在这篇文章中，我们提供了一个知识图谱的全面介绍，在需要开发多样化、动态、大规模数据集的场景中，知识图谱最近获得了行业和学术界的重大关注。在总体介绍之后，我们激发和对比了各种基于图谱的数据模型和用于知识图谱的查询语言。我们讨论了模式、身份和上下文在知识图谱中的作用。我们解释如何使用演绎和归纳技术的组合来表示和提取知识。我们总结了知识图谱的创建、充实、质量评估、细化和发布的方法。我们提供了一个重要的开放知识图和企业知识图的概述，其中还包括它们的应用程序，以及它们如何使用上述技术。我们总结了知识图谱未来的研究方向。

1. 介绍

虽然“知识图谱”这个词至少从1972年开始就被用于文献,但这个短语的现代化身源于2012年谷歌知识图谱的宣布,其次是由Airbnb、亚马逊、eBay、 Facebook、LinkedIn、微软、IBM、Uber发表的关于知识图谱发展的进一步公告。工业界对这一概念的不断吸收证明学术界很难忽视:越来越多的关于知识图的科学文献正在出版，其中包括书籍，以及概述定义、新技术和知识图谱具体方面调查的论文。

所有这些发展背后的核心思想是使用图谱来表示数据，通常通过某种方式来更加清楚地表示知识。结果最常用于涉及大规模集成、管理和从不同数据源提取价值的应用程序场景。在这样的环境中，与关系模型或NoSQL替代品相比，使用基于图谱的知识抽象有很多好处。图谱为各种领域提供了一种简洁而直观的抽象，其中边缘捕获了社会数据、生物交互、书目引用和合著、交通网络等固有实体之间的(潜在周期性)关系。图谱允许维护者推迟模式的定义，允许数据及其范围以一种比关系模式中通常可能的更灵活的方式演进，特别是对于捕获不完整的知识。与(其他)NoSQL模型不同，专门的图谱查询语言不仅支持标准的关系操作符(连接、并、投影等)，而且还支持用于递归地查找通过任意长度路径连接的实体的导航操作符。标准的知识表示形式——比如本体和规则——可以用来定义和推理用于标记和描述图谱中的节点和边的术语的语义。图谱分析的可扩展框架可以用于计算中心性、集群、摘要等，以获得对所描述领域的洞察力。各种表示也被开发出来，支持直接在图上应用机器学习技术。

总之，构建和使用知识图谱的决定开辟了一系列技术，这些技术可以用于集成和从不同的数据源中提取价值。然而，我们还没有看到一个描述知识图谱如何被使用、它们使用了什么技术以及它们如何与现有的数据管理主题相关联的通用统一总结。

本篇指导性论文的目标是激励和给出一个知识图谱的全面介绍：描述其基础数据模型和如何查询；讨论与模式、身份和上下文相关的表示；讨论使知识清晰的演绎和归纳的方法；介绍各种技术，可用于创建和丰富使用图谱结构的数据；描述如何识别知识图谱的质量以及如何对其进行细化；讨论发布知识图谱的标准和最佳实践；并提供在实践中发现的现有知识图谱的概述。我们的目标受众包括新进入知识图谱领域的研究人员和实践者。因此，我们不假设读者对知识图有特定的专长。

*知识图谱*：“知识图谱”的定义仍然存在争议，出现了许多(有时是相互冲突的)定义，从具体的技术建议到更具包容性的一般性建议；我们在附录a中讨论了这些先前的定义。在此，我们采用了一个包含性的定义，其中，我们将知图谱视为一个数据图，旨在积累和传递现实世界的知识，其节点表示感兴趣的实体，其边表示这些实体之间的关系。数据的图(又称数据图)符合基于图的数据模型，可以是有向边标记图、属性图等(我们在第2节讨论具体的替代方案)。所谓知识，我们指的是已知的东西。这些知识可以从外部资源中积累，也可以从知识图谱本身提取。知识可以由简单的语句组成，如“圣地亚哥是智利的首都”，也可以由量化的语句组成，如“所有首都都是城市”。简单的语句可以在数据图谱中作为边累积。如果知识图想要积累量化的语句，那么就需要一种更有表现力的方式来表示知识——比如本体或规则。演绎方法可以用来继承和积累进一步的知识(例如，“圣地亚哥是一个城市”)。额外的知识——基于简单的或量化的语句——也可以使用归纳方法通过知识图谱提取和积累。

知识图谱通常由许多来源组合而成，因此，在结构和粒度方面可能非常不同。为了解决这种多样性，模式、身份和上下文的表示通常起到关键作用，其中模式为知识图谱定义了高级结构，身份表示图谱中的节点(或外部资源中的节点)指向相同的现实世界实体，而上下文可能表示一个特定的背景，其中某些知识单元是被证明正确的。如前所述，知识图谱需要有效的提取、丰富、质量评估和细化方法，才能随着时间的推移而成长和完善。

*实践*：知识图谱的目的是作为组织或社区中不断发展的共享知识的基础。在实践中，我们将知识图谱分为两种类型:开放知识图谱和企业知识图谱。开放的知识图谱在线上发布，使其内容为公众服务。最著名的例子——DBpedia，Freebase， Wikidata，YAGO等——涵盖许多领域，要么是从维基百科中提取的，要么是由志愿者社区建立的。开放知识图谱也已在特定领域发表，如媒体、政府、地理、旅游、生命科学等。企业知识图谱通常是公司内部的，并应用于商业用例。使用企业知识图谱的突出行业包括网络搜索(如Bing，谷歌)，商务(如Airbnb，亚马逊，eBay，Uber)，社交网络(如Facebook，LinkedIn)，金融(如埃森哲，意大利银行，彭博，Capital One，富国银行)，等等。应用包括搜索，推荐，个人代理，广告，商业分析，风险评估，自动化等。我们将在第10节中提供更多关于知识图谱在实践中的使用的细节。

*运行的例子*：为了保持讨论的易懂性，在整篇论文中，我们在一个与智利旅游业有关的假设知识图谱的背景下给出了具体的例子(灵感大致来自，[263,308])。该知识图谱由旅游委员会管理，旨在增加该国的旅游业，并促进战略地区的新景点。知识图谱本身最终将描述旅游景点、文化活动、服务和商业，以及城市和城市间的旅行路线。本组织设想的一些应用是：

* 创建一个旅游门户，允许游客搜索景点、即将发生的事件和其他相关服务(使用多种语言)；
* 了解旅游人口统计数据，包括季节、国籍等；
* 分析市民对各种景点及活动的意见，包括正面评论、对活动及服务的投诉摘要、罪案举报等；
* 了解旅游轨迹:游客参观景点、事件等的顺序；
* 参考现有航班/巴士的飞行轨迹，以建议新的策略性路线；
* 提供个性化的旅游景点推荐；

*结构*：论文其余部分的结构如下：

第二部分：概述图谱数据模型和可用于查询它们的语言。

第三部分：描述知识图谱中模式、标识和上下文的表示形式。

第四部分：提出演绎形式理论，通过该理论知识可以被表示和蕴涵。

第五部分：描述归纳技术，通过这些技术可以提取额外的知识。

第六部分：讨论从外部资源创建和丰富知识图谱。

第七部分：列举了可以评估知识图谱的质量维度。

第八部分：讨论知识图谱细化的各种技术。

第九部分：讨论了发布知识图谱的规则和协议。

第十部分：调查了一些重要的知识图谱及其应用。

第十一部分：最后对知识图谱的研究进行了总结和展望。

附录A：提供知识图谱的历史背景和以前的定义。

附录B：列举将从论文主体引用的正式定义。

11.摘要和结论

我们对知识图谱进行了全面的介绍，近年来它受到了越来越多的关注。在知识图谱旨在积累和传达现实世界知识的数据图的定义下，其节点代表感兴趣的实体，它们的边代表这些实体之间的关系，我们讨论了数据可以结构化为图的模型；模式、身份和上下文的表示；利用演绎和归纳知识的技术；知识图谱的创建、丰富、质量评估和改进的方法；发布知识图谱的规则和标准；最后，在现实世界中使用知识图谱。

知识图最终作为一个组织或社区中共同的知识基础，使知识的表示、积累、管理和传播成为可能。在这个角色中，知识图谱已经被广泛应用于各种各样的用例中，从商业应用程序——包括语义搜索、用户推荐、会话代理、目标广告、运输自动化等——到用于公共利益的开放知识图谱。在应用方面的一般趋势包括：（1）利用知识图谱大规模地整合和利用来自不同来源的数据（2）演绎形式主义（逻辑规则、本体论等）的组合，以及归纳技术（机器学习、分析等）来表示和积累知识（3）知识图的自下而上开发，通过间歇地或持续地应用提取、丰富和细化过程，以随着时间的推移更新、扩展或改进知识图谱。

因此，对知识图谱的研究应该（理想情况下）成为来自不同领域的技术的融合，其共同目标是知识最大化——从而提高价值——可以使用基于图的数据抽象来大规模地从不同的来源中提取出来。追求这一目标将受益于图形数据库、知识表示、逻辑、机器学习、图形算法和理论、本体工程、数据质量、自然语言处理、信息提取、隐私和安全等方面的专业知识。鉴于质量和采用程度不断提高的开放知识图谱的可用性，以及企业知识图谱在各个行业的迅速采用，未来对知识图谱的研究有可能促进社会各方面的关键进展。

致谢：...