知识图谱部分相关文献综述

1. 知识图谱定义相关

1.1 知识图谱定义与背景

知识图谱（Knowledge Graph）在维基百科中的定义为：知识图谱是Google用于增强其搜索引擎功能的知识库。互联网自从被发明至今，被认为经历了三个主要阶段：以“文档互联”作为其主要特征的Web1.0，以“数据互联”作为其主要特征的Web2.0，以及当下正在逐步进入的，以“知识互联”作为其主要特征的Web3.0。这里的“知识”与学生日常在学校中学习到的“知识”的定义并不相同，其着重点是能被计算机理解。当然，这种知识的定义并非是新兴的概念。早在2004年，为了给数量庞大甚至臃肿的互联网页添加语义，让计算机能够自动处理网页上的信息，作为这种概念前身的语义网就被提出，并产生了RDF与OWL等形式化本体语言。万维网联盟（W3C）将它们立为推荐标准，并呼吁在互联网中广泛使用这类包含隐藏语义的语言。RDF和OWL都是基于XML的形式化、结构化语言，其创造本义就不是为了让人，而是为了让计算机理解的。正是基于对语义网的研究，逐渐引出了知识图谱这一新的领域。

2012年，Google收购了原本致力于构建一个世界级规模的开放Web语义知识库的Metaweb公司后，正式提出了知识图谱的概念，并将知识图谱应用于其搜索引擎之中，以提供语义搜索的功能。从此，知识图谱被越来越多地应用于各个领域，国内也同样有搜狗的“搜立方”及百度的“知心”等搜索引擎建立于知识图谱技术之上。

1.2 知识图谱相关概念

一般而言，知识图谱的整体结构分为“模式层”与“数据层”两层。模式层为抽象的上层，是知识图谱的核心，主要包括“实体-关系-实体”与“实体-属性-属性值”。其中的“属性”也可以转化为“具有”这一关系，即将“实体-属性-属性值”转换为“实体-关系-实体”。数据层即为模式层的具象化，包含真实的数据，这些数据以上两种三元组形式进行存储，最终形成一张庞大的数据图，即为知识图谱。其中，“实体”是图中的节点，可以是诸如时间、地点、人名、事件等等，而“关系”则是图中连接两节点的边，也是模式层中定义的关系的一个实例。

知识图谱可以从不同角度进行分类。从覆盖范围角度看，知识图谱可以被分为通用类知识图谱与领域类知识图谱。通用类知识图谱更注重知识的广度，追求囊括更多更全面的知识。类似百度、谷歌这样的搜索引擎需要的正是此种类型的知识图谱。相对应的，领域类知识图谱则更聚焦于某一专业领域，如经济、医学等，力求细致准确地收纳这一专业的专业知识，以供相关软件与公司使用。从构建方式角度看，一部分知识图谱主要依靠专业人士手动构建。人工构建带来的是极高的准确度与完整性，但代价是规模一定偏小，因此更适合用来构建某一专业的领域类知识图谱。相对应的，另一部分知识图谱主要依靠人工智能的自动抽取，从互联网数十亿的网页中抽取数据，进行自动化的知识图谱构建。这样的方式带来的是极大的知识与数据量，并可以不断迭代更新，但同样会导致知识驳杂，质量层次不齐，准确度也无法保证。通用类知识图谱由于其规模所限，一般只能采用这样的方式构建。当然，以上的方法也可以相互融合，比如大的通用类知识图谱将小型的专业类知识图谱吸纳整合等。

1. 知识图谱构建相关

一般而言，知识图谱的构建是迭代性的过程，每一轮迭代中主要包含：知识抽取，知识融合，知识存储，知识加工等步骤，经过多次迭代才能形成一个规模较大，准确度较高的可用的知识图谱。

2.1 知识抽取

互联网上的知识主要可分为结构化数据、半结构化数据以及非结构化数据三种。

2.1.1 结构化数据

结构化数据主要指已存储在关系型数据库内的数据，数据都已经有明确的实体名以及实体之间的对应关系，是使用起来最方便的数据，因此是构建知识图谱的重要知识来源。结构化数据的知识抽取主要关注如何将关系型数据库的存储形式转化为知识图谱可用的数据类型。目前已经有一些比较成熟的用于将关系型数据库转化为RDF或OWL数据的方式。典型的rdb2rdf（关系型数据库数据转化为RDF数据）包括直接映射和R2RML。直接映射采用一种默认的，约定俗成的算法将数据表转化为RDF图，包括表、主键、外键等。而R2RML则是在其基础上给予用户自定义转化、映射方式的工具，更为灵活。总而言之，R2RML功能更为强大，包含了直接映射的功能。

2.1.2 半结构化数据

半结构化数据主要指百科类网站上的网页信息数据。这类网页结构有很高的一致性，因此抽取有一定的规律。同时，百科网页数据量大，数据质量有保障，因此对于构建知识图谱同样至关重要。

半结构化数据的抽取主要有三种方式：人工抽取，有监督的自动抽取与无监督的自动抽取。人工抽取顾名思义，依靠人工对网页结构进行分析，手动构建合适的数据抽取脚本来对某一百科网站进行数据抽取。这种方式好处在于准确性与效率极高，缺点在于可移植性差，消耗大量人力，同时如果网页内容有所更新，脚本也需要同步更新。有监督的自动抽取指先选取一定的已标注的训练集用于机器学习的训练，然后将训练完成的模型用于数据的抽取。这种抽取方式的效率和准确率很大程度上取决于用于训练的训练集选取的好坏。最后，无监督的自动抽取指先将网页聚类分组，然后依靠机器学习自动发掘其中数据的分布规律，并进行数据的抽取。这是对人力资源消耗最少的方法，但缺点是抽取来的数据可能有大量的噪声。

2.1.3 非结构化数据

非结构化数据指一般网页上的数据。虽然百科网页数据已经较为全面，但仍有大量的知识散布在普通的一般网页之中。这些网页结构多样，内容复杂繁多，因此是三种知识抽取中最具挑战性的。

简单来说，非结构化数据需要进行实体抽取、关系抽取与属性抽取等步骤，每一步都有多种方法，如基于规则的匹配，机器学习、深度学习等等。其中基于规则的匹配比较类似手工抽取，需要由专家进行规则与模板的编写，在小数据集上准确率极高，但同样具有可移植性、可扩展性差的问题，无法被广泛应用。机器学习等方法则更为自动化，但仍依赖训练集的选择决定最终模型的好坏。

2.2 知识融合

由于知识抽取来源的多元性与异构性，最终得到的结构化数据可能存在各种歧义问题。知识融合就是将多种来源的数据相互融合，组成质量高、覆盖广的知识图。

知识融合要解决的是不同来源数据间对同一实体称呼方式不同，导致最终的查询出现结果不符、结果缺失的现象。如“复旦”、“复旦大学”、“Fudan University”指代的可能都是相同的实体。同时，知识融合也需要解决实体之间的歧义问题，如“我的苹果坏了”与“我喜欢吃苹果”，前者指代的是苹果手机，而后者指代的是水果苹果。最后，知识融合还需要解决指代间的歧义问题，如人称代词“他”指代的究竟是上下文中的哪个具体实体人等。

2.3 知识存储

知识存储即将已完善过的知识进行存储的方式。主流的存储方式主要有：RDF数据存储与图存储。

2.3.1 RDF数据存储

RDF数据存储与传统关系型数据库比较类似，主要包括三元组表存储、垂直划分存储、水平划分存储及属性表存储等。

三元组表存储就是将每一个关系三元组的主语、谓语、宾语作为关系型数据表中的三列进行存储，其查询非常直接方便，缺点是在查询时会进行大量的连接操作，在数据量大时性能会受到很大影响。典型的使用三元组表存储方式的系统是3store。

垂直划分存储即建立与谓语数量相同的关系型数据表，并将所有主语与宾语作为表的两列进行存储。这种方式在谓语较多时需要维护大量的数据表，且数据表之间的相互连接也十分频繁。典型的使用垂直划分存储的系统是SW-Store。

水平划分存储即建立一整张关系型数据表，每行对应一个主语，每列对应一个谓语，表格内容就是该主语与谓语对应的宾语。这样的存储方式查询方便，但主要有两个问题。首先，表中会存在大量空值；其次，对于每一个确定的主语及谓语，至多只能有一个宾语存在，但在事实的知识图谱中并非如此。

属性表存储则在水平划分存储上有所优化，先将主语分类，然后将相同类型的主语置入同一张表中。这种方式的假设是同类主语的谓语也会比较相似，因此表中的空值数量会被大大减少。然而事实上在即便在相同类型的主语之间谓语依然有可能出现较大差异，因此空值问题并没有被很好地缓解，同时也没有解决一一对应问题。典型的使用属性表存储的系统是Jena。

2.3.2 图存储

每种不同的图数据库都为存储三元组设计了不同的，有针对性的底层存储结构。典型的图数据库有Neo4j。Neo4j的存储主要包括实体（节点）、关系（边）、属性和标签。图数据库是目前知识图谱构建主流选择的存储方式，因为其底层设计具有针对性，同时充分借鉴了现有的大规模数据库查询系统的经验，因此在性能上，尤其是大规模、复杂的数据查询性能上具有较大的优势。

1. 知识图谱应用相关

3.1 智能搜索

如前文所述，知识图谱最初被提出的目的就是为了强化Google的搜索引擎功能。传统的搜索引擎会将用户输入的关键词与其数据库中的网页进行匹配，并以关联度高低将结果返回给用户，而在知识图谱帮助下的语义搜索则可以在此基础上完成更复杂的功能，比如将用户搜索的实体“复旦大学”的相关信息（所在地、邮编、专业排名等等）一并返回。同时，知识图谱可以通过实体对齐的方式解析用户的真正意图，比如搜索“复旦大学校长”，搜索引擎并不简单地返回与该关键词相关的网页，而是将其对齐至实体“金力”，并直接返回其相关信息。同时，搜索引擎还能将历任复旦大学校长的信息作为补充一并返回。

3.2 智能问答

问答系统则可以看成搜索引擎的强化使用版本。为了回答用户的自然语言式问题，问答系统需要先利用NLP（自然语言处理）相关技术对问题进行解析，抽取出其中的实体与谓语，再在知识图谱中进行检索。简而言之，问答系统就是将知识图谱看作一个庞大的知识数据库。

在此基础上，用户可能会提出更多种更难的问题，如“多条问答”：“与复旦大学在同一个城市的大学中最大的是哪一所？”需要先解析出复旦大学所在城市上海，然后再通过上海的大学关联，比较占地面积，返回最终结果。“多轮问答”：“复旦大学在哪个城市？”“它的邮编是多少？”需要智能系统能够保留前一轮甚至前几轮使用到的实体，包括问题与答案中的实体，然后进行指代解析等等。

3.3 推荐系统

推荐系统目前也广泛被应用于各大电商平台。其主要内涵在于，先构建商品之间相互关联的知识图谱，然后跟踪用户的购买与使用软件习惯，有针对性地推荐类似的产品，以达到最好的推销效果。同时，推荐系统还会根据收集大量顾客的购买信息，找到与用户类似的其他用户的购买行为，进行相对应的推荐。这些也就是常见的“猜你喜欢”，“其他人还在搜”等功能。

1. 参考文献

徐增林,盛泳潘,贺丽荣,等. 知识图谱技术综述[J]. 电子科技大学学报,2016,45(4):589-606. DOI:10.3969/j.issn.1001-0548.2016.04.012.

黄恒琪,于娟,廖晓,等. 知识图谱研究综述[J]. 计算机系统应用,2019,28(6):1-12. DOI:10.15888/j.cnki.csa.006915.

于浏洋,郭志刚,陈刚,等. 面向知识图谱构建的知识抽取技术综述[J]. 信息工程大学学报,2020,21(2):227-235. DOI:10.3969/j.issn.1671-0673.2020.01.017.

葛唯益,王振宇,王羽,等. 主流知识图谱存储系统试验对比[J]. 指挥信息系统与技术,2019,10(5):28-33,75. DOI:10.15908/j.cnki.cist.2019.05.006.