10 实践中的知识图谱

在本节中，我们将讨论过去几年出现的一些最突出的知识图谱。 我们首先讨论开放知识图谱，这些知识图谱已根据第 9 节中描述的指南和协议发布在 Web 上。稍后我们将讨论公司为各种应用程序创建的企业知识图谱。

**10.1 开放知识图谱**

我们所说的开放知识图谱，特指在开放数据哲学下发布的知识图谱，即“开放意味着任何人都可以出于任何目的自由访问、使用、修改和共享（最多以保持出处和开放的要求为前提） )”.35 许多开放知识图已经以链接开放数据集 [216] 的形式发布，这些图是根据开放数据哲学的链接数据原则（参见第 9.1.2 节）发布的（RDF）图。 许多最著名的开放知识图谱——包括 DBpedia [291]、YAGO [481]、Freebase [50] 和 Wikidata [515]——涵盖多个领域，代表了广泛多样的实体和关系； 我们首先依次讨论这些。 稍后我们将讨论一些其他（特定的）领域，这些领域目前可以使用开放知识图。 我们在本节中讨论的大多数开放知识图都是在 RDF 中建模的，按照关联数据原则发布，并通过转储 (RDF)、节点查找（关联数据）、图形模式 (SPARQL) 以及在某些情况下提供对其数据的访问。案例，边缘图案（三重图案碎片）。

**10.1.1 DBpedia**

DBpedia项目的开发是为了提取嵌入在维基百科文章[20]中的半结构化数据的图形结构表示，从而以统一的方式集成、处理和查询这些数据。通过链接到外部开放资源（包括图像、网页和外部数据集，如DailyMed、DrugBank、GeoNames、MusicBrainz、纽约时报和WordNet[291]），生成的知识图进一步丰富。DBpedia extraction framework由几个组件组成，对应于维基百科文章源的抽象、图形存储和序列化目的地、维基标记提取器、解析器和抽取管理器[44]。特定的提取器用于处理标签、摘要、中介语链接、图像、重定向、歧义消除页面、外部链接、内部页面链接、主页、类别和地理坐标。DBpedia知识图中的内容不仅是多领域的，而且是多语言的：截至2012年，DBpedia包含多达97种不同语言的标签和摘要[327]。DBpedia中的实体使用四种不同的模式进行分类，以满足不同的应用需求[44]。这些模式包括维基百科类别的简单知识组织系统（SKOS）表示、另一个伟大的本体论（YAGO）分类模式（在下文中讨论）、上层映射和绑定交换层（UMBEL）本体论分类模式，以及一个名为DBpedia本体论的定制模式，其中包含个人、地点、组织和工作等类[291]。DBpedia还支持实时同步，以便与动态维基百科文章保持一致[291]。

**10.1.2 Yet Another Great Ontology**

YAGO 同样从 Wikipedia 中提取图形结构数据，然后将其与 WordNet 的层次结构统一，以创建“具有高质量和覆盖率的轻量级和可扩展本体”[481]。 该知识图谱旨在应用于各种信息技术任务，例如机器翻译、词义消歧、查询扩展、文档分类、数据清洗、信息集成等。虽然早期的方法使用模式匹配从文本中自动提取结构化知识，但自然 语言处理（NLP）和统计学习，与通过手动构建可能实现的内容相比，生成的内容往往缺乏质量[481]。 然而，手动构建成本高昂，难以实现广泛覆盖并保持数据最新。 为了提取高覆盖率和高质量的数据，YAGO（如DBpedia）主要从维基百科信息框和类别页面中提取数据，其中分别包含特定类别的基本实体信息和文章列表； 这些又与来自 WordNet [482] 的分层概念统一起来。 一个模式——称为 YAGO 模型——提供了在 RDFS 中定义的词汇； 该模型允许将单词表示为实体，捕获同义词和歧义[481]。 该模型进一步支持具体化、n 元关系和数据类型 [482]。 YAGO 中采用的细化机制包括规范化，其中每个边和节点都映射到一个唯一标识符并删除重复元素，以及类型检查，其中消除了无法通过演绎或归纳方法分配给类的节点 [482]。 YAGO 将在以后的几年中扩展以支持时空上下文 [232] 和多语言维基百科 [313]。

**10.1.3 Freebase**

Freebase 是人类知识的一般集合，旨在解决与语义 Web 的分散性质相关的一些大规模信息集成问题，例如采用不均衡、实施挑战和分布式查询性能限制 [51]。 与 DBpedia 和 YAGO 不同——它们大多是从 Wikipedia/WordNet 中提取的——Freebase 直接从人类编辑那里征求意见。 Freebase 平台中包含一个具有版本控制机制的可扩展数据存储； 用于存储文本、图像和媒体文件的大型数据对象存储 (LOB)； 可以使用 Metaweb 查询语言 (MQL) 查询的 API； 网络用户界面； 和一个轻量级的打字系统[51]。 后一种打字系统旨在支持协作过程。 该系统不是强制本体正确性或逻辑一致性，而是作为一个松散的结构化机制集合来实现的——基于数据类型、语义类、属性、模式定义等——允许不兼容的类型和属性同时共存[51]。 内容可以通过 Web 用户界面以交互方式添加到 Freebase，或者通过利用 API 的写入功能以自动方式添加。 Freebase 已于 2010 年被 Google 收购，其中 Freebase 的内容构成了 2012 年宣布的 Google 知识图谱的重要组成部分 [459]。 当 Freebase 在 2015 年 3 月变为只读时，知识图谱包含超过 30 亿条边。 大部分内容随后被迁移到 Wikidata [382]。

**10.1.4 Wikidata**

正如 DBpedia 和 YAGO 所利用的那样，维基百科包含大量嵌入在信息框、列表、表格等中的半结构化数据。然而，这些数据传统上是在不同的文章和语言中手动管理和更新的； 例如，智利足球运动员的进球可能需要在数百种语言版本中手动更新球员文章、锦标赛文章、球队文章、最佳射手名单等。 人工管理导致了各种数据质量问题，包括不同文章、语言等中的矛盾数据。因此，维基媒体基金会提议将维基数据作为一个集中的、协作编辑的知识图谱，为维基百科和任意其他客户提供数据。 在这个愿景下，可以将一个事实添加到 Wikidata 一次，从而触发 Wikipedia 中跨不同语言的潜在大量文章的自动更新 [515]。与 Wikipedia 一样，Wikidata 也被认为是包含应引用主要来源的声明的二级来源，尽管声明也可以最初添加而无需参考 [390]。 Wikidata 进一步允许在潜在矛盾（引用）的声明方面有不同的观点 [515]。 Wikidata 是多语言的，其中节点和边被分配了与语言无关的 Qxx 和 Pxx 代码（参见图 33），随后与各种语言的标签、别名和描述相关联 [261]，从而允许以这些语言出现声明。 不仅允许在数据级别进行协作编辑，还允许在模式级别进行协作编辑，允许用户添加或修改轻量级语义公理 [391]——包括子类、子属性、逆属性等——以及形状 [56]。 Wikidata 提供各种访问协议 [315] 并已获得广泛采用，被 Wikipedia 用于在某些域中生成信息框 [431]，得到 Google [382] 的支持，并已被用作 Apple 等著名应用程序的数据源 Siri 等 [315]。

**10.1.5 其他开放的跨领域知识图谱**

多年来，已经开发了许多其他跨领域知识图谱。 BabelNet [351]——以与 YAGO 类似的方式——基于统一 WordNet 和 Wikipedia，但集成了其他知识图谱，例如 Wikidata，并专注于创建多语言词汇形式的知识图谱（组织成多语言同义词集） 通过将字典资源（如维基词典和 OmegaWiki）转换为知识图谱。 与其他知识图谱相比，BabelNet 等词汇化知识图谱将维基百科中的百科信息与单语和双语词典中常见的词典信息结合在一起。 Cyc 项目 [294] 旨在以机器可读的方式对常识知识进行编码，自 1986 年以来，超过 900 人年的努力 [319] 已经创建了 220 万个事实和规则。 尽管 Cyc 是专有的，但已经发布了一个名为 OpenCyc 的开放子集，我们参考了 Färber 等人的比较。 [148] 的 DBpedia、Freebase、OpenCyc 和 YAGO 以获取更多详细信息。 自 2010 年以来，永无止境的语言学习 (NELL) 项目 [335] 使用 OIE 方法从网页文本中提取了 1.2 亿条边的图形（参见第 6 节）。 每个这样的开放知识图谱都将本文中讨论的语言和技术的不同组合应用于不同的来源，并产生不同的结果。

**10.1.6 特定领域的开放知识图谱**

开放知识图谱已在各种特定领域发布。Schmachtenberg 等人。 [438] 确定关联数据上下文中最突出的领域如下：与新闻、电视、广播等相关的媒体（例如，BBC 世界服务档案 [406]）； 政府，与发布数据以促进透明度和发展有关（例如，由美国 [222] 和英国 [450] 政府）； 与各个学科的学术文献相关的出版物（例如，OpenCitations [385]、SciGraph [246]、Microsoft Academic Knowledge Graph [147]）； 地理的，与感兴趣的地方和区域有关（例如，LinkedGeoData [472]）； 生命科学，与蛋白质、基因、药物、疾病等有关（例如，Bio2RDF [79]）； 和用户生成的内容，与评论、开源项目等有关（例如 Revyu [217]）。 开放知识图谱也已在其他领域发表，包括文化遗产[245]、音乐[407]、法律[337]、神学[451]，甚至旅游[11、263、308、540]。 此类知识图谱的设想应用与它们产生的领域一样多种多样，但通常与集成 [79、407]、推荐 [308、407]、透明度 [222、450]、归档 [245、406]、 去中心化 [217]、多语言支持 [451]、法规遵从性 [337] 等。

**10.2 企业知识图**

许多公司已经宣布创建专有的“企业知识图谱”，并考虑到各种目标，其中包括：提高搜索能力 [83、205、280、457、459]、提供用户推荐 [83、205]、 实施会话/个人代理 [392]、增强定向广告 [214]、增强业务分析能力 [214]、连接用户 [214、365]、扩展多语言支持 [214]、促进研究和发现 [34]、评估和降低风险 [107, 495]，跟踪新闻事件 [326]，以及提高运输自动化 [223] 等（许多）。 正如 Noy 等人的讨论所反映的，这些企业知识图虽然高度多样化，但确实遵循了一些高级趋势。 [365]：（1）数据通常从各种外部和内部来源（通常涉及文本）集成到知识图谱中； (2) 企业知识图谱往往非常庞大，有数百万甚至数十亿的节点和边，在可扩展性方面提出挑战； (3) 初始知识图谱的细化——添加新链接、合并重复实体等——对于提高质量很重要； (4) 使知识图谱与领域保持同步的技术通常是至关重要的； (5) 为了从企业知识图谱中得出结论，本体和机器学习表示的混合经常被组合或用于不同的情况； (6) 使用的本体往往是轻量级的，通常是表示类或概念层次结构的简单分类法。

我们现在讨论部署企业知识图的主要行业。

**10.2.1 网络搜索**

Web 搜索引擎传统上专注于将查询字符串与 Web 文档中的子字符串进行匹配。 谷歌知识图 [365, 459] 提倡“事物不是字符串”的范式——类似于语义搜索 [199]——搜索引擎现在将尝试识别特定搜索可能表示感兴趣的实体。 知识图本身描述了这些实体以及它们如何相互关联。 谷歌知识图的主要面向用户的应用程序之一是“知识面板”，它在（一些）搜索结果的右侧显示一个窗格，描述搜索似乎正在寻找的主要实体，包括一些图像， 属性-值对，以及用户也搜索的相关实体列表。 谷歌知识图谱是普及“知识图谱”这一短语的现代用法的关键（见附录 A）。 其他主要搜索引擎，例如 Microsoft Bing36 [457]，随后将按照类似的思路发布知识图谱。

**10.2.2 Commerce**

主要关注销售或租赁商品和服务的公司也发布了企业知识图谱。 这种知识图谱的一个突出例子是亚马逊 [127, 280] 使用的图谱，它描述了在其在线市场上销售的产品。 该知识图谱的主要目标之一是为产品启用更高级的（语义）搜索功能，以及改进对其在线市场用户的产品推荐。 eBay [392] 宣布了另一个商业知识图，它对产品描述和购物行为模式进行编码，并用于为会话代理提供支持，帮助用户通过自然语言界面找到相关产品。 Airbnb [83] 还描述了一个知识图谱，它对租金、地点、事件、体验、社区、用户、标签等进行编码，在其之上定义了分类模式。 该知识图用于向潜在客户推荐特定房屋附近可用的景点、事件和活动。 优步 [205] 同样宣布了一个知识图谱，专注于他们的“优步外卖”送货服务的食品和餐馆。 目标再次为不确定他们正在寻找什么样的食物的用户提供语义搜索功能和推荐。

**10.2.3 社交网络**

企业知识图谱也出现在社交网络服务的背景下。Facebook [365] 聚集了一个知识图谱，不仅描述了关于用户的社交数据，还描述了他们感兴趣的实体，包括名人、地方、电影、音乐等，以便将人们联系起来，了解他们的兴趣，并提供建议。LinkedIn [214] 宣布了一个包含用户，工作，技能，公司，地点，学校等的知识图谱，在其上定义了分类模式。知识图谱用于提供重要概念的多语言翻译，改进有针对性的广告，为求职和人员搜索提供高级功能，同样也提供与人员匹配的工作建议 (反之亦然)。Pinterest [188] 创建了另一个知识图谱，描述了用户及其兴趣，后者被组织到分类法中。知识图谱的主要用例是帮助用户更轻松地找到他们感兴趣的内容，并通过有针对性的广告来增加收入。

**10.2.4 金融**

金融部门也看到了企业知识图谱的部署。 其中，Bloomberg [326] 提出了一个知识图谱，为金融数据分析提供动力，包括基于当前新闻报道和推文的公司情绪分析、问答服务以及检测可能影响股票价值的新兴事件。 Thompson Reuters (Refinitiv) [495] 同样宣布了一个知识图谱，使用分类模式来组织这些实体，对人员、组织、股权工具、行业分类、合资企业和联盟、供应链等的“金融生态系统”进行编码 . 他们提到的知识图的一些应用包括供应链监控、风险评估和投资研究。 Banca d'Italia [32] 在学术环境中也探索了知识图谱，使用基于规则的推理来确定，例如，不同利益相关者对公司的所有权百分比。 其他探索金融知识图谱的公司包括埃森哲 [368]、第一资本 [65]、富国银行 [355] 等。

**10.2.5 其它行业**

企业也一直在积极开发知识图谱，以在各种其他行业中实现新的应用，包括：医疗保健，IBM 正在探索药物发现的用例 [365] 和从包装插页中提取信息 [173]，而阿斯利康 [34] 正在使用知识图谱来推进基因组学研究和疾病理解； 运输，博世正在探索用于驾驶自动化的场景和位置的知识图[223]； 石油和天然气，其中 Maana [107] 正在使用知识图来执行数据集成，以减轻油井和钻井的风险；还有更多。