户问句输入映射为目标形式化语言。为了保证形式语言的合法性，相关研究均为神经网络添加不同的语法约束、语义约束 [Ge and Mooney 2009, Dong and Lapata 2016, Yin and Neubig 2018]。但当问题趋于复杂时，一步到位生成目标语言较为困难，Hu 等人 [Hu et al. 2018] 以语义查询图为媒介，运用基于转移的方法生成操作序列构造语义图，最后使用子图匹配或生成 SparQL 的方法即可在知识库上查询答案。除了语义图之外，也有工作专注于使用模板作为中间媒介，包括手工设计模板 [Jia et al. 2018]、动态生成模板 [Abujabal et al. 2017]、以及用持续学习的方法适应并获取无标注领域的新模板 [Abujabal et al. 2018] 等。另一方面，从问句端要识别复杂问题的语义组合关系也较为困难，Luo 等人 [Luo et al. 2018] 使用依存句法分析结果来增强问句的语义表示，并将依存结果用于语义子图的消歧。针对现有依存分析工具可能引入噪音的问题，Sun 等人 [Sun et al. 2020b] 定义了一种骨架依存语法，并额外标注了数据，以大规模预训练模型实现了该语法的各个分析步骤，所得的分析结果可以很容易地具化（grounding）到知识库上。受到上述复杂问句组合分析、目标语言媒介设计等既有工作的启发，有的工作更进一步，使用结构敏感的编码器表征目标实体以强化它与问句的联系 [Zhu et al. 2020]，或优先完成问句的抽象分析再约束形式语言生成 [Chen et al. 2020b]。Kapanipathi 等人 [Kapanipathi et al. 2021] 分别为问句和目标语言选择了 AMR 和一阶逻辑表示，由于 AMR 是领域无关的，他们通过实体链接、关系链接及基于路径的图转换模块得到逻辑表示，最终再转化为 SparQL 语句。Das 等人 [Das et al. 2021] 则另辟蹊径，通过从训练集中检索出有用的问句及其对应的 SparQL 语句，使用基于案例的推理方法，在面向复杂问题的数据集 ComplexWebQuestion 上取得了很好的结果，并可以泛化到未训练的实体或关系上。

2．基于检索排序的知识库问答方法

基于检索排序的知识库问答方法把知识库问答看做是一个语义匹配的过程。通过表示学习，基于检索的方法能够将用户的自然语言问题转换为一个低维空间中的数值向量（分布式语义表示），同时知识库中的实体、概念、类别以及关系也能够表示成为同一语义空间的数值向量。这样一来，知识库问答任务就可以看成问句语义向量与知识库中实体、边的语义向量相似度计算的过程。基于检索排序的知识库问答方法最初由 Bordes 等人在 2014 年提出[Bordes et al. 2014]，在此之后得到了持续和广泛的关注和研究。基于检索排序的方法回避了语义分析的难题，一般只需要问题-答案对的标注即可采用端到端的形式进行训练，泛化性强，有效缓解了规范语义表达式难以标注的问题，但是其缺点是可解释性较差 [Lan et al. 2021]。基于检索排序的方法核心在于问题和候选答案的表示，以及排序打分模块的设计。为此，一种思路是利用问题和知识库中实体之间的交互信息来增强表示，例如，Dong 等人[Dong et al. 2015]利用多列卷积神经网络、Hao 等人 [Hao et al. 2017] 利用问题和候选答案之间的交互注意力，来增强问题和候选实体表示。另一种思路是挖掘其他形式的信息来增强知识库中的实体表示，从相关文本语料中找到非结构化知识作为补充证据。例如，Han 等人[Han et al. 2020] 提出利用图神经网络，将检索到的文本信息融合到知识图谱实体表示中，以缓解知识图谱不完备的问题。Saxena 等人 [Saxena et al. 2020] 提出利用预训练的知识库表示来增强实体表示，以提高多跳问答的性能。近年来，更多的研究工作开始围绕回答复杂问题展开。He 等人 [He et al. 2021] 针对多跳问答缺少中间步骤监督信号的问题，提出一种教师-学生模型，其中，学生模型用于预测答案，教师模型则用于提供中间步骤监督信号，有效提升了多跳知识库问答的性能。在排序网络方面，主流技术已经从原来的简单打分网络，演化到图神经网络等复杂网络，这使得问句和答案候选的表示也和排序网络的优化高度融合在一起。

3．融合知识图谱的文本问答方法

在文本问答的过程中引入知识图谱中的结构知识，一直以来都是研究人员的关注重点。从机器阅读理解数据集角度，从 2018 年以来，已经有多个数据集涉及到常识问答，这些很多都是需要背景知识图谱的支撑。例如，CommonsenseQA[Talmor et al. 2019] 是一个多项选择题的常识问答数据集，每个问题都包含来自 ConceptNet[Liu and Singh 2004] 的一个实体，ConceptNet 是一个大型常识知识图谱，可以提供人类世界的常识信息，帮助机器回答问题。CosmosQA[Huang et al. 2019] 数据集包含 35600 个问题，其中约 94％需要常识，其专注于解决需要上下文的推理问题。Social IQA[Sap et al. 2019] 是首个面向社交常识推理的问答数据集，包含 38000 个覆盖日常社交情感的问题，其问题包括在特定情境中完成对人物行为

的推理等。OpenBookQA 包含大约 6000 个问题，需要结合科学事实或常识知识来回答。OpenBookQA 提供了约 1300 个科学事实的 “open book”，每个事实都与问题直接相关，期望模型能够合理使用这些知识来正确回答问题。

在方法层面，有以下几种典型的方法来利用外部知识：首先是在预训练语言模型中融合知识。例如，Ye 等人 [Ye et al. 2019] 自动构造了一个常识相关的多项选择问答数据集用于预训练语言表示模型。另一种思路是使用图神经网络来融合知识，例如，Feng 等人 [Feng et al. 2020] 提出了一种多跳图关系网络 MHGRN，把基于路径的推理方法和图神经网络统一在一起，通过引入结构化关系注意力机制，对消息传递路径进行建模，实现了更好的解释性和泛化性。

4．基于对话语义理解的任务型对话方法

任务型对话系统通过与后台知识图谱交互，辅助用户完成特定领域的任务，如订餐、查询天气或导航等。其核心步骤为对话语义理解，也称为对话状态跟踪，主要包括意图识别和对话状态识别 [Liu and Lane 2016]。例如，对于用户需求“帮我订一张明天去北京的机票”，意图识别需要从固定类别集合中正确预测“订机票”这个意图，对话状态识别需要对预定义的“目的地”、“日期”等若干槽抽取“北京”、“明天”等值。根据解决方案的不同，现有的任务型对话系统可以分为两大类，流水线式任务型对话系统和端到端式任务型对话系统。流水线式任务型对话系统依次建模对话意图识别、对话状态识别等多个模块。端到端式任务型对话系统使用一个模型完成整个对话任务。随着数据驱动的深度学习技术的进步，任务型对话系统逐步从基于规则的或基于传统机器学习的方法向基于深度学习的方法发展。在流水线任务型对话中，早期方法使用判别式模型识别意图和对话状态 [Jang et al. 2016, Mrksic et al. 2017]，它将预定义的值作为类别，每次预测选择一个预定义的值作为输出。尽管判别式对话状态跟踪方法在简单的数据集上能够取得不错的效果，但是它无法处理未登录值，即不在预定义列表中的值。为了解决未登录值问题，生成式对话状态跟踪方法从对话历史中抽取文本片段作为值 [Rastogi et al. 2017, Xu and Hu 2018]。目前，对话状态跟踪任务在不断地往多领域发展，以希望模型能够具备处理多领域对话状态跟踪的能力。在多领域对话状态跟踪任务 [Budzianowski et al. 2018] 中，领域转换和槽间关系建模等问题成为重要的研究内容。

流水线式任务型对话系统由多个独立的模块组成，这增加了研究人员设计、开发和维护任务型对话系统的成本。因此，很多研究人员开始研究端到端式任务型对话系统 [Eric et al. 2017, Rojas-Barahona et al. 2017]。这类系统通过一个端到端的深层神经网络模型完成整个对话任务。如何有效地融合知识库是这类系统的核心问题。为此，之前的工作主要在两方面进行优化，即如何表示知识库以及如何从知识库中检索知识。Eric 等人 [Eric et al. 2017] 使用

键值对（key-value pairs）的形式表示知识库，这种方法可以建模实体以及它的一跳关系信息。Madotto 等人 [Madotto et al. 2018] 使用内存网络表示知识库，该方法可以在内存单元中存储实体的相关信息以提升实体表示能力。Reddy 等人 [Reddy et al. 2019] 使用多层内存网络将知识库表示为查询语句、查询到的实体和键值对形式的知识，以融合多层级的信息促进富含知识的回复的生成。Wen 等人 [Wen et al. 2018] 将对话状态的隐层表示融入到查询向量中，该方法可以提升知识检索的准确率。Wu 等人 [Wu et al. 2019a] 提出全局-局部指针网络，它可以利用全局指针筛选实体，以去除与回复不相关的实体。

近年来，随着大规模预训练语言模型及提示学习（Prompt Learning）学习范式的进展，很多研究人员开始研究基于预训练模型的端到端任务型对话系统。现有工作主要可以分为如何利用现有预训练模型构建对话系统和如何设计适用于任务型对话的预训练模型这两类研究思路。Le 等人 [Le et al. 2020] 和 Hosseini-Asl 等人 [Hosseini-Asl et al. 2020] 将多个对话子任务联合建模为一个统一的序列生成任务，并使用 GPT-2 预训练语言模型 [Radford et al. 2019] 学习该任务。Peng 等人 [Peng et al. 2020] 在对话语料上预训练对话系统，并使用自回归语言模型学习多个对话子任务。Mehri 等人 [Mehri et al. 2019] 设计了多个预训练任务，如下句检索、下句生成和掩盖句子检索等，以预训练对话系统。Wu 等人 [Wu et al. 2019b] 设计了基于对话顺序的自监督任务以预训练模型，并采用生成对抗网络从预训练模型中学习有效的信息。Yang 等人 [Yang et al. 2022] 利用提示学习实现少样本的对话状态跟踪。

5．融合知识图谱的生成式对话方法

构建能够与人类进行自然交流的对话系统是人工智能和自然语言处理的长远目标之一，受限于技术和资源瓶颈，对话系统提出后的较长一段时间内发展缓慢。近年来，以循环神经网络 [Mikolov et al. 2010]、Transformer[Vaswani et al. 2017] 等为基础的自然语言生成模型可以根据原始数据自动学习生成目标序列，如文本序列，这种生成模型开始在对话系统等众多任务中崭露头角，表现出了优异性能。同时，互联网上广泛存在着大规模人与人之间的原始对话数据，结合计算机的运算能力显著提升，生成式对话方法逐渐普及，它是在编码器-解码器（encoder-decoder）[Cho et al. 2014] 框架下自动习得对话的回复生成模式。在这个框架下，编码器可以将对话历史（上下文）编码成为数值化的分布式表示方式，解码器基于编码结果与目前已经生成的词语，逐词生成目标词语序列。

上述完全数据驱动的端到端对话生成模型容易生成通用无意义的答复，因此，在生成式对话系统中引入知识，一方面可以提高回复内容的知识性和质量，另一方面也可以增强模型的可解释能力。例如，为了对需求“洛国富是哪儿人”并生成答案“洛国富出生于巴西，他现在是中国人”，需要融合利用“洛国富”的“性别”、“出生地”、“国籍”等知识内容。为

了实现上述目标并推动相关研究，人们构建了多个富含知识信息的对话数据集并提出了大量融合知识图谱的生成式对话方法。在数据集构建方面，不同机构基于 Wikipedia 文本、结构化知识图谱等资源，利用众包等方式，构建了覆盖电影等不同领域的 Wizard[Dinan et al. 2019]、CMU_DoG[Zhou et al. 2018b]、Holl-E[Moghe et al. 2018]、DoConv[Zhou et al. 2020]、DyKgChat[Tuan et al. 2019]、NaturalConv[Wang et al. 2021] 等知识型人机对话数据集。在模型方面，人们针对不同知识类型并针对不同目标提出了各种融合知识图谱的生成式对话方法。为了捕获通用的背景知识，Vougiouklis 等人 [Vougiouklis et al. 2016] 先在带有类别的维基百科上预训练了一个句子表示模型，以此作为背景知识模型，对话生成时用预训练的模型获取当前输入上下文的表示，再拼接到普通 Seq2Seq 解码器的隐状态中生成回复。此外，Yin 等人 [Yin et al. 2016] 和 He 等人 [He et al. 2017] 将问答任务与对话生成任务结合，弥补了传统对话生成任务中存在知识匮乏的问题，他们在 Seq2Seq 框架下提出了生成式的问答模型 GenQA 和 COREQA。Zhou 等人 [Zhou et al. 2018a] 在对话生成中引入了常识知识库，常识知识也被表示成三元组的形式如“(玻璃，属性，易碎)”，采用静态注意力机制增强输入上下文的表示，并利用动态注意力机制按一定权重选取常识知识图谱中的知识来生成回复，取得了不错的效果。近年来，人们开始在预训练语言模型中融入知识图谱 [Liu et al. 2020, Sun et al. 2020a]，致力于对话回复生成等知识密集型任务的性能的提升 [Petroni et al. 2021]。

四、 技术展望与发展趋势

纵观知识问答和对话研究发展的态势和技术现状，以下研究方向或问题将可能成为未来整个领域和行业重点关注的方向：

1. 回答多跳、时间推理等复杂类型问答

问答系统是知识图谱乃至人工智能领域的重要应用，它能够利用大量文本、知识库、问答对话等异构数据，通过匹配、推理等手段为用户返回精准答案。目前，简单事实型问答已经取得明显进展，例如，最新方法在每个问题“仅需利用一个事实”的 SimpleQuestions 数据集已经取得 95.7% 的准确率；基于大规模预训练的问答模型在“文本抽取型”简单问答数据集 SQuAD 中已能超越人类水平。但是，在需要多个事实、涉及逻辑推理的复杂问答任务中，当前方法性能不佳，例如，在典型复杂问答数据集 ComplexWebQuestions 上，最新的使用结构化知识库和非结构化文本的问答模型分别只能取得 70.4% 和 34.2% 的准确率，离人类水平还有非常大的距离，而当前在医疗、司法、国防、生活等真实应用场景中亟需复杂问题的求解能力。为了提升问答系统的服务范围，人们从最初的简单事实型问答任务开始往多事实问答、路径类问答、时间推理类问答、数值计算类问答等类型更多样、难度更复杂、

更接近真实应用场景的复杂问答任务，例如，2015 年提出的 SimpleQuestions 数据集关注简单事实型问答，后续提出的 WebQuestionsSP、ComplexWebQuestions、LC-QuAD、KQA Pro 等问答数据集开始关注包含多个事实的复杂问题，2020 年提出的 DuSQL 和 2021 年提出的 CRONQUESTIONS 数据集分别关注数值计算类和时间推理类问题。

2. 文本、图谱、表格等多元知识的联合利用

随着问答研究的深入，人们愈发意识到问答已经不再局限于某个单一的知识来源，因此，结合文本、知识图谱、表格等多元知识的混合式问答近年来得到了更多的关注。在数据集层面，2020 年，Chen 等人 [Chen et al. 2020a] 提出了文本、表格混合数据集 HybridQA，包含了 13000 个表格和 293269 篇文章，其中表格和文章的对应关系是已知的。该数据集需要同时处理表格和文本信息才能回答问题，因此需要模型同时具备表格和文本理解能力，并且，有的问题需要多跳才能回答，增加了数据集的难度。OTT-QA[Chen et al. 2021] 在 HybridQA 数据集的基础上，不再提供表格、文章问题之间的对应关系，需要模型检索获得，进一步提高了模型对于检索和理解的能力要求。在方法层面，Hybrider 模型以表格中的单元格为单位，将相应文章的信息扩充到表格信息中，然后将表格信息线性化，计算和问题的相关程度。OTT-QA 模型分为检索器（Retriever）和阅读器（Reader）两部分。模型首先对每个表格检索相关的文章，然后将文章和表格拼接在一起形成一个块，作为检索的基本单位。阅读器将所有检索到的块拼接起来，使得不同块之间的信息可以相互补充，最后送入模型中得到结果。DuRePa[Li et al. 2021] 模型将表格和文章分开处理，可以在某些问题上利用生成的 SQL 语句，更好地利用表格信息中的结构化信息。

3. 符号推理与数值计算相结合的问答方法

知识问答与对话利用知识图谱对用户信息需求产生精准答复内容，从技术路线上看可以分为基于符号逻辑（知识驱动）的方法和基于数值运算（数据驱动）的方法。通过对现有工作的总结，可以发现：针对限定域的知识图谱，基于符号逻辑的技术得到了广泛的使用；而随着互联网上知识图谱规模的不断增长，问答与对话利用的知识图谱逐渐从限定域转向大规模多领域上，基于数值运算的技术可扩展性强，受到了更多的重视。这类方法的优势在于把传统问答与对话的复杂步骤转变为一个可学习的过程，虽然取得了一定的效果，但是训练过程容易受到训练数据质量的影响，缺乏已有知识的约束，学习的数值模型无法应用于新的领域（领域泛化问题），也无法处理新的语义组合（组合泛化问题）。因此，如何将深度学习等数值计算与符号语义组合等传统语义分析方法想结合，使这两种技术路线相互融合、相互约束，提升知识问答与对话的效果，是一个很值得深入研究的方向。

4. 更鲁棒、更具解释性的问答对话模型

随着深度学习在人工智能的各个领域取得一系列的进展，各项任务的性能得到了长足的进步，在很多问答场景下，机器在评价指标上的表现已经超过了人类。但是，仍然有两个问题不容忽视，需要在未来研究中加以重视。一方面，深度学习模型的不透明性也成为了人们关心的重要问题，这也使得问答对话模型无法获取足够的信任，人们迫切地需要更具可解释性的问答对话模型。另一方面，复杂的模型十分脆弱，在受到对抗攻击时，表现欠佳，即便是看似简单的干扰也会使得模型的输出发生不可预测的变化，人们迫切地需要更加鲁棒的模型。

5. 跨模态真实场景的知识服务

互联网信息的不断丰富已经使人们不再满足于单纯的文本形式的问答，其他诸如图像、声音等多模态的信息也是用户信息需求的一部分，跨模态的知识服务亟需进一步研究。视觉问答是在跨模态方面得到最多研究的一项任务，要求和回答图片相关的自然语言问题。一个好的视觉问答数据集应该足够大，以捕捉真实世界场景中问题和图像内容的各种可能性。目前，很多数据集都包含来自 Microsfot COCO[Lin et al. 2014] 的图像，该数据集包含 328000 张图片，91 种对象类型，共有 250 万个标注实例。DAQUAR 数据集 [Malinowski and Fritz 2014] 和 VQA 数据集 [Antol et al. 2015] 等都是研究人员常用的数据集。MULTIMODALQA[Talmor et al. 2021] 是一个跨模态问答数据集，需要结合图像、文本和表格来回答问题。该复杂问答场景符合知识服务的未来趋势。

五、总结

问答与对话系统不仅是知识图谱和自然语言处理的典型应用，也是验证系统智能水平的有效评测手段，其研究可以追溯至人工智能诞生伊始。与人工智能发展历史类似，问答与对话系统也经历从知识驱动（如专家系统）到数据驱动（如检索式问答、社区问答等），以及到目前的知识与数据联合驱动的发展历程。整体上，问答与对话技术的发展趋势是从限定领域向开放领域、从回答简单问题到处理复杂问题、从单独问答到多轮对话、从应用单模态单结构知识内容到融合利用多模态多结构知识系统、从浅层语义分析向深度知识推理不断推进。我们有理由相信，随着大规模多领域知识资源和问答对话学习数据的不断积累，以及自然语言处理、深度学习、知识工程等相关技术的飞速发展，知识问答与对话技术在未来有可能取得相当程度的突破和应用。

参考文献

[Abujabal et al. 2018] Abujabal, A., Saha Roy, R., Yahya, M., and Weikum, G. (2018). Neverending learning for open-domain question answering over knowledge bases. In Proceedings of the 2018 World Wide Web Conference, WWW '18, page 1053–1062, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[Abujabal et al. 2017] Abujabal, A., Yahya, M., Riedewald, M., and Weikum, G. (2017). Automated template generation for question answering over knowledge graphs. In Proceedings of the 26th International Conference on World Wide Web, WWW '17, page 1191–1200, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[Antol et al. 2015] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 2425–2433.

[Bordes et al. 2014] Bordes, A., Chopra, S., and Weston, J. (2014). Question answering with subgraph embeddings. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 615–620, Doha, Qatar. Association for Computational Linguistics.

[Budzianowski et al. 2018] Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). MultiWOZ - a large-scale multi-domain Wizard-ofOz dataset for task-oriented dialogue modelling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 5016–5026, Brussels, Belgium.

Association for Computational Linguistics.

[Chen et al. 2021] Chen, W., Chang, M.-W., Schlinger, E., Wang, W. Y., and Cohen, W. W. (2021). Open question answering over tables and text. In International Conference on Learning Representations.

[Chen et al. 2020a] Chen, W., Zha, H., Chen, Z., Xiong, W., Wang, H., and Wang, W. Y. (2020a). HybridQA: A dataset of multi-hop question answering over tabular and textual data. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1026–1036, Online. Association for Computational Linguistics.

[Chen et al. 2020b] Chen, Y., Li, H., Hua, Y., and Qi, G. (2020b). Formal query building with query structure prediction for complex question answering over knowledge base. In Bessiere, C., editor, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-

20, pages 3751–3758. International Joint Conferences on Artificial Intelligence Organization. Main track.

[Cho et al. 2014] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder– decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

[Das et al. 2021] Das, R., Zaheer, M., Thai, D., Godbole, A., Perez, E., Lee, J. Y., Tan, L., Polymenakos, L., and McCallum, A. (2021). Case-based reasoning for natural language queries over knowledge bases. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9594–9611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Dinan et al. 2019] Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2019). Wizard of wikipedia: Knowledge-powered conversational agents. ArXiv, abs/1811.01241.

[Dong and Lapata 2016] Dong, L. and Lapata, M. (2016). Language to logical form with neural attention. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33–43, Berlin, Germany. Association for Computational Linguistics.

[Dong et al. 2015] Dong, L., Wei, F., Zhou, M., and Xu, K. (2015). Question answering over Freebase with multi-column convolutional neural networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 260–269, Beijing, China. Association for Computational Linguistics.

[Eric et al. 2017] Eric, M., Krishnan, L., Charette, F., and Manning, C. D. (2017). Key-value retrieval networks for task-oriented dialogue. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

[Etzioni and Oren 2011] Etzioni and Oren (2011). Search needs a shake-up. Nature, 476(7358):25.

[Feng et al. 2020] Feng, Y., Chen, X., Lin, B. Y., Wang, P., Yan, J., and Ren, X. (2020). Scalable multi-hop relational reasoning for knowledge-aware question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1295–1309,

Online. Association for Computational Linguistics.

[Gangi Reddy et al. 2019] Gangi Reddy, R., Contractor, D., Raghu, D., and Joshi, S. (2019). Multi-level memory for task oriented dialogs. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3744–3754, Minneapolis, Minnesota. Association for Computational Linguistics.

[Ge and Mooney 2009] Ge, R. and Mooney, R. (2009). Learning a compositional semantic parser using an existing syntactic parser. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 611–619, Suntec, Singapore. Association for Computational Linguistics.

[Han et al. 2020] Han, J., Cheng, B., and Wang, X. (2020). Open domain question answering based on text enhanced knowledge graph with hyperedge infusion. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1475–1481, Online. Association for Computational Linguistics.

[Hao et al. 2017] Hao, Y., Zhang, Y., Liu, K., He, S., Liu, Z., Wu, H., and Zhao, J. (2017). An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 221–231, Vancouver, Canada. Association for Computational Linguistics.

[He et al. 2021] He, G., Lan, Y., Jiang, J., Zhao, W. X., and Wen, J.-R. (2021). Improving multi-hop knowledge base question answering by learning intermediate supervision signals. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM ’21, page 553–561, New York, NY, USA. Association for Computing Machinery.

[He et al. 2017] He, S., Liu, C., Liu, K., and Zhao, J. (2017). Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 199–208, Vancouver, Canada. Association for Computational Linguistics.

[Hosseini-Asl et al. 2020] Hosseini-Asl, E., McCann, B., Wu, C.-S., Yavuz, S., and Socher, R. (2020). A simple language model for task-oriented dialogue. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20, Red Hook, NY, USA. Curran Associates Inc.

[Hu et al. 2018] Hu, S., Zou, L., and Zhang, X. (2018). A state-transition framework to answer complex questions over knowledge base. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2098–2108, Brussels, Belgium. Association for Computational Linguistics.

[Huang et al. 2019] Huang, L., Le Bras, R., Bhagavatula, C., and Choi, Y. (2019). Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

[Jang et al. 2016] Jang, Y., Ham, J., Lee, B.-J., Chang, Y., and Kim, K.-E. (2016). Neural dialog state tracker for large ontologies by attention mechanism. In 2016 IEEE Spoken Language Technology Workshop (SLT), pages 531–537.

[Jia et al. 2018] Jia, Z., Abujabal, A., Saha Roy, R., Strötgen, J., and Weikum, G. (2018). Tequila: Temporal question answering over knowledge bases. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, page 1807–1810, New York, NY, USA. Association for Computing Machinery.

[Kapanipathi et al. 2021] Kapanipathi, P., Abdelaziz, I., Ravishankar, S., Roukos, S., Gray, A., Fernandez Astudillo, R., Chang, M., Cornelio, C., Dana, S., Fokoue, A., Garg, D., Gliozzo, A., Gurajada, S., Karanam, H., Khan, N., Khandelwal, D., Lee, Y.-S., Li, Y., Luus, F., Makondo, N., Mihindukulasooriya, N., Naseem, T., Neelam, S., Popa, L., Gangi Reddy, R., Riegel, R., Rossiello, G., Sharma, U., Bhargav, G. P. S., and Yu, M. (2021). Leveraging Abstract Meaning Representation for knowledge base question answering. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3884–3894, Online. Association for Computational Linguistics.

[Lan et al. 2021] Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W. X., and Wen, J.-R. (2021). A survey on complex knowledge base question answering: Methods, challenges and solutions. In Zhou, Z.-H., editor, Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pages 4483–4491. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

[Le et al. 2020] Le, H., Sahoo, D., Liu, C., Chen, N., and Hoi, S. C. (2020). UniConv: A unified conversational neural architecture for multi-domain task-oriented dialogues. In Proceedings of the

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1860–1877, Online. Association for Computational Linguistics.

[Li et al. 2021] Li, A. H., Ng, P., Xu, P., Zhu, H., Wang, Z., and Xiang, B. (2021). Dual reader-parser on hybrid textual and tabular evidence for open domain question answering. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4078–4088, Online. Association for Computational Linguistics.

[Lin et al. 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, Computer Vision – ECCV 2014, pages 740–755, Cham. Springer International Publishing.

[Liu and Lane 2016] Liu, B. and Lane, I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. pages 685–689.

[Liu and Singh 2004] Liu, H. and Singh, P. (2004). Conceptnet—a practical commonsense reasoning tool-kit. BT technology journal, 22(4):211–226.

[Liu et al. 2020] Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., and Wang, P. (2020). K-bert: Enabling language representation with knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 2901–2908.

[Luo et al. 2018] Luo, K., Lin, F., Luo, X., and Zhu, K. (2018). Knowledge base question answering via encoding of complex query graphs. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2185–2194, Brussels, Belgium. Association for Computational Linguistics.

[Madotto et al. 2018] Madotto, A., Wu, C.-S., and Fung, P. (2018). Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1468–1478, Melbourne, Australia. Association for Computational
Linguistics.

[Malinowski and Fritz 2014] Malinowski, M. and Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1, NIPS'14, page 1682–1690, Cambridge, MA, USA. MIT Press.

[Mehri et al. 2019] Mehri, S., Razumovskaia, E., Zhao, T., and Eskenazi, M. (2019). Pretraining methods for dialog context representation learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3836–3845, Florence, Italy. Association for Computational Linguistics.

[Mikolov et al. 2010] Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In Kobayashi, T., Hirose, K., and Nakamura, S., editors, INTERSPEECH, pages 1045–1048. ISCA.

[Misra and Artzi 2016] Misra, D. K. and Artzi, Y. (2016). Neural shift-reduce CCG semantic parsing. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1775–1786, Austin, Texas. Association for Computational Linguistics.

[Moghe et al. 2018] Moghe, N., Arora, S., Banerjee, S., and Khapra, M. M. (2018). Towards exploiting background knowledge for building conversation systems. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

[Mrkšić et al. 2017] Mrkšić, N., Ó Séaghdha, D., Wen, T.-H., Thomson, B., and Young, S. (2017). Neural belief tracker: Data-driven dialogue state tracking. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

[Peng et al. 2021] Peng, B., Li, C., Li, J., Shayandeh, S., Liden, L., and Gao, J. (2021). Soloist: Building task bots at scale with transfer learning and machine teaching. Transactions of the Association for Computational Linguistics, 9:807–824.

[Petroni et al. 2021] Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., and Riedel, S. (2021). KILT: a benchmark for knowledge intensive language tasks. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2523–2544, Online. Association for Computational Linguistics.

[Radford et al. 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

[Rastogi et al. 2017] Rastogi, A., Hakkani-Tür, D. Z., and Heck, L. (2017). Scalable multidomain dialogue state tracking. pages 561–568.

[Sap et al. 2019] Sap, M., Rashkin, H., Chen, D., Le Bras, R., and Choi, Y. (2019). Social IQa:

Commonsense reasoning about social interactions. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

[Saxena et al. 2020] Saxena, A., Tripathi, A., and Talukdar, P. (2020). Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4498–4507, Online. Association for Computational Linguistics.

[Steedman 1997] Steedman, M. (1997). Surface structure and interpretation. In Linguistic inquiry.

[Sun et al. 2020a] Sun, T., Shao, Y., Qiu, X., Guo, Q., Hu, Y., Huang, X., and Zhang, Z. (2020a). CoLAKE: Contextualized language and knowledge embedding. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3660–3670, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[Sun et al. 2020b] Sun, Y., Zhang, L., Cheng, G., and Qu, Y. (2020b). Sparqa: skeletonbased semantic parsing for complex questions over knowledge bases. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8952–8959.

[Talmor et al. 2019] Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2019). CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

[Talmor et al. 2021] Talmor, A., Yoran, O., Catav, A., Lahav, D., Wang, Y., Asai, A., Ilharco, G., Hajishirzi, H., and Berant, J. (2021). Multimodalqa: complex question answering over text, tables and images. In International Conference on Learning Representations.

[Tuan et al. 2019] Tuan, Y.-L., Chen, Y.-N., and Lee, H.-y. (2019). DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1855–1865, Hong Kong, China. Association for Computational Linguistics.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V.,

Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

[Vougiouklis et al. 2016] Vougiouklis, P., Hare, J., and Simperl, E. (2016). A neural network approach for knowledge-driven response generation. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3370–3380, Osaka, Japan. The COLING 2016 Organizing Committee.

[Wang et al. 2021] Wang, X., Li, C., Zhao, J., and Yu, D. (2021). Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 14006–14014.

[Wen et al. 2018] Wen, H., Liu, Y., Che, W., Qin, L., and Liu, T. (2018). Sequence-tosequence learning for task-oriented dialogue with dialogue state representation. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3781–3792, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[Wen et al. 2017] Wen, T.-H., Vandyke, D., Mrkšić, N., Gašić, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2017). A network-based end-to-end trainable taskoriented dialogue system. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

[Wu et al. 2019] Wu, J., Wang, X., and Wang, W. Y. (2019). Self-supervised dialogue learning. pages 3857–3867.

[Xu and Hu 2018] Xu, P. and Hu, Q. (2018). An end-to-end approach for handling unknown slot values in dialogue state tracking. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1448–1457, Melbourne, Australia. Association for Computational Linguistics.

[Yang et al. 2022] Yang, Y., Lei, W., Cao, J., Li, J., and Chua, T.-S. (2022). Prompt learning for few-shot dialogue state tracking. ArXiv, abs/2201.05780.

[Ye et al. 2019] Ye, Z., Chen, Q., Wang, W., and Ling, Z. (2019). Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. ArXiv, abs/1908.06725.

[Yin et al. 2016] Yin, J., Jiang, X., Lu, Z., Shang, L., Li, H., and Li, X. (2016). Neural generative question answering. In Proceedings of the Workshop on Human-Computer Question Answering,

pages 36–42, San Diego, California. Association for Computational Linguistics.

[Yin and Neubig 2018] Yin, P. and Neubig, G. (2018). TRANX: A transition-based neural abstract syntax parser for semantic parsing and code generation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 7–12, Brussels, Belgium. Association for Computational Linguistics.

[Zettlemoyer and Collins 2005] Zettlemoyer, L. S. and Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence, UAI'05, page 658–666, Arlington, Virginia, USA. AUAI Press.

[Zhou et al. 2018a] Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2018a). Commonsense knowledge aware conversation generation with graph attention. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pages 4623–4629. International Joint Conferences on Artificial Intelligence Organization.

[Zhou et al. 2020] Zhou, H., Zheng, C., Huang, K., Huang, M., and Zhu, X. (2020). KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7098–7108, Online. Association for Computational Linguistics.

[Zhou et al. 2018b] Zhou, K., Prabhumoye, S., and Black, A. W. (2018b). A dataset for document grounded conversations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

[Zhu et al. 2020] Zhu, S., Cheng, X., and Su, S. (2020). Knowledge-based question answering by tree-to-sequence learning. Neurocomputing, 372:64–72.

[赵军 et al. 2022] 赵军, 刘康, 何世柱, 陈玉博, and 张元哲 (2022). 《知识图谱：算法与实践》 . 高等教育出版社.

第十二章 基于知识的搜索与推荐

程龚

南京大学 计算机科学与技术系，南京 210023

一、任务定义、目标和研究意义

我们几乎每天都在使用搜索和推荐服务，包括百度、必应、谷歌等通用搜索引擎，电商、新闻、学术等垂直领域网站和 App 的搜索和推荐功能，个人计算机和手机等设备操作系统内置的搜索功能等。尽管搜索和推荐的应用场景十分多样，其核心任务是相通的：如何准确捕获用户意图及喜好，返回相匹配的结果。

“知识”很早便被应用于搜索和推荐，例如，基于语言学知识库 WordNet 对用户提交的关键词查询进行同义词扩展，以提高查全率，但这并非本章关注的重点。在本章中，我们将“知识”的范围主要限定于描述世界知识或领域知识的知识图谱，其描述的实体及其关系本身往往就是搜索和推荐的对象，同时也可作为辅助资源用于增强搜索和推荐服务的能力。

“基于知识的搜索与推荐”是较为广泛的研究主题，可以从不同的角度进行细分。我们根据用户需求的不同目标类型，将现有工作主要分为以下三类。

（1）实体搜索与推荐：目标是从知识图谱中找出用户需要的实体。这是目前最常见、应用最广泛的一类工作，例如，电商领域的商品搜索与推荐，学术领域的论文搜索与推荐，电影和音乐的搜索与推荐等，都属于这类工作。

（2）实体关系搜索：目标是从知识图谱中找出用户关注的一组实体之间的关系。目前，这类工作的典型应用包括：商业领域的企业关系搜索，社交领域的人际关系链推荐，安全领域的特定目标关系搜索等。

（3）基于关键词的知识探索：目标是从知识图谱中找出用户感兴趣的子图。这是上述两类工作的泛化形式，主要应用场景是帮助用户探索和理解知识图谱的内容和结构，寻找用户感兴趣的信息。

上述三类工作以知识图谱的组成元素作为搜索和推荐的目标。与之不同的是，还有一类工作将知识图谱作为一种外部辅助资源，用于增强传统搜索引擎和推荐系统的能力，这些内容不作为本章的重点。

二、研究内容和关键科学问题

上述三类工作的具体任务目标不同，因此，其研究内容和面临的关键科学问题也有所差

异，以下分别阐述。

1. 实体搜索与推荐

实体搜索通常可以追溯到2003年发表在WWW上的一篇论文[Guha et al., 2003]，这篇论文描述了一个实体搜索原型系统，可以视作如今谷歌知识图谱的雏形。实体搜索与推荐的输入可以是表达用户需求的一组关键词，也可以是表达用户兴趣的一个或一组实体；输出是用户可能要找的一个排序的实体列表。

实体搜索与推荐的现有研究主要关注以下问题。

（1）实体搜索与推荐模型：即如何将用户输入映射为实体输出，例如，如何将关键词与实体描述进行语义匹配，或者如何从用户感兴趣的实体发现更多的相似实体。

（2）实体排序算法：在上述的模型中通常已经考虑了与输入相关的排序，这里的排序主要指与输入无关的排序，例如，如何度量实体自身的重要性。

（3）实体集探索方法：当用户需求不够精确时，如何提供便捷的交互手段帮助用户对输出的大量实体进行探索，例如，如何对实体分组以便批量过滤掉无关实体。

（4）实体摘要算法：对于输出的每个实体，除了实体名称以外，还应当选择呈现哪些属性作为其完整描述的摘要，以便用户快速判定相关性，甚至直接满足用户需求。

2. 实体关系搜索

在知识图谱中，实体之间不仅可以通过边直接相连，也可以通过路径、子图等更复杂的结构间接相连，统称为实体关系[Gong Cheng, 2020]。实体关系搜索通常可以追溯到2003年发表在WWW上的一篇论文[Anyanwu and Sheth, 2003]，这篇论文定义了实体在知识图谱中的多种语义关系。实体关系搜索的输入是用户感兴趣的两个或多个实体，输出是这些实体之间的关系。

实体关系搜索的现有研究主要关注以下问题。

（1）实体关系搜索算法：即实体关系应定义为何种图结构，以及如何在大规模知识图谱中高效地搜索实体关系。

（2）实体关系排序算法：如何计算实体关系的重要性并排序。

（3）实体关系集探索方法：当用户需求不够精确时，如何提供便捷的交互手段帮助用户对输出的大量实体关系进行探索，例如，如何对结果分组以便批量过滤掉无关结果。

3. 基于关键词的知识探索

用户表达需求较为便捷的方式是关键词查询，其目标可能是实体，可能是实体关系，也可能并没有明确目标，而是试图探索和理解知识图谱的内容[Lissandrini et al., 2020]。因此，

作为知识图谱搜索的最一般形式，基于关键词的搜索的输入是一组关键词，输出是知识图谱中包含这些关键词的子图。

基于关键词的知识探索的现有研究主要关注的问题是：子图应定义为何种图结构，以及如何在大规模知识图谱中高效地搜索子图。

三、技术方案和研究现状

1. 实体搜索与推荐

1）实体搜索与推荐模型

对于用户输入的关键词查询，最直接的搜索方法是关键词匹配。然而，用户很少在查询中直接指明目标实体的名称，更多的是通过关键词描述目标实体的属性。因此，在为实体构建索引时，需要索引实体的全部属性，构成实体的虚拟文档[Cheng and Qu, 2009]，继而可以采用传统的文档搜索技术，如向量空间模型（VSM）和词频-反文档频率（TF-IDF）等实现实体搜索。这种做法的局限性是无法区分实体的不同属性。为此，可以采用 BM25F 模型[Blanco et al., 2011]，这是经典的 BM25 模型的一种扩展，支持为不同的属性赋予不同的权重。类似地，也可采用 FSDM 模型[Zhiltsov et al., 2015]，这是序列依赖模型（SDM）的一种扩展，也支持对属性加权。这些模型面临的共同问题是如何确定属性的权重。进一步地，可以采用排序学习技术，将上述模型输出的实体得分作为特征，综合得到实体的最终得分[Chen et al., 2016]。此外，还可以利用知识图谱的图结构，例如，采用激活传播技术对实体排序[Rocha et al., 2004, Lukovnikov and Ngomo, 2014]，通过主题模型统一表示实体的文本和结构信息[Hong et al., 2020]，或者基于随机游走扩充实体的表示[Takahiro Komamizu, 2020, Nikolaev and Kotov, 2020]。也有一些模型专门利用了实体的类型层次[Ma et al., 2018, Lin and Lam, 2018, Lin et al., 2018]。目前，实体搜索的常用评测集是 DBpedia-Entity v2[Hasibi et al., 2017]。

用户也可能输入一个或一组实体，需要推荐与之相似或相关的其它实体。对于相似实体的推荐，一类方法从用户输入的实体出发，在知识图谱中随机游走，计算到达其它实体的稳态概率作为相似度[Balmin et al., 2004]，并且可以限制随机游走遵循具有指定模式的路径或子图以表达特定的相似关系[Sun et al., 2011, Xiong et al., 2015, Shi et al., 2017, Huang et al., 2016]。另一类方法从知识图谱中抽取用户输入的实体集共同具有的结构特征，例如共同的类型和属性[Metzger et al., 2017]，到其它类型实体的距离[Lim et al., 2013]，共同关联的路径模式等[Yu et al., 2012, Zhang et al., 2017, Chen et al., 2018, Shi et al., 2021]，再据此扩展实体集。对于相关实体的推荐，通常假设相关关系可以被表示为一组路径模式（称作元路径），根据用户输入的正例自动学习元路径的权重，元路径可以预定义[Bu et al., 2014]，也可以自

动生成[Lao and Cohen, 2010, Wang et al., 2016, Meng et al., 2015, Gu et al., 2019]，后者对于富模式的知识图谱（例如实体和关系类型多样的百科领域）尤为关键。也有一些方法采用了图嵌入技术[Liu et al., 2017, Liu et al., 2018]或者生成模型[Rastogi et al., 2019, Zhou et al., 2020]。

从更广的涉及多用户的视角来看，基于知识图谱的实体推荐方法总体上可以分为三类[Guo et al., 2020]：基于嵌入的方法[Zhang et al., 2016, Huang et al., 2018, Wang et al., 2019, Cao et al., 2019, Liu et al., 2021, Tu et al., 2021]采用图嵌入技术将知识图谱编码到低维向量，推荐系统的用户也表示为知识图谱中的顶点；基于路径的方法[Xian et al., 2019, Sun et al., 2018, Anelli et al., 2021]将用户和被推荐实体之间的关系表示为图，分析实体间的连通模式；上述两类方法可以融合[Wang et al., 2019, Wang et al., 2018, Wang et al., 2019, Wang et al., 2021, Wang et al., 2021, Cao et al., 2021]，将嵌入表示沿着图的结构传播并迭代优化。也有工作将推荐转化为基于知识图谱的问答[Chen et al., 2021]。此外，还有一些新闻推荐的工作利用知识图谱来增强新闻文本的表示[Wang et al., 2018, Liu et al., 2020, Lee et al., 2020, Qi et al., 2021]。

2）实体排序算法

实体除了与用户输入的相关性之外，其自身的重要性在排序中也需要考虑。实体作为知识图谱中的顶点，可以通过度量顶点在图结构中的中心性作为实体的重要性。常见的中心性度量方法包括顶点的度、介度以及 HITS[Jon M. Kleinberg, 1999]、PageRank[Page et al., 1999]等算法：度表示顶点关联的边的数量；介度表示经过顶点的其它顶点间最短路的数量；HITS 和 PageRank 算法认为与重要顶点相邻的顶点才重要，并分别设计了迭代式的中心性计算方法。其中，PageRank 基于随机游走的思想，对游走目标的选择是等概率的，可以对这个模型进行扩展[Diligenti et al., 2004]，用概率值来表达某种有助于重要性计算的其它信息。进一步地，可以采用排序学习技术，将上述各种重要性作为特征，学习得到实体的最终排序[Dali et al., 2012]。

在基于语义网和链接数据规范的知识图谱中，实体以全局唯一的国际化资源标识符（IRI）作为标识符，因此，不同的知识图谱可以方便地描述同一个实体。在这种场景下，实体重要性的度量就超越了单个知识图谱的范围，需要综合考虑多个知识图谱。一种直接的方法是[Harth et al., 2009]：将实体的重要性定义为所有提及该实体的知识图谱的重要性之和，为此，构建了知识图谱之间的引用关系图，边表示一个知识图谱提及了另一个知识图谱中定义的实体，在这个图上运行 PageRank 算法计算知识图谱的重要性。进一步地，跨知识图谱的实体关系可能有多种，参考 TF-IDF 的思想对关系加权，局部频繁而全局不常见的关系更重要，在随机游走中被赋予更高的概率[Delbru et al., 2010]。还可以构建一种混合图，包括所有知识图谱的全部内容，将每个知识图谱也作为一个顶点加入图中，并与其提及的实体通过边相

连，这个混合图便同时表示了知识图谱内部和知识图谱之间的关系，在这个图上运行PageRank 算法对实体排序[Hogan et al., 2006]。

3）实体集探索方法

不够精确的用户需求可能搜索到或者推荐出大量的实体，对这些实体的进一步浏览和筛选是一种探索式搜索[Gary Marchionini, 2006]。

比较常见的一种探索式搜索模式称作分面搜索（faceted search），即按属性值对实体进行分组，用户选择特定的属性值对分组进行筛选过滤。对于电商等垂直领域，用于分组的属性可以预先设定；对于开放域场景，用于分组的属性的选取及其取值的呈现方式有多种自动化方法。例如，一种简单通用的做法是按照类型对实体进行层次化的分组[Cheng and Qu, 2009]。也可以由用户通过拖拽等方式自由选取属性用于分组[Schraefel et al., 2006]，但当可选属性较多时，比较难以操作。为此，可以自动选取用于分组的属性，这样的属性具有三个特征[Oren et al., 2006]：不同取值对应的实体数量分布较为均衡，这样分组过滤的效率较高；不同取值的种类不多，否则用户难以选取；实体集内的大部分实体都有该属性，否则过滤的覆盖面不足。考虑到一些属性的取值是比较长的文本，不宜整体作为分组项，可以从文本中提取词作为分组项[Sinha and Karger, 2005]。对于数值型属性，可以根据取值的分布情况，从均衡的角度自动分段作为分组项[Wagner et al., 2011]。对于取值为实体的属性，可以采用实体的类型作为粒度更大的分组项[Hildebrand et al., 2006]。还有一些方法允许按属性序列分组[Arenas et al., 2016, Sherkhonov et al., 2017]，可以表达多跳关系。对于大规模知识图谱，当可选属性与取值较多时，分面搜索可能面临性能问题，需要预先对图谱做一些离线分析[Moreno-Vega and Hogan, 2018]。

另一种较常见的探索式搜索模式是对搜索结果自动聚类成为分组，目前这类工作还比较少[Cheng et al., 2010, Zheng et al., 2018]，主要难点既包括如何计算实体的相似性，也包括如何为聚类生成有明确含义的、用户可以理解的分组标签。

4）实体摘要算法

搜索和推荐的每个实体，在知识图谱中可能有大量的属性描述，在结果呈现时从中自动选取一个最优子集的方法称作实体摘要[Liu et al., 2021]。不同的实体摘要方法，利用了实体的不同特征，并采用不同的框架组合特征。总体而言，特征可以分为通用特征和特定特征：通用特征包括与频率和中心性有关的特征，与信息量有关的特征，与多样性有关的特征等；特定特征包括与特定领域、特定上下文以及特定用户等有关的特征。组合多种特征的框架更是种类繁多，以下介绍一些代表性方法。

LinkSUM[Thalhammer et al., 2016]对实体的所有属性根据其在知识图谱中的出现频率排

序，属性值则采用 PageRank 等方法排序。RELIN[Cheng et al., 2011]采用了加权的 PageRank 算法，将每个属性值作为一个顶点，计算属性值的信息量和属性值之间的相关度作为在属性值之间随机游走的概率。CES[Yan et al., 2016]是对 RELIN 的扩展，在计算概率时加入了与当前会话上下文的相关性。DIVERSUM[Sydow et al., 2013]避免选取同一个属性的不同取值，以增强摘要的多样性。FACES[Gunaratna et al., 2015]及其扩展版本 FACES-E[Gunaratna et al., 2016]基于文本特征对属性值聚类分组，尽量从不同分组中选取属性值以提高多样性，组内排序主要基于属性值的信息量和频率。MMR-QSFS[Zhang et al., 2012]贪心地迭代选取属性值，每轮选取与查询相关性最大并且与其它已选属性值相似性最小的一个属性值。还有一些方法将实体摘要建模为组合优化问题，例如，ESSTER[刘庆霞 et al., 2020]采用二阶背包模型，最大化属性值的重要性和可读性，最小化冗余性； C3D+P[Cheng et al., 2015]、REMES[Gunaratna et al., 2017]、COMB[Cheng et al., 2015]采用二阶背包、二阶多维背包等模型同时为多个实体生成摘要，选取的属性值侧重于关联或者区分不同实体。

随着深度学习技术的广泛应用，近年来也出现了一些基于深度神经网络的实体摘要方法。例如，ESA[Wei and Liu, 2019]采用 TransE 和 BiLSTM 编码属性及属性值，再通过注意力机制对属性值排序。DeepLENS[Liu et al., 2020]则忽略了知识图谱的图结构，只对属性和属性值的文本编码。NEST[Li et al., 2020]同时编码文本和结构，继而分别计算属性值的重要性和摘要的多样性，并且不同于上述有监督方法在训练时对标记数据的依赖，NEST 利用知识图谱 DBpedia 和维基百科之间的对应关系自动生成大量的标记数据用于训练，是一种弱监督方法。此外，还有方法利用用户反馈，基于深度强化学习技术持续优化实体摘要的质量[Liu et al., 2020]。

目前，实体摘要的常用评测集是 ESBM[Liu et al., 2020]。

2. 实体关系搜索

1）实体关系搜索算法

在知识图谱中，两个实体之间的关系通常被定义为连接两个顶点的路径。对于需要快速搜索出任意一条实体关系的应用，一种简单的做法是离线计算并存储知识图谱的广度优先搜索树，在线搜索时便可以通过拼接两个顶点到树根的路径得到一条实体关系[Lehmann et al., 2007]。更多的应用需要搜索最重要的实体关系，重要性可以通过边权来表达，问题便转化为带权图上的最短路问题，可以通过经典的 Dijkstra 算法找出最短路；需要注意的是，考虑到大规模知识图谱往往存储于磁盘等不支持随机存取的存储器中，Dijkstra 算法的具体实现有不少性能优化技巧[Gubichev and Neumann, 2011]。还有一些应用需要搜索出所有的实体关系，当然，考虑到两个顶点之间的路径数量是指数级的，通常只搜索限定长度的所有路径，

可以采用双向搜索算法[Janik and Kochut, 2005]。除了路径以外，两个实体之间的关系还有一些其它定义方法。例如，有一些工作认为某些特定的子图结构是必需且不可分解的，将其作为实体关系搜索的目标[Fang et al., 2011]。

对于多个实体之间的关系搜索，除了简单地搜索两两关系再拼接以外[Heim et al., 2010]，目前有三类具有代表性的方法，采用不同的实体关系定义。第一类方法的搜索目标是连通所有输入实体的极小子图，且子图直径不超过给定阈值[Cheng et al., 2016, Cheng et al., 2021]。具体搜索算法是从每个输入实体同时开始搜索直至相遇于同一顶点，在此过程中，可以利用直径上限进行剪枝。第二类方法的搜索目标是连通所有输入实体的最优极小子图，且包含的顶点数量不超过给定阈值[Tong and Faloutsos, 2006, Chen et al., 2011]。这是一个 NP 难问题，为了达到近似最优，具体搜索算法综合了运用贪心、动态规划、粒子群优化等技术。第三类方法的搜索目标是连通所有输入实体的最小权子图，即斯坦纳树（Steiner tree）[Kasneci et al., 2009]。这是图论中一个经典的 NP 难问题，相关研究较为丰富，具体搜索算法可以采用局部搜索等技术。

当用户输入的实体在知识图谱中的距离较远甚至不连通时，上述搜索算法有可能返回空结果，对用户不够友好。为了解决该问题，一种解决方法是对输入的实体集进行松弛，在搜索前先计算出一个（在限定直径内）可以连通的最大实体子集，再以该子集作为输入调用上述搜索算法[Li et al., 2020, Li et al., 2020]。

2）实体关系排序算法

实体关系作为一个子图，其重要性通常被定义为包含的顶点和边的重要性之和。顶点即实体的重要性在前面的章节中已经有所讨论，边的重要性有以下几种常见的度量方法。第一种方法是计算边的类型在知识图谱中的出现频率[Tartari and Hogan, 2018]。第二种方法是计算边的信息量[Anyanwu et al., 2005]。第三种方法是计算边的排他度[Hulpus et al., 2015]，即边的类型在其关联的两个顶点上的出现频率。

另一些方法并不注重于分别度量每个顶点或每条边的重要性，而是将子图作为一个整体来评价。例如，一般认为规模越小的子图刻画了越紧密的关系，因此越重要。此外，还可以分析顶点之间或者边之间的相互关系，例如，顶点实体之间的平均相似度、边类型的多样度等。有实验表明[Cheng et al., 2017]，这些基于整体视角的排序算法的实际效果更显著。

此外，也有一些基于机器学习的实体关系排序算法，采用了排序学习[Chen and Prasanna, 2012]、主动学习[Bianchi et al., 2017]等技术，能在一定程度上满足用户的个性化需求。还有方法通过选取包含不同顶点的实体关系，提高搜索结果的多样性[Aebeloe et al., 2018]。

3）实体关系集探索方法

对于可能搜索到的大量的实体关系，其探索方法主要包括分面搜索和聚类搜索两种。

分面搜索从实体关系中抽取特征作为分组过滤项。例如，可以由用户指定实体关系必须包含的关键词[Zhou et al., 2011]，或者必须包含的顶点和边的类型[Cheng et al., 2014]，一些方法还可以自动对用户指定的类型进行扩展[Giuseppe Pirrò, 2019]。

聚类搜索目前主要根据顶点和边的类型，即本体中定义的类和属性，提取出实体关系的本体图模式，将模式相同的实体关系聚为一类，并以模式作为聚类的标签。现有方法主要关注选择哪些模式呈现给用户作为分组过滤项[Cheng et al., 2016, Cheng et al., 2021, Cheng et al., 2014, Giuseppe Pirrò, 2015, Gu et al., 2018]。例如，重要的模式应当具有较高的频度；具有较高的信息量，即由较具体的类和属性组成；并且相互之间的重叠度较低。进一步地，考虑到类和属性在本体中具有上下位层次关系，模式也可以相应地组织为层次结构，即形成实体关系的层次化聚类[Zhang et al., 2013, Zhang et al., 2013]。此外，还有一些工作关注性能问题，即对于用户指定的模式，如何高效搜索出符合该模式的实体关系[Liang et al., 2016]。

3. 基于关键词的知识探索

早期的一些工作尝试解释关键词查询的含义，将关键词自动转化为知识图谱上的形式化查询[Tran et al., 2007, Zhou et al., 2007, Wang et al., 2008, Tran et al., 2009, Tran et al., 2009, Ladwig and Tran, 2010, Fu et al., 2011, Fu and Anyanwu, 2011, Tran et al., 2011, Pound et al., 2012]，如 SPARQL 查询，这类方法后来逐步发展为基于知识图谱的问答系统，在报告的其它章节已有所讨论，这里不再赘述。其它工作则主要尝试直接从知识图谱中抽取包含关键词的子图。

这个问题可以建模为图论中的组斯坦纳树（group Steiner tree）问题，简称 GST 问题：每个关键词匹配到知识图谱中的一组顶点，每条边的权值越小则越重要，目标是找出一棵边权和最小的树，包含每组顶点中的至少一个顶点，即匹配所有关键词。GST 问题是一个 NP 难问题，其精确算法可以采用最佳优先搜索和动态规划[Ding et al., 2007]、A*搜索[Li et al., 2016]等技术，但这些算法的性能难以适用于大规模知识图谱。为此，出现了很多近似算法，大多以优化子图内顶点间的距离和为目标。例如，BANKS-II 采用双向搜索，并通过激活传播技术来确定顶点的搜索顺序[Kacholia et al., 2005]。BLINKS 构建了距离索引，用于引导搜索过程[He et al., 2007]。还有一些方法先找出权和较小的团，再从团中提取 GST[Kargar and An, 2011]。为了提高搜索性能，也可以利用知识图谱的本体模式作为摘要来辅助剪枝[Le et al., 2014]。最新的 KeyKG 算法则以距离和最短路计算作为核心步骤，并同时构建静态和动态的中心标记索引实现高效的距离和最短路计算，用有限的空间代价换取了较高的搜索性能[Shi et al., 2020]。

除了 GST 以外，还有一些其它的建模方式。例如，有的方法从查询中解析出答案类型、修饰词等，进而相应地要求子图满足特定的条件[Shan et al., 2017]。有的方法要求子图具有较高的内聚性，内聚性既可以指边的密集程度[Zhu et al., 2018]，也可以指顶点间的相似性[Shi et al., 2021, Shi et al., 2021]。有的方法并不以一个最优子图作为搜索目标，而是按序搜索出所有子图[Golenberg and Sagiv, 2016]。还有一些方法考虑了不同的应用场景，例如知识图谱动态演化[Xu et al., 2013]、高并发搜索等[Yang et al., 2019]。此外，当返回多个子图时，提高搜索结果的多样性至关重要[Qin et al., 2012]。而当关键词较多时，为了避免子图规模过大导致用户难以阅读，可以预先对关键词进行松弛，自动识别并删除部分关键词[Cheng et al., 2020]。

四、技术展望和发展趋势

实体搜索与推荐由于具有广泛而明确的应用场景，在各互联网企业的大力推动下，技术相对已经较为成熟，未来的研究重点可能在于如何提高搜索和推荐结果的可解释性。实体关系搜索尽管有着同样长的研究历史，但早期受搜索性能等因素的制约，应用发展较为缓慢，如今随着技术成熟度的持续提高，有望成为知识图谱技术的新一轮应用增长点。

本章讨论的搜索与推荐技术，大多局限于一个给定的知识图谱内部，显然这远未达到此前研究者对于语义网和链接数据技术的预期。事实上，在网络开放环境下，搜索具体知识之前，首先需要解决的问题是：应当选择哪个知识图谱来搜索？即知识图谱本身的搜索，而这个问题的更一般形式便是数据集的搜索[Chapman et al., 2020]，目前，国内外均已经出现了一些原型系统[Pietriga et al., 2018, Brickley et al., 2019, Wang et al., 2021, Wang et al., 2022]，但其功能还远不能满足需要，并且尚未与实体搜索、实体关系搜索等功能实现整合，因此，这个方向被认为具有广阔的研究空间。

此外，搜索技术的下沉，也可能成为一个发展趋势。搜索已经成为一种必不可少的通用技术，正逐步推动着知识图谱管理系统将搜索技术内化为其组成部分之一，知识图谱的分布式搜索、并发搜索等相关问题的研究将愈发迫切。

参考文献

[Guha et al., 2003] Ramanathan V. Guha, Rob McCool, Eric Miller: Semantic search. WWW 2003: 700-709

[Gong Cheng, 2020] Gong Cheng: Relationship search over knowledge graphs. SIGWEB Newsl. 2020(Summer): 3:1-3:8 (2020)

[Anyanwu and Sheth, 2003] Kemafor Anyanwu, Amit P. Sheth: ρ-Queries: enabling querying for semantic associations on the semantic web. WWW 2003: 690-699

[Lissandrini et al., 2020] Matteo Lissandrini, Torben Bach Pedersen, Katja Hose, Davide Mottin: Knowledge graph exploration: where are we and where are we going?" SIGWEB Newsl. 2020: 4:1-4:8

[Cheng and Qu, 2009] Gong Cheng, Yuzhong Qu: Searching Linked Objects with Falcons: Approach, Implementation and Evaluation. Int. J. Semantic Web Inf. Syst. 5(3): 49-70 (2009)

[Blanco et al., 2011] Roi Blanco, Peter Mika, Sebastiano Vigna: Effective and Efficient Entity Search in RDF Data. ISWC (1) 2011: 83-97

[Zhiltsov et al., 2015] Nikita Zhiltsov, Alexander Kotov, Fedor Nikolaev: Fielded Sequential Dependence Model for Ad-Hoc Entity Retrieval in the Web of Data. SIGIR 2015: 253-262

[Chen et al., 2016] Jing Chen, Chenyan Xiong, Jamie Callan: An Empirical Study of Learning to Rank for Entity Search. SIGIR 2016: 737-740

[Devezas and Nunes, 2021] José Devezas, Sérgio Nunes: A Review of Graph-Based Models for Entity-Oriented Search. SN Computer Science 2: 437 (2021)

[Rocha et al., 2004] Cristiano Rocha, Daniel Schwabe, Marcus Poggi de Aragão: A hybrid approach for searching in the semantic web. WWW 2004: 374-383

[Lukovnikov and Ngomo, 2014] Denis Lukovnikov, Axel-Cyrille Ngonga Ngomo: SESSA – Keyword-Based Entity Search through Coloured Spreading Activation. NLIWoD 2014

[Hong et al., 2020] Yu Hong, Suo Feng, Yanghua Xiao: EntityLDA: A Topic Model for Entity Retrieval on Knowledge Graph. ICKG 2020: 388-395

[Takahiro Komamizu, 2020] Takahiro Komamizu: Random walk-based entity representation learning and re-ranking for entity search. Knowl. Inf. Syst. 62(8): 2989-3013 (2020)

[Nikolaev and Kotov, 2020] Fedor Nikolaev, Alexander Kotov: Joint Word and Entity Embeddings for Entity Retrieval from a Knowledge Graph. ECIR (1) 2020: 141-155

[Ma et al., 2018] Denghao Ma, Yueguo Chen, Kevin Chen-Chuan Chang, Xiaoyong Du, Chuanfei Xu, Yi Chang: Leveraging Fine-Grained Wikipedia Categories for Entity Search. WWW 2018: 1623-1632

[Lin and Lam, 2018] Xinshi Lin, Wai Lam: Entity Retrieval via Type Taxonomy Aware Smoothing. ECIR 2018: 773-779

[Lin et al., 2018] Xinshi Lin, Wai Lam, Kwun Ping Lai: Entity Retrieval in the Knowledge Graph

with Hierarchical Entity Type and Content. ICTIR 2018: 211-214

[Hasibi et al., 2017] Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, Jamie Callan: DBpedia-Entity v2: A Test Collection for Entity Search. SIGIR 2017: 1265-1268

[Balmin et al., 2004] Andrey Balmin, Vagelis Hristidis, Yannis Papakonstantinou: ObjectRank: Authority-Based Keyword Search in Databases. VLDB 2004: 564-575

[Sun et al., 2011] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, Tianyi Wu: PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. Proc. VLDB Endow. 4(11): 992-1003 (2011)

[Xiong et al., 2015] Yun Xiong, Yangyong Zhu, Philip S. Yu: Top-k Similarity Join in Heterogeneous Information Networks. IEEE Trans. Knowl. Data Eng. 27(6): 1710-1723 (2015)

[Shi et al., 2017] Yu Shi, Po-Wei Chan, Honglei Zhuang, Huan Gui, Jiawei Han: PReP: Path-Based Relevance from a Probabilistic Perspective in Heterogeneous Information Networks. KDD 2017: 425-434

[Huang et al., 2016] Zhipeng Huang, Yudian Zheng, Reynold Cheng, Yizhou Sun, Nikos Mamoulis, Xiang Li: Meta Structure: Computing Relevance in Large Heterogeneous Information Networks. KDD 2016: 1595-1604

[Metzger et al., 2017] Steffen Metzger, Ralf Schenkel, Marcin Sydow: QBEES: query-by-example entity search in semantic knowledge graphs based on maximal aspects, diversity-awareness and relaxation. J. Intell. Inf. Syst. 49(3): 333-366 (2017)

[Lim et al., 2013] Lipyeow Lim, Haixun Wang, Min Wang: Semantic queries by example. EDBT 2013: 347-358

[Yu et al., 2012] Xiao Yu, Yizhou Sun, Brandon Norick, Tiancheng Mao, Jiawei Han: User guided entity similarity search using meta-path selection in heterogeneous information networks. CIKM 2012: 2025-2029

[Zhang et al., 2017] Xiangling Zhang, Yueguo Chen, Jun Chen, Xiaoyong Du, Ke Wang, Ji-Rong Wen: Entity Set Expansion via Knowledge Graphs. SIGIR 2017: 1101-1104

[Chen et al., 2018] Jun Chen, Yueguo Chen, Xiangling Zhang, Xiaoyong Du, Ke Wang, Ji-Rong Wen: Entity set expansion with semantic features of knowledge graphs. J. Web Semant. 52-53: 33-44 (2018)

[Shi et al., 2021] Chuan Shi, Jiayu Ding, Xiaohuan Cao, Linmei Hu, Bin Wu, Xiaoli Li: Entity set

expansion in knowledge graph: a heterogeneous information network perspective. Frontiers Comput. Sci. 15(1): 151307 (2021)

[Bu et al., 2014] Shaoli Bu, Xiaoguang Hong, Zhaohui Peng, Qingzhong Li: Integrating meta-path selection with user-preference for top-k relevant search in heterogeneous information networks. CSCWD 2014: 301-306

[Lao and Cohen, 2010] Ni Lao, William W. Cohen: Relational retrieval using a combination of path-constrained random walks. Mach. Learn. 81(1): 53-67 (2010)

[Wang et al., 2016] Chenguang Wang, Yizhou Sun, Yanglei Song, Jiawei Han, Yangqiu Song, Lidan Wang, Ming Zhang: RelSim: Relation Similarity Search in Schema-Rich Heterogeneous Information Networks. SDM 2016: 621-629

[Meng et al., 2015] Changping Meng, Reynold Cheng, Silviu Maniu, Pierre Senellart, Wangda Zhang: Discovering Meta-Paths in Large Heterogeneous Information Networks. WWW 2015: 754-764

[Gu et al., 2019] Yu Gu, Tianshuo Zhou, Gong Cheng, Ziyang Li, Jeff Z. Pan, Yuzhong Qu: Relevance Search over Schema-Rich Knowledge Graphs. WSDM 2019: 114-122

[Liu et al., 2017] Zemin Liu, Vincent W. Zheng, Zhou Zhao, Fanwei Zhu, Kevin Chen-Chuan Chang, Minghui Wu, Jing Ying: Semantic Proximity Search on Heterogeneous Graph by Proximity Embedding. AAAI 2017: 154-160

[Liu et al., 2018] Zemin Liu, Vincent W. Zheng, Zhou Zhao, Fanwei Zhu, Kevin Chen-Chuan Chang, Minghui Wu, Jing Ying: Distance-Aware DAG Embedding for Proximity Search on Heterogeneous Graphs. AAAI 2018: 2355-2362

[Rastogi et al., 2019] Pushpendre Rastogi, Adam Poliak, Vince Lyzinski, Benjamin Van Durme: Neural variational entity set expansion for automatically populated knowledge graphs. Inf. Retr. J. 22(3-4): 232-255 (2019)

[Zhou et al., 2020] Tianshuo Zhou, Ziyang Li, Gong Cheng, Jun Wang, Yuang Wei: GREASE: A Generative Model for Relevance Search over Knowledge Graphs. WSDM 2020: 780-788

[Guo et al., 2020] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, Qing He: A Survey on Knowledge Graph-Based Recommender Systems. CoRR abs/2003.00911 (2020)

[Zhang et al., 2016] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, Wei-Ying Ma: Collaborative Knowledge Base Embedding for Recommender Systems. KDD 2016: 353-362

[Huang et al., 2018] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, Edward Y. Chang: Improving Sequential Recommendation with Knowledge-Enhanced Memory Networks. SIGIR 2018: 505-514

[Wang et al., 2019] Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie, Minyi Guo: Multi-Task Feature Learning for Knowledge Graph Enhanced Recommendation. WWW 2019: 2000-2010

[Cao et al., 2019] Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, Tat-Seng Chua: Unifying Knowledge Graph Learning and Recommendation: Towards a Better Understanding of User Preferences. WWW 2019: 151-161

[Liu et al., 2021] Yi Liu, Bohan Li, Yalei Zang, Aoran Li, Hongzhi Yin: A Knowledge-Aware Recommender with Attention-Enhanced Dynamic Convolutional Network. CIKM 2021: 1079-1088

[Tu et al., 2021] Ke Tu, Peng Cui, Daixin Wang, Zhiqiang Zhang, Jun Zhou, Yuan Qi, Wenwu Zhu: Conditional Graph Attention Networks for Distilling and Refining Knowledge Graphs in Recommendation. CIKM 2021: 1834-1843

[Xian et al., 2019] Yikun Xian, Zuohui Fu, S. Muthukrishnan, Gerard de Melo, Yongfeng Zhang: Reinforcement Knowledge Graph Reasoning for Explainable Recommendation. SIGIR 2019: 285-294

[Sun et al., 2018] Zhu Sun, Jie Yang, Jie Zhang, Alessandro Bozzon, Long-Kai Huang, Chi Xu: Recurrent knowledge graph embedding for effective recommendation. RecSys 2018: 297-305

[Anelli et al., 2021] Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio, Antonio Ferrara, Alberto Carlo Maria Mancino: Sparse Feature Factorization for Recommender Systems with Knowledge Graphs. RecSys 2021: 154-165

[Wang et al., 2019] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, Tat-Seng Chua: KGAT: Knowledge Graph Attention Network for Recommendation. KDD 2019: 950-958

[Wang et al., 2018] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, Minyi Guo: RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems. CIKM 2018: 417-426

[Wang et al., 2019] Hongwei Wang, Fuzheng Zhang, Mengdi Zhang, Jure Leskovec, Miao Zhao, Wenjie Li, Zhongyuan Wang: Knowledge-aware Graph Neural Networks with Label Smoothness Regularization for Recommender Systems. KDD 2019: 968-977

[Wang et al., 2021] Xiang Wang, Tinglin Huang, Dingxian Wang, Yancheng Yuan, Zhenguang Liu,

Xiangnan He, Tat-Seng Chua: Learning Intents behind Interactions with Knowledge Graph for Recommendation. WWW 2021: 878-887

[Wang et al., 2021] Chunyang Wang, Yanmin Zhu, Haobing Liu, Wenze Ma, Tianzi Zang, Jiadi Yu: Enhancing User Interest Modeling with Knowledge-Enriched Itemsets for Sequential Recommendation. CIKM 2021: 1889-1898

[Cao et al., 2021] Xianshuai Cao, Yuliang Shi, Han Yu, Jihu Wang, Xinjun Wang, Zhongmin Yan, Zhiyong Chen: DEKR: Description Enhanced Knowledge Graph for Machine Learning Method Recommendation. SIGIR 2021: 203-212

[Chen et al., 2021] Yu Chen, Ananya Subburathinam, Ching-Hua Chen, Mohammed J. Zaki: Personalized Food Recommendation as Constrained Question Answering over a Large-scale Food Knowledge Graph. WSDM 2021: 544-552

[Wang et al., 2018] Hongwei Wang, Fuzheng Zhang, Xing Xie, Minyi Guo: DKN: Deep Knowledge-Aware Network for News Recommendation. WWW 2018: 1835-1844

[Liu et al., 2020] Danyang Liu, Jianxun Lian, Shiyin Wang, Ying Qiao, Jiun-Hung Chen, Guangzhong Sun, Xing Xie: KRED: Knowledge-Aware Document Representation for News Recommendations. RecSys 2020: 200-209

[Lee et al., 2020] Dongho Lee, Byungkook Oh, Seungmin Seo, Kyong-Ho Lee: News Recommendation with Topic-Enriched Knowledge Graphs. CIKM 2020: 695-704

[Qi et al., 2021] Tao Qi, Fangzhao Wu, Chuhan Wu, Yongfeng Huang: Personalized News Recommendation with Knowledge-aware Interactive Matching. SIGIR 2021: 61-70

[Jon M. Kleinberg, 1999] Jon M. Kleinberg: Authoritative Sources in a Hyperlinked Environment. J. ACM 46(5): 604-632 (1999)

[Page et al., 1999] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd: The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab, 1999.

[Diligenti et al., 2004] Michelangelo Diligenti, Marco Gori, Marco Maggini: A Unified Probabilistic Framework for Web Page Scoring Systems. IEEE Trans. Knowl. Data Eng. 16(1): 4-16 (2004)

[Dali et al., 2012] Lorand Dali, Blaz Fortuna, Duc Thanh Tran, Dunja Mladenic: Query-Independent Learning to Rank for RDF Entity Search. ESWC 2012: 484-498

[Harth et al., 2009] Andreas Harth, Sheila Kinsella, Stefan Decker: Using Naming Authority to Rank Data and Ontologies for Web Search. ISWC 2009: 277-292

[Delbru et al., 2010] Renaud Delbru, Nickolai Toupikov, Michele Catasta, Giovanni Tummarello,

Stefan Decker: Hierarchical Link Analysis for Ranking Web Data. ESWC (2) 2010: 225-239

[Hogan et al., 2006] Aidan Hogan, Andreas Harth, Stefan Decker: ReConRank: A Scalable Ranking Method for Semantic Web Data with Context. SSWS 2006

[Gary Marchionini, 2006] Gary Marchionini: Exploratory search: from finding to understanding. Commun. ACM 49(4): 41-46 (2006)

[Schraefel et al., 2006] Monica M. C. Schraefel, Max L. Wilson, Alistair Russell, Daniel A. Smith: mSpace: improving information access to multimedia domains with multimodal exploratory search. Commun. ACM 49(4): 47-49 (2006)

[Oren et al., 2006] Eyal Oren, Renaud Delbru, Stefan Decker: Extending Faceted Navigation for RDF Data. ISWC 2006: 559-572

[Sinha and Karger, 2005] Vineet Sinha, David R. Karger: Magnet: Supporting Navigation in Semistructured Data Environments. SIGMOD Conference 2005: 97-106

[Wagner et al., 2011] Andreas Wagner, Günter Ladwig, Thanh Tran: Browsing-Oriented Semantic Faceted Search. DEXA (1) 2011: 303-319

[Hildebrand et al., 2006] Michiel Hildebrand, Jacco van Ossenbruggen, Lynda Hardman: /facet: A Browser for Heterogeneous Semantic Web Repositories. ISWC 2006: 272-285

[Arenas et al., 2016] Marcelo Arenas, Bernardo Cuenca Grau, Evgeny Kharlamov, Sarunas Marciuska, Dmitriy Zheleznyakov: Faceted search over RDF-based knowledge graphs. J. Web Semant. 37-38: 55-74 (2016)

[Sherkhonov et al., 2017] Evgeny Sherkhonov, Bernardo Cuenca Grau, Evgeny Kharlamov, Egor V. Kostylev: Semantic Faceted Search with Aggregation and Recursion. ISWC (1) 2017: 594-610

[Moreno-Vega and Hogan, 2018] José Moreno-Vega, Aidan Hogan: GraFa: Scalable Faceted Browsing for RDF Graphs. ISWC (1) 2018: 301-317

[Cheng et al., 2010] Gong Cheng, Wei Hu, Sulong Xu, Yuzhong Qu: digO: Clustering-based Interactive Entity Search. ESWC 2010 Posters

[Zheng et al., 2018] Liang Zheng, Yuzhong Qu, Xinqi Qian, Gong Cheng: A hierarchical co-clustering approach for entity exploration over Linked Data. Knowl. Based Syst. 141: 200-210 (2018)

[Liu et al., 2021] Qingxia Liu, Gong Cheng, Kalpa Gunaratna, Yuzhong Qu: Entity summarization: State of the art and future challenges. J. Web Semant. 69: 100647 (2021)

[Thalhammer et al., 2016] Andreas Thalhammer, Nelia Lasierra, Achim Rettinger: LinkSUM: Using

Link Analysis to Summarize Entity Data. ICWE 2016: 244-261

[Cheng et al., 2011] Gong Cheng, Thanh Tran, Yuzhong Qu: RELIN: Relatedness and Informativeness-Based Centrality for Entity Summarization. ISWC (1) 2011: 114-129

[Yan et al., 2016] Jihong Yan, Yanhua Wang, Ming Gao, Aoying Zhou: Context-Aware Entity Summarization. WAIM (1) 2016: 517-529

[Sydow et al., 2013] Marcin Sydow, Mariusz Pikula, Ralf Schenkel: The notion of diversity in graphical entity summarisation on semantic knowledge graphs. J. Intell. Inf. Syst. 41(2): 109-149 (2013)

[Gunaratna et al., 2015] Kalpa Gunaratna, Krishnaprasad Thirunarayan, Amit P. Sheth: FACES: Diversity-Aware Entity Summarization Using Incremental Hierarchical Conceptual Clustering. AAAI 2015: 116-122

[Gunaratna et al., 2016] Kalpa Gunaratna, Krishnaprasad Thirunarayan, Amit P. Sheth, Gong Cheng: Gleaning Types for Literals in RDF Triples with Application to Entity Summarization. ESWC 2016: 85-100

[Zhang et al., 2012] Lanbo Zhang, Yi Zhang, Yunfei Chen: Summarizing highly structured documents for effective search interaction. SIGIR 2012: 145-154

[刘庆霞 et al., 2020] 刘庆霞, 程龚, 瞿裕忠: 一种高可读低冗余实体摘要的生成方法. 中国科学:信息科学 50(6): 845-861 (2020)

[Cheng et al., 2015] Gong Cheng, Danyun Xu, Yuzhong Qu: C3D+P: A summarization method for interactive entity resolution. J. Web Semant. 35: 203-213 (2015)

[Gunaratna et al., 2017] Kalpa Gunaratna, Amir Hossein Yazdavar, Krishnaprasad Thirunarayan, Amit P. Sheth, Gong Cheng: Relatedness-based Multi-Entity Summarization. IJCAI 2017: 1060-1066

[Cheng et al., 2015] Gong Cheng, Danyun Xu, Yuzhong Qu: Summarizing Entity Descriptions for Effective and Efficient Human-centered Entity Linking. WWW 2015: 184-194

[Wei and Liu, 2019] Dongjun Wei, Yaxin Liu: ESA: Entity Summarization with Attention. EYRE 2019

[Liu et al., 2020] Qingxia Liu, Gong Cheng, Yuzhong Qu: DeepLENS: Deep Learning for Entity Summarization. DL4KG@ESWC 2020

[Li et al., 2020] Junyou Li, Gong Cheng, Qingxia Liu, Wen Zhang, Evgeny Kharlamov, Kalpa Gunaratna, Huajun Chen: Neural Entity Summarization with Joint Encoding and Weak Supervision.

IJCAI 2020: 1644-1650

[Liu et al., 2020] Qingxia Liu, Yue Chen, Gong Cheng, Evgeny Kharlamov, Junyou Li, Yuzhong Qu: Entity Summarization with User Feedback. ESWC 2020: 376-392

[Liu et al., 2020] Qingxia Liu, Gong Cheng, Kalpa Gunaratna, Yuzhong Qu: ESBM: An Entity Summarization BenchMark. ESWC 2020: 548-564

[Lehmann et al., 2007] Jens Lehmann, Jörg Schüppel, Sören Auer: Discovering Unknown Connections - the DBpedia Relationship Finder. CSSW 2007: 99-109

[Gubichev and Neumann, 2011] Andrey Gubichev, Thomas Neumann: Path Query Processing on Very Large RDF Graphs. WebDB 2011

[Janik and Kochut, 2005] Maciej Janik, Krys J. Kochut: BRAHMS: A WorkBench RDF Store and High Performance Memory System for Semantic Association Discovery. ISWC 2005: 431-445

[Fang et al., 2011] Lujun Fang, Anish Das Sarma, Cong Yu, Philip Bohannon: REX: Explaining Relationships between Entity Pairs. Proc. VLDB Endow. 5(3): 241-252 (2011)

[Heim et al., 2010] Philipp Heim, Steffen Lohmann, Timo Stegemann: Interactive Relationship Discovery via the Semantic Web. ESWC (1) 2010: 303-317

[Cheng et al., 2016] Gong Cheng, Daxin Liu, Yuzhong Qu: Efficient Algorithms for Association Finding and Frequent Association Pattern Mining. ISWC (1) 2016: 119-134

[Cheng et al., 2021] Gong Cheng, Daxin Liu, Yuzhong Qu: Fast Algorithms for Semantic Association Search and Pattern Mining. IEEE Trans. Knowl. Data Eng. 33(4): 1490-1502 (2021)

[Tong and Faloutsos, 2006] Hanghang Tong, Christos Faloutsos: Center-piece subgraphs: problem definition and fast solutions. KDD 2006: 404-413

[Chen et al., 2011] Chen Chen, Guoren Wang, Huilin Liu, Junchang Xin, Ye Yuan: SISP: a new framework for searching the informative subgraph based on PSO. CIKM 2011: 453-462

[Kasneci et al., 2009] Gjergji Kasneci, Maya Ramanath, Mauro Sozio, Fabian M. Suchanek, Gerhard Weikum: STAR: Steiner-Tree Approximation in Relationship Graphs. ICDE 2009: 868-879

[Li et al., 2020] Shuxin Li, Gong Cheng, Chengkai Li: Relaxing relationship queries on graph data. J. Web Semant. 61-62: 100557 (2020)

[Li et al., 2020] Shuxin Li, Zixian Huang, Gong Cheng, Evgeny Kharlamov, Kalpa Gunaratna: Enriching Documents with Compact, Representative, Relevant Knowledge Graphs. IJCAI 2020: 1748-1754

[Tartari and Hogan, 2018] Gonzalo Tartari, Aidan Hogan: WiSP: Weighted Shortest Paths for RDF Graphs. VOILA@ISWC 2018: 37-52

[Anyanwu et al., 2005] Kemafor Anyanwu, Angela Maduko, Amit P. Sheth: SemRank: ranking complex relationship search results on the semantic web. WWW 2005: 117-127

[Hulpus et al., 2015] Ioana Hulpus, Narumol Prangnawarat, Conor Hayes: Path-Based Semantic Relatedness on Linked Data and Its Use to Word and Entity Disambiguation. ISWC (1) 2015: 442-457

[Cheng et al., 2017] Gong Cheng, Fei Shao, Yuzhong Qu: An Empirical Evaluation of Techniques for Ranking Semantic Associations. IEEE Trans. Knowl. Data Eng. 29(11): 2388-2401 (2017)

[Chen and Prasanna, 2012] Na Chen, Viktor K. Prasanna: Learning to Rank Complex Semantic Relationships. Int. J. Semantic Web Inf. Syst. 8(4): 1-19 (2012)

[Bianchi et al., 2017] Federico Bianchi, Matteo Palmonari, Marco Cremaschi, Elisabetta Fersini: Actively Learning to Rank Semantic Associations for Personalized Contextual Exploration of Knowledge Graphs. ESWC (1) 2017: 120-135

[Aebeloe et al., 2018] Christian Aebeloe, Gabriela Montoya, Vinay Setty, Katja Hose: Discovering Diversified Paths in Knowledge Bases. Proc. VLDB Endow. 11(12): 2002-2005 (2018)

[Zhou et al., 2011] Mo Zhou, Yifan Pan, Yuqing Wu: Conkar: constraint keyword-based association discovery. CIKM 2011: 2553-2556

[Cheng et al., 2014] Gong Cheng, Yanan Zhang, Yuzhong Qu: Explass: Exploring Associations between Entities via Top-K Ontological Patterns and Facets. ISWC (2) 2014: 422-437

[Giuseppe Pirrò, 2019] Giuseppe Pirrò: Building relatedness explanations from knowledge graphs. Semantic Web 10(6): 963-990 (2019)

[Giuseppe Pirrò, 2015] Giuseppe Pirrò: Explaining and Suggesting Relatedness in Knowledge Graphs. ISWC (1) 2015: 622-639

[Gu et al., 2018] Yu Gu, Yue Liang, Gong Cheng, Daxin Liu, Ruidi Wei, Yuzhong Qu: Diversified and Verbalized Result Summarization for Semantic Association Search. WISE (1) 2018: 381-390

[Zhang et al., 2013] Yanan Zhang, Gong Cheng, Yuzhong Qu: Towards Exploratory Relationship Search: A Clustering-Based Approach. JIST 2013: 277-293

[Zhang et al., 2013] Yanan Zhang, Gong Cheng, Yuzhong Qu: RelClus: Clustering-based Relationship Search. ISWC (Posters & Demos) 2013: 1-4

[Liang et al., 2016] Jiongqian Liang, Deepak Ajwani, Patrick K. Nicholson, Alessandra Sala,

Srinivasan Parthasarathy: What Links Alice and Bob?: Matching and Ranking Semantic Patterns in Heterogeneous Networks. WWW 2016: 879-889

[Tran et al., 2007] Thanh Tran, Philipp Cimiano, Sebastian Rudolph, Rudi Studer: Ontology-Based Interpretation of Keywords for Semantic Search. ISWC/ASWC 2007: 523-536

[Zhou et al., 2007] Qi Zhou, Chong Wang, Miao Xiong, Haofen Wang, Yong Yu: SPARK: Adapting Keyword Query to Semantic Search. ISWC/ASWC 2007: 694-707

[Wang et al., 2008] Haofen Wang, Kang Zhang, Qiaoling Liu, Thanh Tran, Yong Yu: Q2Semantic: A Lightweight Keyword Interface to Semantic Search. ESWC 2008: 584-598

[Tran et al., 2009] Thanh Tran, Haofen Wang, Sebastian Rudolph, Philipp Cimiano: Top-k Exploration of Query Candidates for Efficient Keyword Search on Graph-Shaped (RDF) Data. ICDE 2009: 405-416

[Tran et al., 2009] Thanh Tran, Haofen Wang, Peter Haase: Hermes: Data Web search on a pay-as-you-go integration infrastructure. J. Web Semant. 7(3): 189-203 (2009)

[Ladwig and Tran, 2010] Günter Ladwig, Thanh Tran: Combining Query Translation with Query Answering for Efficient Keyword Search. ESWC (2) 2010: 288-303

[Fu et al., 2011] Haizhou Fu, Sidan Gao, Kemafor Anyanwu: CoSi: context-sensitive keyword query interpretation on RDF databases. WWW (Companion Volume) 2011: 209-212

[Fu and Anyanwu, 2011] Haizhou Fu, Kemafor Anyanwu: Effectively Interpreting Keyword Queries on RDF Databases with a Rear View. ISWC (1) 2011: 193-208

[Tran et al., 2011] Thanh Tran, Daniel M. Herzig, Günter Ladwig: SemSearchPro - Using semantics throughout the search process. J. Web Semant. 9(4): 349-364 (2011)

[Pound et al., 2012] Jeffrey Pound, Alexander K. Hudek, Ihab F. Ilyas, Grant E. Weddell: Interpreting keyword queries over web knowledge bases. CIKM 2012: 305-314

[Ding et al., 2007] Bolin Ding, Jeffrey Xu Yu, Shan Wang, Lu Qin, Xiao Zhang, Xuemin Lin: Finding Top-k Min-Cost Connected Trees in Databases. ICDE 2007: 836-845

[Li et al., 2016] Rong-Hua Li, Lu Qin, Jeffrey Xu Yu, Rui Mao: Efficient and Progressive Group Steiner Tree Search. SIGMOD Conference 2016: 91-106

[Kacholia et al., 2005] Varun Kacholia, Shashank Pandit, Soumen Chakrabarti, S. Sudarshan, Rushi Desai, Hrishikesh Karambelkar: Bidirectional Expansion For Keyword Search on Graph Databases. VLDB 2005: 505-516

[He et al., 2007] Hao He, Haixun Wang, Jun Yang, Philip S. Yu: BLINKS: ranked keyword searches

on graphs. SIGMOD Conference 2007: 305-316

[Kargar and An, 2011] Mehdi Kargar, Aijun An: Keyword Search in Graphs: Finding r-cliques. Proc. VLDB Endow. 4(10): 681-692 (2011)

[Le et al., 2014] Wangchao Le, Feifei Li, Anastasios Kementsietsidis, Songyun Duan: Scalable Keyword Search on Large RDF Data. IEEE Trans. Knowl. Data Eng. 26(11): 2774-2788 (2014)

[Shi et al., 2020] Yuxuan Shi, Gong Cheng, Evgeny Kharlamov: Keyword Search over Knowledge Graphs via Static and Dynamic Hub Labelings. WWW 2020: 235-245

[Shan et al., 2017] Yi Shan, Mingda Li, Yi Chen: Constructing target-aware results for keyword search on knowledge graphs. Data Knowl. Eng. 110: 1-23 (2017)

[Zhu et al., 2018] Yuanyuan Zhu, Qian Zhang, Lu Qin, Lijun Chang, Jeffrey Xu Yu: Querying Cohesive Subgraphs by Keywords. ICDE 2018: 1324-1327

[Shi et al., 2021] Yuxuan Shi, Gong Cheng, Trung-Kien Tran, Evgeny Kharlamov, Yulin Shen: Efficient Computation of Semantically Cohesive Subgraphs for Keyword-Based Knowledge Graph Exploration. WWW 2021: 1410-1421

[Shi et al., 2021] Yuxuan Shi, Gong Cheng, Trung-Kien Tran, Jie Tang, Evgeny Kharlamov: Keyword-Based Knowledge Graph Exploration Based on Quadratic Group Steiner Trees. IJCAI 2021: 1555-1562

[Golenberg and Sagiv, 2016] Konstantin Golenberg, Yehoshua Sagiv: A Practically Efficient Algorithm for Generating Answers to Keyword Search Over Data Graphs. ICDT 2016: 23:1-23:17

[Xu et al., 2013] Yanwei Xu, Jihong Guan, Fengrong Li, Shuigeng Zhou: Scalable continual top-k keyword search in relational databases. Data Knowl. Eng. 86: 206-223 (2013)

[Yang et al., 2019] Yueji Yang, Divyakant Agrawal, H. V. Jagadish, Anthony K. H. Tung, Shuang Wu: An Efficient Parallel Keyword Search Engine on Knowledge Graphs. ICDE 2019: 338-349

[Qin et al., 2012] Lu Qin, Jeffrey Xu Yu, Lijun Chang: Diversifying Top-K Results. Proc. VLDB Endow. 5(11): 1124-1135 (2012)

[Cheng et al., 2020] Gong Cheng, Shuxin Li, Ke Zhang, Chengkai Li: Generating Compact and Relaxable Answers to Keyword Queries over Knowledge Graphs. ISWC (1) 2020: 110-127

[Chapman et al., 2020] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, Paul Groth: Dataset search: a survey. VLDB J. 29(1): 251-272 (2020)

[Pietriga et al., 2018] Emmanuel Pietriga, Hande Gözükan, Caroline Appert, Marie Destandau, Sejla

Cebiric, François Goasdoué, Ioana Manolescu: Browsing Linked Data Catalogs with LODAtlas. ISWC (2) 2018: 137-153

[Brickley et al., 2019] Dan Brickley, Matthew Burgess, Natasha F. Noy: Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. WWW 2019: 1365-1375

[Wang et al., 2021] Xiaxia Wang, Tengteng Lin, Weiqing Luo, Gong Cheng, Yuzhong Qu: Content-Based Open Knowledge Graph Search - A Preliminary Study with OpenKG.CN. CCKS 2021: 104-115

[Wang et al., 2022] Xiaxia Wang, Tengteng Lin, Weiqing Luo, Gong Cheng, Yuzhong Qu: CKGSE: A Prototype Search Engine for Chinese Knowledge Graphs. Data Intell. 4(1): 41-65 (2022)

第十三章 知识图谱交叉前沿

张文 2，毕祯 1，朱渝珊 1，李娟 1，陈卓 1，陈华钧 1

1. 浙江大学 计算机科学与技术学院，浙江省 杭州市 310007

2.浙江大学 软件学院，浙江省 宁波市 315048

一、知识图谱交叉前沿简介

知识图谱技术是知识图谱建立和应用的技术，是语义 Web、自然语言处理和机器学习等的交叉学科。在实际应用中，知识图谱在知识融合、语义搜索和推荐、问答和对话系统以及大数据分析和决策等以数据为中心的应用中已经凸显出越来越重要的应用价值。不仅如此，近年来知识图谱技术还被应用到了不同的学科中，帮助解决领域特定的问题。

下面我们就知识图谱在区块链、物联网、软件工程以及视觉分析四个领域中的应用进行知识图谱交叉前沿的介绍，每个部分包括领域简介、知识图谱赋能领域应用、业界应用案例以及技术展望和发展趋势。

二、知识图谱 + 区块链

1. 区块链和分布式账本简介

区块链使用分布式账本的技术，它可以在开放的 P2P 网络内对电子数据共享、复制和同步。每个节点都可以参与监控交易的合法性，并为交易结果作证。区块链构成一个多中心网络，对完整的交易日志和执行结果达成共识，具有不可变、可追溯和可确权的特点，为区块链技术奠定了坚实的可信基础。

利用分布式账本技术，区块链可以记录公开知识的产生、发展和推演，全面追踪公开知识的价值和归属。比如多中心化的区块链网络提供可信的基础设施，跟踪开放知识发展的过程，并保证数据的真实性。

2. 知识图谱赋能区块链应用

知识是对事实信息的描述，也是对世界的认识和理解。日常生活中，知识是通过感受发现和学习获得的。随着万维网的蓬勃发展，将知识库作为开放数据发布已引起了极大的关注。知识的生产、转化、交换和消费形成了社会知识的价值链。

构建在去中心化分布式网络上的开放知识图谱会有诸多挑战，包括激励机制、所有权管理、可追溯性、信任和隐私机制等。然而现有的知识图谱相关的平台没有考虑这些问题，既

阻碍了知识的共享和互联，也无法保证知识的真实性和及时性。

同时知识图谱的价值网络不仅包括知识的贡献者，还包括知识的使用者等。知识构建和消费的全过程将逐步丰富知识网络，本质上增加知识价值。这一过程对多重粒度的知识奖励、自主知识、对抗性攻击和知识问责等提出了新的要求。对于知识图谱赋能区块链应用，构建信任价值链来支持知识的生命周期是核心问题。

3. 业界应用案例

- Knowledge Market [Lin et al., 2019]是基于区块链技术的知识有偿共享平台。它使用联盟链架构，知识商品是利用传感器获取的数据，经过训练的机器学习模型在链上存储并且交易。
- OpenKG Chain [Chen et al., 2021]以 OpenKG 社区为主导，是国内首个基于区块链技术开发的知识共享基础架构设施。它通过利用粗粒度知识（共享知识图谱数据集和工具集平台 OpenKG.CN）和细粒度知识（知识众包平台 OpenBase），构建了基于知识共享的区块链平台。OpenKG Chain 同时首次定义了知识衡量指标 K-Point 和参与贡献指标 OpenKG Token，其作为底层的基础结构目前已经应用在 OpenKG 社区中相关服务中。
- FactChain [朱 et al., 2022]是基于区块链技术的众包知识融合和知识管理系统。它使用联盟链架构，在技术层面设计了适应区块链架构的置信度加权等算法，为去中心化多源知识共享场景提出了新的解决方案。
- Knowledge Blockchain [Fill et al., 2018]用于以不可变和防篡改的方式存储企业模型中表达的知识。区块链允许通过恢复分布式计算和加密技术以防篡改和不可撤销的方式存储信息，Knowledge Blockchain 应用于知识管理领域，基于以企业模型的形式对知识进行解释。
- EpiK Protocol 铭识协议是全球首个 AI 数据的分布式存储协议。EpiK Protocol 旨在构建去中心化的超大规模知识图谱，通过去中心化存储技术、DAO 和通证经济模型，组织并激励全球社区成员将人类各领域知识梳理成知识图谱。同时该协议设想通过设计更合理的数据封装、更加宽容的惩罚措施，和 E2P 的数据上传模式等等来解决算力竞赛的问题。

4. 技术展望与发展趋势

知识是有价值的，知识之间建立关联可以进一步增加知识的价值。知识图谱区块链技术有极大的应用前景。从知识生产层而言，知识众包、知识抽取、知识校验、知识对齐融合、

知识补全等过程，可以利用区块链的技术体系去完善知识生产的不确定性。当知识已经产生，从知识传播层的角度来看，知识确权、知识溯源等也是结合设计区块链合约等所必须要考虑的因素。最后从知识消费的角度来看，知识图谱与区块链的交叉应用未来将会在知识问答系统、联邦知识学习等等与知识图谱相关的任务中拥有非常广阔的应用价值。

三、知识图谱 + 物联网

1. 物联网简介

物联网（Internet of things，IoT）即“万物相连的互联网”，是互联网基础上的延伸和扩展的网络，将各种信息传感设备与网络结合起来而形成的一个巨大网络，实现任何时间、任何地点，人、机、物的互联互通。物联网是新一代信息技术的重要组成部分，物联网的发展为各种物理对象提供无处不在的智能和普遍的互联。

知识图谱可服务于物联网中的设备管理和数据管理、推荐系统、设备预测、智慧医疗等领域。

2. 知识图谱赋能物联网应用

在物联网设备、数据的管理应用中，常常面临以下问题：不同的物联网设备无缝集成到物联网系统中不同设备间存在通信鸿沟和异构性，Xie 等人提出了利用基于知识图谱的新型多层物联网中间件方法，通过构建物联网设备到知识图谱的映射，以及物联网知识图谱服务几个部分，完成复杂物联网系统中的物联网设备管理，消除了不同设备通信间存在的鸿沟和异构性[Xie et al., 2021]；当前物联网设备产生的不同质量数据的评估方式并不通用，Khokhlov 等人提出了一个基于知识图谱的通用数据质量（DQ）评估框架，可将 DQ 评估纳入广泛的物联网应用程序[Khokhlov et al., 2020]，该框架支持选择 DQ 指标并实现它们的积分计算，有助于在物联网应用程序设计中实现更高级别的 DQ 评估集成自动化；工业领域的信息互动慢，机器维护过程困难，Hossayni 等人提出一种知识图谱 SemKoRe，旨在改善工业领域的机器维护[Hossayni et al., 2020]。SemKoRe 与供应商无关，它帮助原始设备制造商捕获、共享和利用其位于世界各地的客户机器产生的故障知识，并通过减少故障诊断时间和集中由世界各地的专家和技术人员提供的机器维护知识，显著增强了维护过程。知识图谱作为统一的数据管理方式，可以克服数据之间的异质性，进行统一的数据质量管控，快速进行不同类信息的查询和检索。因此，知识图谱可应用于物联网设备、数据的管理。

在物联网推荐系统中，和其他领域的推荐系统类似，在用户查询几个关键词的情况下，准确找到本体类一直是个难题，同时新成立的物联网平台存在的冷启动问题，Wang 等人提

出了一种基于语义和知识图谱构建的物联网本体类推荐方法来解决上述挑战[Wang et al., 2021]。首先从科技期刊中收集与物联网相关的文章。然后通过从文章中提取实体和关系，构建三元组，形成知识图谱。使用构建的知识图谱作为补充关系，以推荐与用户输入文本语义不同但与之相关的本体类。最后通过整合基于语义的推荐和基于知识图谱的推荐，可以得到最终推荐的物联网本体类。现有的推荐方法常常忽略了用户偏好与其他群组偏好之间可能存在的相关性，忽略用户隐式偏好，难以满足用户的精准推荐需求，Yao 等人提出了一种基于协同过滤和知识图谱的组发现方法[Yao et al., 2022]，首先利用注意力机制从知识图谱和用户与服务的交互中学习服务实体的嵌入，从而实现用户自己的偏好嵌入。考虑到相似用户的偏好将有助于获得准确的目标用户偏好，然后通过协同过滤和 word2vec 方法训练用户的最终偏好嵌入。最后，通过两阶段聚类方法将用户划分为不同的组。知识图谱通过关联不同类型的实体，构建实体间的关系，能够为推荐任务提供额外的背景知识，缓解新平台面临的推荐冷启动问题，同时作为用户和群组信息组织的媒介，不同用户和群组在图上的连通性可以用于挖掘用户和群组的偏好。

在物联网设备运维中，在实现物联网的无人值守操作时，现有方法难以准确预测设备将执行的动作并满足用户的个性化需求，You 等人提出了一种基于物联网时间知识图谱（TKG）和长短期记忆（LSTM）模型来预测设备状态的新方法[You et al., 2020]。首先为物联网构建了一个 TKG，它除了描述物联网静态对象的语义概念和关系，还将时间作为一个独立的维度来描述物联网中不断变化的时间序列数据，为物联网中的对象和不断变化的时间序列数据提供丰富的语义信息。然后，利用 LSTM 在序列学习方面的优势，学习 TKG 中语义信息的时序特征，挖掘隐含在 TKG 语义信息中的用户行为习惯，实现对设备状态的智能预测；由于设备、消费者或过程的动态性，难以准确预测设备状态并改进系统行为，Gómez-Berbís 等人提出了一种基于物联网数据管理和知识图谱的语义数字孪生方法[Gómez-Berbís et al., 2019]。该方法使用企业知识图来支持数字孪生，其基于物联网数据和知识图谱构建数字孪生的方法步骤包括设置数字孪生参数、采集数据、在知识图谱中明确定义单元时间、速率和生产流程等概念、定义算法、优化系统，最终实现预测设备状态并改进系统。在物联网设备预测服务中应用动态知识图谱对应用场景进行建模，不仅可以捕捉单位时间内事物之间的关联关系，还可以捕捉单位时间之间相同事物的变化状态，辅助准确的设备状态预测。

物联网智慧健康领域发展迅速，但尚未形成稳定的作者和机构网络，该领域的知识库也已初步形成，但基于此的研究尚不全面，Yang 等人采用文献计量法，基于 2003-2019 年该领域 9561 份文献数据，对时间分布、空间分布、文献共引、关键词等方面的分析进行可视化分析[Yang et al., 2020]。该研究为相关领域的研究人员提供全景知识支持，帮助他们了解

基于物联网的智慧健康研究领域的研究现状、未来趋势和热点；目前还缺乏适用于远程医疗或移动设备的中文综合医学知识图谱，Liu 等人提出了一种新的中国医疗保健知识图谱，可用于基于物联网的移动设备[Liu et al., 2021]。该智能医疗知识图谱利用深度神经网络结合自注意力生成，它不仅可以对疾病进行分类，还可以提出治疗建议。更重要的是，该医学知识图谱可以在为患者和医生服务的远程会诊平台中应用到实践中，大大方便了疾病诊断和治疗。

3. 业界应用案例

西门子艾闻达（原西门子物联网服务事业部）以知识图谱为抓手，结合机器学习的预测性维护系统 SiePA，成功帮助企业推进数字化转型，为企业建立了涵盖智能预警到智能诊断的生产管理闭环，保证了生产的可靠性和安全性。

美国初创公司 Stardog 开发了企业级的知识图谱平台，具备高度的灵活性和可用性，结合图形存储和虚拟化功能，Stardog 为企业提供了统一、查询、搜索和分析数据的一系列解决方案，企业可以通过知识图谱来统一数据，实现高效的数据集。目前 Stardog 技术已经成功应用于施耐德电气，构建了智能建筑领域的物联网知识图谱，融合了建筑物管理数据、建筑物舒适度调节数据以及电源监控数据，集成了无数物联网传感器和系统的信息，帮助实现更优化的智能建筑运营。

4. 技术展望与发展趋势

随着物联网（IoT）技术的飞速发展，知识图谱在赋能物联网的相关应用上发展着重要作用，知识图谱在服务于物联网设备管理、设备预测、物联网推荐、智慧医疗等方面都非常有帮助。在物联网设备和数据管理方面，基于知识图谱的物联网中间件提供了灵活的物联网设备访问，有助于实现自适应远程监控，相关方案还可以应用于其他物联网相关行业，如环境监测、保护区监测、山林、农村农业等。在物联网推荐系统中，利用时间因素来分析用户的短期和长期利益，是未来值得研究的方法。在物联网设备预测服务中，针对智能家居的状态预测是一个具有广泛前景的应用场景。在智慧医疗中，利用知识图谱为相应的患者推荐更合适准确的医生仍然是一个需要解决的问题。此外，进行多模态智慧医疗知识图谱的图文数据融合也是一个方向，模型加入图像数据有利于实现更全面的预测，为患者和医生服务。

四、知识图谱 + 软件工程

1. 软件工程简介

软件工程的概念于 1968 年 NATO（北大西洋公约组织）在德国 Garmish 召开的学术会议上由 Feitz Bauer 首先提出。软件工程是运用工程的、数学的、计算机等科学的概念、方法和原理来指导软件开发和维护的一门学科，或者说是研究如何开发软件的一门学科。现代软件工程定义又包括：运用现代科学技术知识来设计并构造计算机程序及为开发、运行和维护这些程序创建必需的相关文件资料；开发、运行、维护和修复软件的系统方法；建立并使用完善的工程化原则，以较经济的手段获得能在实际机器上有效运行的可靠软件的一系列方法等。其本质特征涵盖了关注于大型程序的构造、中心课题是控制复杂性、软件经常变化、开发效率非常重要、和谐合作是开发软件的关键等多个方面。

软件工程的目标有：付出较低的开发成本，达到要求的软件功能，取得较好的软件性能，开发的软件易于移植，需要较低的维护费用，能按时完成开发工作，及时交付使用。此外，软件工程的特点涉及到多学科、多目标、多阶段。其中多目标不仅关注软件成品的功能，并且关注软件开发过程的成本、进度、可靠性以及可维护性。软件工程方法学中主要有三要素：过程、方法和工具。过程指开发一个软件产品所需步骤、需完成的各项任务及对这些任务的组织和管理；方法指完成软件工程项目的技术手段，如软件需求分析、设计、编码、测试和维护等；工具指自动或半自动地支持软件的开发和管理文档的生成。软件生命周期则包含了问题定义、可行性研究、需求分析、总体设计、详细设计、编码和单元测试、综合测试（集成测试和验收测试）、使用与维护。

2. 知识图谱赋能软件工程应用

软件复用是软件开发中避免重复劳动的解决方案。虽然软件复用可提高软件开发效率，但随着软件规模不断扩大以及软件复杂度日益提高，开发者的学习成本越来越高，复用软件项目变难。为提高软件开发效率，如何有效利用软件项目整个生命周期中积累的大量数据，如源代码、邮件列表、缺陷报告和问题文档等成为关键。针对这些多源异构数据呈分散形态，缺乏全局、统一的组织整理，彼此之间也缺乏关联；大量信息是以无结构文本的形式表示的，如代码标识符、代码注释、邮件、用户手册、缺陷描述，文本信息具有很高的随意性和模糊性，复用者可以通过关键词来对无结构文本进行检索的效果不尽如人意等问题，为有效地组织和利用多源异构软件大数据，更好地进行软件理解和复用，知识图谱赋能的软件工程应用被关注。知识图谱赋能的软件工程应用主要包含软件项目知识图谱的构造方法，基于软件项目知识图谱的智能问答方法，以及代码语义标签自动生成方法。

软件项目知识图谱的构建方法考虑针对种类繁多，而且可能层出不穷的各种扩展知识和新来源、新格式的数据，如何设计一个框架以满足知识图谱构造的自动性和可扩展性。目前的工作[邹 et al., 2021]设计了知识抽取和知识融合两个阶段，基于不同数据源，不同的知识自动抽取算法对应到特定的数据类型，自动抽取实体和关系形成特定数据源的子图谱；这些子图谱则通过不同知识融合手段建立跨数据源的关联，将原本独立的子知识图谱整合成一个完整、统一的软件项目知识图谱。构建的知识图谱又可以通过知识推理方法推断更多可能的知识。另一个方法[李 et al., 2017]则提出更完整的，覆盖了知识图谱应用的软件知识图谱构建框架。该框架不仅是知识提取以及知识融合，也包括了知识图谱的存储管理模块与软件知识检索模块构成，设计了形式化检索和文本检索两种机制展现检索结果。

基于软件项目知识图谱的智能问答方法旨在将知识图谱融入机器对无结构文本的处理过程中，从而为复用者提供准确有效的智能问答服务，进而提高软件复用过程效率和质量。针对许多软件知识图谱的数据量较大、数据类型较多，开发人员在查询过程中存在较高的学习成本等困难。早期工作提出通过构造推理子图实现基于自然语言的知识图谱查询方法，对自然语言查询语句进行解析后与知识图谱中元素匹配，将自然语言转换为知识图谱上的推理子图后，将推理子图转化为 Cypher 查询语句到 Neo4j 图数据库上执行。但基于关键词匹配的方法不仅需要对大量搜索结果进行筛选且准确率低，智能问答方法的提出[邹 et al., 2021]希望通过软件项目的知识图谱计算不同单词的潜在语义相关度对候选文本进行评估。该方法识别出与一段软件文本相关的代码实体集合，将软件文本的语义结构化地表示为一个带权重的代码实体的集合。基于知识图谱表示学习技术，以软件文本间语义相似度为核心，度量软件文本间的语义相似度，选择最合适的文本片段返回给复用者。这种利用代码实体之间的结构依赖关系实现了对文本之间的潜在语义关联的更直接、更有效的挖掘与利用，从而显著地改进文档搜索的效果。

代码语义标签自动生成方法考虑到代码是重要的软件开发资源，现有通过代码文本的信息检索难以实现精准的代码搜索等问题，以及反映代码整体意图和主题的语义标签对于改进代码搜索、辅助代码理解的重要意义，希望生成高质量的代码语义标签。该方法[邢 et al., 2021]通过基于 API 文档和软件开发问答文本的概念和关系抽取构造软件知识图谱作为代码语义标签生成的基础。针对给定的代码,该方法识别并抽取出通用 API 调用或概念提及,并链接到软件知识图谱中的相关概念上。在此基础上,进一步识别与所链接的概念相关的其他概念作为候选,然后按照多样性和代表性排序产生最终的代码语义标签。

3. 业界应用案例

目前已有部分公司针对企业关注的软件项目，构建了对应的软件项目知识图谱，并对数

据来源，应用领域和知识图谱中实体和关系数作了相关统计。以神州数码信息系统有限公司为例，该公司智慧城市公共信息服务开发平台规模大、功能繁多复杂、平台文档庞杂，这使得软件开发人员熟练使用该平台进行软件项目开发的学习成本较高。通过对大规模软件项目数据（源代码、需求文档、设计文档、测试文档、用户手册及参考文献）的知识化，有效提高了智慧城市领域软件构造的效率与质量。除了公司，面向 Apache 开源社区的软件项目知识图谱也通过[邹 et al., 2021]中提到的方法被构建，设计到的相关数据有：源代码、Git 版本控制数据、邮件列表数据、JIRA 事务跟踪数据和 Stack Overflow 上的相关问答对。

4. 技术展望与发展趋势

当前知识图谱赋能软件工程项目的相关应用，对于软件项目的复用有着重要作用，因此知识图谱和软件工程的结合必不可少。随着软件开发过程中的更多数据类型，软件知识图谱中知识实体的扩充，实体之间更多语义关联的建立，探索更精准的交互式智能问答服务、智能推荐服务以及挖掘更多跟软件复用强相关的应用等也十分重要。此外，由于软件知识图谱的相关应用依赖于构建的图谱质量，高质量软件项目知识图谱的自动构建也是一个需要不断突破的方向。

五、知识图谱 + 视觉分析

1. 视觉分析简介

计算机视觉是人工智能界在 21 世纪进步最大、发展最快的领域之一。根据 Global VIEW 的研究，全球计算机视觉市场规模在 2020 的价值为 113 亿 2000 万美元，预计从 2021 到 2028 的复合年增长率为 7.3%。人工智能计算机视觉的使用案例几乎不计其数，其中最受欢迎的是无人机以及自动和半自动车辆。此外，由于计算机视觉的最新进展，人工智能现在已成为各个行业的必需品，例如教育、医疗保健、机器人、消费电子、零售、制造等。

尽管自 20 世纪 60 年代以来计算机视觉已经取得了很大进展，但就研究和开发而言，它仍然是一个很大程度上尚未开发的领域。这主要是因为人类视觉本身极其复杂，而计算机视觉系统相比之下就受到了影响。即使是在不同的年龄段，人们也需要几秒钟的时间才能认出照片中的朋友，我们记忆和存储面孔以供未来识别的能力似乎是无限的。然而，很难想象一台计算机处理几乎相似的事情需要多少工作量。

为了突破其在知识层面的瓶颈，很多研究者都将计算机领域的研究从纯感知层转向认知层，也就是具有一定视觉推理能力，当然视觉推理在很多实际应用中也非常重要，而实现这一目标的其中一个目前看起来会比较有效的方法就是结合知识图谱。具体来说，计算机视觉

可以细分为多个方向，这其中，与知识图谱结合较紧密的，具有代表性的方向我们总结为一个大类：视觉推理（Visual Reasoning）。

2. 知识图谱赋能视觉分析应用

视觉推理是一个很广泛的概念，凡是视觉领域涉及到逻辑/知识/推理的场景，我们均可以归纳到视觉推理的范畴。下面对知识图谱有所应用的图像注释、视觉常识生成以及视觉问答三种任务进行展开介绍。

1）图像注释（Image Caption）

Image Caption（图像注释）仅仅是对图片进行描述，目的只是把图片内容描述正确、清楚，一般而言是不带有感情色彩的。在描述一张图片时，可能只利用图片上的知识是不够的。知识图谱的建立与应用是迅速发展的一个领域，因此在 Image Caption 中通过知识图谱引入外部知识是值得研究的一个方向。在这方面的尝试包括[Lu et al., 2018]，其采用类似于 Neural Baby Talk 的方法，先利用 Encoder-Decoder 获得一个含有实体空槽的语言模板，再使用实体填槽。该工作采用将训练数据的标签相近的图片的描述作为上下文，从中抽取命名实体输入知识图谱，选择图谱中概率最高的实体组合作为插槽输入。这种引入外部知识的方法大大提高了语义丰富度。

和图像注释不同，对 Visual Storytelling 而言讲故事是需要有前后逻辑和一定的想象力的，对比 Image Caption,它的挑战大很多，不仅是对图像中的物体进行客观描述，还要找寻图片和图片之间有关联物体的关系，即不能简单把图片客观翻译出来，而是要联系多张图片的内容，找出之间的关联，再来写故事，就像我们小学的时候看图写话差不多，给你几张图编一段故事。难点在于要让计算机去理解图片内容，并且联系起来，但是这几年也有一些这方面的工作做的不错的，通过各种中间层来进行图片——故事的连接。KG-Story [Hsu et al., 2020]把知识图谱加入到 Storytelling 的过程中从每张图片中提取一组单词，然后在知识图谱中进行搜素，找出图像中词对之间的潜在关系，使用故事生成器，利用前两个阶段得到的所有词汇和知识生成故事。在这过程中知识图谱起到一个知识库的作用。

传统知识图谱以外，多模态知识图谱也发挥了很大作用。Zhao 等人研究了实体感知的图像描述（看图说话）问题，旨在通过利用相关文章中的背景知识来描述与图像相关的命名实体和事件[Zhao et al., 2021]。由于命名实体的长尾分布，很难建立命名实体和视觉信息之间的关联。此外，从背景文章中提取细粒度的实体和关系以生成有关图像的描述非常具有挑战性。为了应对这些挑战，这项研究构建了一个多模态知识图谱，将视觉对象与命名实体相关联，并从网络收集的外部知识中同时建立实体之间的关系，然后对多模态知识图谱通过图注意力机制集成到字幕生成模型中。

2）视觉常识生成（Visual Commonsense Generation）

视觉常识生成任务是 2020 年新提出的任务，比看图说话要更难，因为需要常识推理来预测给定图像之前或之后的事件。Xing 等人提出了一个知识增强的多模态 BART（KM-BART）模型[Xing et al., 2021]，这是一种基于 Transformer 的 Seq2Seq 模型，能够从图像和文本的多模态输入中推理常识知识。具体来说这项研究将生成式 BART 架构调整为具有视觉和文本输入的多模态模型，并进一步开发了新颖的预训练任务，以提高视觉常识生成（VCG）任务的模型性能。基于知识的常识生成的预训练任务通过利用在外部常识知识图谱上预训练的大型语言模型中的常识知识，提高了 VCG 任务的模型性能。

回答关于图像的复杂问题是机器智能的一个雄心勃勃的目标，它需要对图像、文本和常识的联合理解，以及强大的推理能力。与视觉常识生成类似的，视觉常识推理任务（Visual Commonsense Reasoning, VCR）于 2018 年由华盛顿大学的研究人员首次提出，任务旨在将图像和自然语言理解二者结合，验证多模态模型高阶认知和常识推理的能力，让机器拥有“看图说话”的能力,例如 VCR 能够通过图片中人物的行为，进一步推理出其动机、情绪等信息。

最近多模态 Transformer 在视觉常识推理任务上取得了很大的进展，通过跨通道注意力层共同理解视觉对象和文本标记。然而，这些方法并没有利用场景的丰富结构和对象之间的交互作用，而这些在回答复杂的常识问题时是必不可少的。Wang 等人提出了一个场景图增强图像-文本学习（SGEITL）框架，将视觉场景图纳入常识推理[Wang et al., 2021]。为了利用场景图结构，在模型结构层次上，作者提出了一种多跳图转换器来正则化各跳间的注意力交互。在预训练方面，提出了一种场景感知的预训练方法，利用视觉场景图中提取的结构知识。

同时，现有的方法仅考虑了区域-词的相似性来实现视觉和语言域之间的语义对齐，忽略了视觉概念和语言词之间的隐式对应（如词-场景、区域-短语和短语-场景）。Wu 等人提出了一种层次语义增强方向图网络，设计了一个模态交互单元（MIU）模块，通过聚合层次视觉-语言关系来捕获高阶跨模态对齐。同时提出了一种新颖的层次语义增强方向图网络用于视觉常识推理任务，该网络能够捕获不同模式间的高阶相关性，并执行清晰的推理过程[Wu et al., 2021]。

3）视觉问答（Visual Question Answering，VQA）

视觉的捕捉与理解，学习与感知知识，诸多信息促进着我们对世界的认知。VQA 作为多模态领域的典型场景，目的是结合视觉的信息来回答提出的问题。其首次在 15 年提出，涉及的方法从联合编码，到双线性融合，再到注意力机制，组合模型，场景图，从而引入外部知识，进行知识推理，以及使用图网络，多模态预训练语言模型等，近年来发展迅速。传

统 VQA 仅凭借视觉与语言信息的组合来回答问题，而近年来许多研究者开始探索外部信息对于解决 VQA 任务的重要性。

推理与知识的实际存储进行分离，是外部知识 VQA 相关论文所持的观点。Wu 等人提出将自动生成的图像描述与外部的知识库融合，以实现对问题的预测[Wu et al., 2016]。其中生成图像描述的方法流程如下：给定一张图像，先预测图像中各种属性，然后再将这些属性代替之前的 CNN 图像特征，输入到 RNN 当中生成语句。这个简单的操作使他们的图像标注模型在当年 COCO 图像标注大赛上排名第一。添加中介属性以减小双模态鸿沟的方法，也用在了本文中。首先用 CNN 提取图片特征属性，然后利用这些检测到的属性，使用 SPARQL 查询语句从知识库比如 DBpedia 中提取出图像相关描述的一个段落，利用 Doc2Vec 对这些段落编码。同时，根据图片特征属性使用 Image Caption 方法形成图像对应的段落特征表达。最后将上面两种信息以及编码的属性结合在一起并输入作为一个 Seq2Seq 模型的初始初始状态，同时将问题编码作为 LSTM 的输入，利用最大似然方法处理代价函数，预测答案。

既然知识和推理对 VQA 都很重要，那么就可以考虑将它们两个结合在一起，进行显式推理。和以往直接把图像加问题直接映射到答案不同，Wang 等人提出的模型的答案是可追溯的，就是通过查询语句在知识图谱中的搜索路径可以得到一个显式的逻辑链[Wang et al., 2018]。这也是一种全新的能够进行显式推理的 VQA 模型。并且，他们提出了一种涉及外部知识的 VQA 任务。它首先会通过解析将问题映射到一个知识图谱查询语句从而能够接入到已有知识库中。同时将提取的视觉概念的图链接到 DBpedia 里面，同期发表的 FVQA 是对其的改进和梳理，并且贡献了这方面很重要的数据集：除了一般的图片、问题、回答以外，这个数据集还提供了支撑这一回答的事实集合（参考数据来源于 DBpedia、Conceptnet、WebChild 三个数据库），共包括 4216 条事实。某种意义上来说，该数据集是基于事实去针对性构建的。在实际的数据中，事实以关系三元组的形式表示，其中的关系使用来自于数据库中已有的定义。

模型的第一部分检测图像中的视觉概念，然后将他们与知识库对齐并连接到子图中。第二步将自然语言式的问题映射到一个查询类型，然后相应地确定关键的关系类型，视觉概念和答案源。再根据上面的信息构建一个特殊的查询会去请求上一步当中建立好的图，找到所有满足条件的事实。最后通过关键词筛选得到对应问题的答案。

前文提出的方法大多类似于组合模型。此外，近几年也有涉及到图来解决外部知识 VQA 问题的方法[Narasimhan et al., 2018]。该文章的作者基于 FVQA 数据集，把之前深度网络筛选事实的这一训练过程用图卷积网络代替，成为一个端到端的推理系统，用于具有知识库的

视觉问题解答。一共分为七个步骤，给定图像和问题，首先使用相似性评分技术根据图像和问题从事实空间获得相关事实。使用 LSTM 模型从问题预测关系，筛选事实来进一步减少相关事实及其实体的集合。然后分别进行图像视觉概念提取，问题的 LSTM 嵌入，以及事实词组的的 LSTM 嵌入，将图像的视觉概念 Multi-Hot 向量和问题的 LSTM 嵌入向量组合，并与每一个实体的 LSTM 嵌入拼接，作为一个实体的特征表示，同时也是作为 GCN 模型里图上的一个节点。图中的边代表实体之间的关系。最后将 GCN 输出的每一个实体节点特征向量作为多层感知机二元分类模型的输入，最后输出的结果通过 argmax 得到最终的决策结果。

在回答给定上下文（例如图像）的问题时，可以将观察到的内容与常识无缝结合在一起[Narasimhan et al., 2018]。对于自然参与我们日常工作的自主代理和虚拟助手，在最常根据上下文和常识回答问题的地方，利用观察到的内容和常识的算法非常有用。许多前述方法集中在问题回答任务的视觉方面，即通过结合问题和图像的表示来预测答案。这与描述的类人方法明显不同，后者将观察与常识相结合。为此，相关研究设计了一种从问题中提取关键字并从知识库中检索包含这些关键字的事实的方法。但是，同义词和同形异义词构成了难以克服的挑战。为了解决这个问题，作者开发了一种基于学习的检索方法。更具体地说，他们的方法学习事实和问题图像对到嵌入空间的参数映射。为了回答问题，他们使用与所提供的问题图像对最一致的事实。知识库中的事实是根据视觉概念（例如，对象，场景和从输入图像中提取的动作）进行过滤的。然后将预测的查询应用于过滤后的数据库，从而获得一组检索到的事实。然后，在检索到的事实和问题之间计算匹配分数，以确定最相关的事实。最正确的事实构成了问题答案的基础。

一些工作[Zhu et al., 2020]的出发点是将图像表示成一个多模态的异构图，其中包含来自不同模态三个层次的信息（分别是视觉图、语义图和事实图），来互相补充和增强 VQA 任务的信息。具体来说，视觉图包含了图像中的物体及其位置关系的表示，语义图包含了用于衔接视觉和知识的高层语义信息，事实图则包含图像对应的外部知识，它的构造思想参考了 Out-of-the-box 模型。然后进行每个模态内的知识选择：在问题的引导下确定每个节点和边在内部图卷积过程中的分数权重占比，然后进行常规的更新操作。也就是说在跨模态之前，先独立选择单个模态内有价值的证据，让和问题相关性强的节点及边，在图内部卷积过程中占更大的权重。这三个模态内部的卷积操作都是相同的，只是节点和边的表示不同。最后，跨模态的知识推理是基于知识选择的结果。考虑到信息的模糊性，不同图很难显式地对齐，所以作者采用一种隐式的基于注意力机制的异构图卷积网络方法来关联不同模态的信息，从不同层的图中自适应地收集互补线索并进行汇聚。包括视觉到事实的卷积和语义到事实的卷积。比如视觉到事实的卷积场景中，对于事实图中的每个节点 v_i，计算视觉图中每个节点 v_j

和它在问题引导下的相似度注意力分数，越互补的节点它的相似度分数就越高，然后根据这个分数对视觉图加权求和，得到事实图中每个节点来自视觉图层的事实互补信息。

现有的许多方法采用流水线的模式，多模块分工进行跨模态知识处理和特征学习，但这种模式下，中间件的性能瓶颈会导致不可逆转的误差传播（Error Cascading）。此外，大多数已有工作都忽略了答案偏见问题——因为长尾效应的存在，真实世界许多答案在模型训练过程中可能不曾出现过（Unseen Answer）。Chen 等人提出了一种适用于零样本视觉问答（ZS-VQA）的基于知识图谱的掩码机制，更好结合外部知识的同时，一定程度缓解了误差传播对于模型性能的影响[Chen et al., 2021]。并在原有 F-VQA 数据集基础上，提供了基于 Seen / Unseen 答案类别为划分依据的零样本 VQA 数据集（ZS-F-VQA）。实验表明，其方法可以在该数据集下达到最佳性能，同时还可以显著增强端到端模型在标准 F-VQA 任务上的性能效果。

总而言之，形形色色的方法各有千秋。在实际应用中，可以根据不同方法的优劣和实际场景的条件选择合适的 VQA 模型。目前来说解决 VQA 问题主要方向主要是三个大方向：改善模型对于文本与图像的表达能力；可解释性与视觉推理；外部知识。其中知识图谱在这三个方向中都有涉及。起到的作用分别对应于：用图网络来捕捉信息联系，通过三元组来提供与描述事实并进行解释与答案追溯，以及引入外部语料库，组织实体关系和 SPARQL 查询语句。此外，现有的模型默认训练集与测试集具有独立同分布的特质，但现实往往不尽如人意，也就是说同分布的假设大概率要打破。正如三位图灵奖得主最近发表的文章 Deep Learning for AI[Bengio et al. 2021]中所强调的核心概念——高层次认知。将现在已经学习的知识或技能重新组合，重构成为新的知识体系，随之也重新构建出了一个新的假想世界（如在月球上开车），这种能力是人类天生就被赋予了的，在因果论中，被称作“反事实”能力。现有的统计学习系统仅仅停留在因果关系之梯的第一层，即观察，观察特征与标签之间的关联，借助知识图谱有望完成更高层次的事情。

3. 业界应用案例

- 百度：多模态语义理解是人工智能领域重要研究方向之一，如何让机器像人类一样具备理解和思考的能力，需要融合语言、语音、视觉等多模态的信息。百度在该领域取得突破，首次将场景图（Scene Graph）知识融入多模态预训练，提出业界首个融合场景图知识的多模态预训练模型 ERNIE-ViL[Yu et al., 2021]。百度研究者将场景图知识融入到视觉-语言模型的预训练过程，学习场景语义的联合表示，显著增强了跨模态的语义理解能力。ERNIE-ViL 还在包括视觉常识推理、视觉问答、引用表达式理解、跨模态图像检索、跨模态文本检索等 5 项典型多模态任务中刷新了世

界最好效果。并在多模态领域权威榜单视觉常识推理任务（VCR）上登顶榜首，并在多模态领域权威榜单 VCR 上超越微软、谷歌、Facebook 等机构。为了验证场景图预测任务是否有效果，作者对比加入场景图预测和不加入场景图预测的效果区别，加入场景图后，效果有显著的提高。

- 阿里：将产品结构化知识作为一种独立于图像和文本的新的模态，称为知识模态，即对于产品数据的预训练，考虑了三种模态的信息:图像模态（产品图像）、文本模态（产品标题）和知识模态（PKG）。提出了一种在电子商务应用中新颖的知识感知的多模态预训练方法 K3M[Zhu et al. 2021]。K3M 通过 3 个步骤学习产品的多模态信息:（1）对每个模态的独立信息进行编码，对应 Modal-Encoding Layer，（2）对模态之间的相互作用进行建模，对应 Modal-Interaction Layer，（3）通过各个模态的监督信息优化模型，对应 Modal-Task Layer。并且在淘宝 4 千万商品上训练。结果显示，饿了么新零售导购算法，离线算法 AUC 提升了 0.2%绝对值；在线 AB-Test 实验，流量 5%，5 天：CTR 平均提高 0.296%，CVR 平均提高 5.214%，CTR+CVR 平均提高：5.51%；淘宝主搜找相似服务，离线算法 AUC 提升 1%，业务方反馈是很大的提升；目前在线 AB 测试中；阿里妈妈年货节商品组合算法，在线算法，基于 Embedding 的实验桶（5.52%）CTR 指标相较于另外 2 个实验桶（5.50%，5.48%）分别提高 0.02%、0.04%的点击率，相对提高分别为 0.363%、0.73%；小蜜算法团队低意愿下的相似商品的推荐，整体增加这一路的召回情况下，转化能有 2.3%到 2.7%左右的提升，相对提升 12.5%。之前版本相对提升 11%。后续扩展到其他场景。

4. 技术展望与发展趋势

随着人工智能技术的不断发展，知识图谱作为人工智能领域的知识支柱，以其强大的知识表示和推理能力受到学术界和产业界的广泛关注。近年来，知识图谱在语义搜索、问答、知识管理等领域得到了广泛的应用。多模态知识图谱与传统知识图谱的主要区别是，传统知识图谱主要集中研究文本和数据库的实体和关系，而多模态知识图谱则在传统知识图谱的基础上，构建了多种模态（例如视觉模态）下的实体，以及多种模态实体间的多模态语义关系。多模态知识图谱的应用场景十分广泛，它极大地帮助了现有自然语言处理和计算机视觉等领域的发展。多模态结构数据虽然在底层表征上是异构的，但是相同实体的不同模态数据在高层语义上是统一的，所以多种模态数据的融合对于在语义层级构建多种模态下统一的语言表示模型提出数据支持。其次多模态知识图谱技术可以服务于各种下游领域，例如多模态实体链接技术可以融合多种模态下的相同实体，可应用于新闻阅读，同款商品识别等场景中，多模态知识图谱补全技术可以通过远程监督补全多模态知识图谱，完善现有的多模态知识图

谱，多模态对话系统可用于电商推荐，商品问答领域。

参考文献

[Lin et al. 2019] Xi Lin, Jianhua Li, Jun Wu, Haoran Liang, Wu Yang. Making knowledge tradable in edge-AI enabled IoT: A consortium blockchain-based efficient and incentive approach. IEEE Transactions on Industrial Informatics, 15(12):6367-6378, 2019

[Chen et al 2021] Huajun Chen, Ning Hu, Guilin Qi, Haofen Wang, Zhen Bi, Jie Li, Fan Yang. OpenKG chain: A blockchain infrastructure for Open Knowledge Graphs. Data Intelligence, 3(2):205-227, 2021

[朱 et al. 2022] 朱向荣, 吴鸿祜, 胡伟. FactChain：一个基于区块链的众包知识融合系统. 软件学报, 33(10):0, 2022

[Fill et al. 2018] Hans-Georg Fill, Härer Felix. Knowledge blockchains: Applying blockchain technologies to enterprise modeling. HICSS, 1-10, 2018.

[Xie et al. 2021] Cheng Xie, Beibei Yu, Zuoying Zeng, Yun Yang, Qing Liu. Multilayer Internet-of-Things Middleware Based on Knowledge Graph. IEEE Internet of Things Journal, 8(4):2635-2648, 2021

[Khokhlov et al. 2020] Igor Khokhlov, Leon Reznik. Knowledge Graph in Data Quality Evaluation for IoT applications. WF-IoT, 1-6, 2020

[Hossayni et al. 2020] Hicham Hossayni, Imran Khan, Mohammad Aazam, Amin Taleghani-Isfahani, Noel Crespi. SemKoRe: Improving Machine Maintenance in Industrial IoT with Semantic Knowledge Graphs. Applied Sciences, 10(18):6325, 2020

[Wang et al. 2021] Xi Wang, Chuantao Yin, Xin Fan, Si Wu, Lan Wang. An IoT Ontology Class Recommendation Method Based on Knowledge Graph. KSEM, 666-678, 2021

[Yao et al. 2022] Kaiming Yao, Haiyan Wang, Yuliang Li, Joel J. P. C. Rodrigues, Victor Hugo C. de Albuquerque. A Group Discovery Method Based on Collaborative Filtering and Knowledge Graph for IoT Scenarios. IEEE Transactions on Computational Social Systems, 9(1):279-290, 2022

[You et al. 2020] Shujuan You, Xiaotao Li, Wai Chen. Intelligent Prediction for Device Status Based on IoT Temporal Knowledge Graph. ICCC, 560-565, 2020

[Gómez-Berbís et al. 2019] Juan Miguel Gómez-Berbís, Antonio de Amescua Seco. SEDIT: Semantic Digital Twin Based on Industrial IoT Data Management and Knowledge Graphs. CITI, 178-188, 2019

[Yang et al. 2020] Xuejie Yang, Xiaoyu Wang, Xingguo Li, Dongxiao Gu, Changyong Liang, Kang Li, Gongrang Zhang, Jinhong Zhong. Exploring emerging IoT technologies in smart health research: a knowledge graph analysis. BMC Medical Informatics and Decision Making, 20(1):260, 2020

[Liu et al. 2021] Wanheng Liu, Ling Yin, Cong Wang, Fulin Liu, Zhiyu Ni. Medical Knowledge Graph in Chinese Using Deep Semantic Mobile Computation Based on IoT and WoT. Wireless Communications and Mobile Computing, 5590754:1-5590754:13, 2021

[邹 et al. 2021] 邹艳珍, 王敏, 谢冰, 等. 基于大数据的软件项目知识图谱构造及问答方法. 大数据, 7(1), 2021

[邢 et al. 2021] 邢双双, 刘名威, 彭鑫. 基于软件知识图谱的代码语义标签自动生成方法. 软件学报, 2021

[李 et al. 2017] 李文鹏, 王建彬, 林泽琦, 等.面向开源软件项目的软件知识图谱构建方法. 计算机科学与探索, 11(6):851-862, 2017

[Lu et al., 2018] Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, Shih-Fu Chang. Entity-aware image caption generation. EMNLP, 4013-4023, 2018

[Hsu et al., 2020] Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao Kenneth Huang, Lun-Wei Ku. Knowledge-enriched visual storytelling. AAAI, 7952-7960, 2020

[Zhao et al., 2021] Wentian Zhao, Yao Hu, Heda Wang, Xinxiao Wu, Jiebo Luo. Boosting Entity-aware Image Captioning with Multi-modal Knowledge Graph. arXiv preprint arXiv:2107.11970, 2021

[Xing et al., 2021] Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, Roger Wattenhofer. KM-BART: Knowledge enhanced multimodal BART for visual commonsense generation. ACL, 525-535, 2021

[Wang et al., 2021] Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, Shih-Fu Chang. SGEITL: Scene Graph Enhanced Image-Text Learning for Visual Commonsense Reasoning. arXiv preprint arXiv:2112.08587, 2021

[Wu et al., 2021] Mingyan Wu, Shuhan Qi, Jun Rao, Jiajia Zhang, Qing Liao, Xuan Wang, Xinxin Liao. Hierarchical Semantic Enhanced Directional Graph Network for Visual Commonsense Reasoning. Trustworthy AI, 27-36, 2021

[Wu et al., 2016] Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources.

CVPR, 4622-4630, 2016

[Wang et al., 2018] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, Anton van den Hengel. Fvqa: Fact-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(10): 2413-2427, 2018

[Narasimhan et al., 2018] Medhini Narasimhan, Svetlana Lazebnik, Alexander G. Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. NIPS, 2659-2670, 2018

[Narasimhan et al., 2018] Medhini Narasimhan, Alexander G. Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. ECCV, 451-468, 2018

[Zhu et al., 2020] Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, Qi Wu. Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. IJCAI, 1097-1103, 2020

[Chen et al., 2021] Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z. Pan, Zonggang Yuan, Huajun Chen. Zero-shot visual question answering using knowledge graph. ISWC, 146-162, 2021

[Bengio et al. 2021] Yoshua Bengio, Yann LeCun, Geoffrey E. Hinton. Deep learning for AI. Commun. ACM 64(7): 58-65 (2021)

[Yu et al., 2021] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, Haifeng Wang. ERNIE-ViL: Knowledge Enhanced Vision-Language Representations through Scene Graphs. AAAI, 3208-3216, 2021.

[Zhu et al. 2021] Yushan Zhu, Huaixiao Zhao, Wen Zhang, Ganqiang Ye, Hui Chen, Ningyu Zhang, Huajun Chen. Knowledge Perceived Multi-modal Pretraining in E-commerce. ACM Multimedia, 2744-2752, 2021.