【干货】最全知识图谱综述#2: 构建技术与典型应用

原创：Xu/Quan et al. 专知 2017-09-29

【导读】知识图谱技术是人工智能技术的组成部分，其强大的语义处理和互联组织能力，为智能化信息应用提供了基础。我们专知的技术基石之一正是知识图谱-构建AI知识体系-专知主题知识树简介。下面我们特别整理了关于知识图谱的技术全面综述，涵盖基本定义与架构、代表性知识图谱库、构建技术、开源库和典型应用。主要基于的参考文献来自[22]和[40], 本人(Quan)做了部分修整。

昨天我们介绍了《知识图谱的概念以及构建技术-知识提取、知识表示、知识融合》，今天介绍知识图谱的知识推理和典型应用。

知识图谱构建的关键技术

1 知识提取
2 知识表示
3 知识融合
4 知识推理

知识推理则是在已有的知识库基础上进一步挖掘隐含的知识，从而丰富、扩展知识库。在推理的过程中，往往需要关联规则的支持。由于实体、实体属性以及关系的多样性，人们很难穷举所有的推理规则，一些较为复杂的推理规则往往是手动总结的。对于推理规则的挖掘，主要还是依赖于实体以及关系间的丰富同现情况。知识推理的对象可以是实体、实体的属性、实体间的关系、本体库中概念的层次结构等。知识推理方法主要可分为基于逻辑的推理与基于图的推理两种类别。

1） 基于逻辑的推理

基于逻辑的推理基于逻辑的推理方式主要包括一阶谓词逻辑(first order logic)、描述逻辑(description logic)以及规则等。一阶谓词逻辑推理是以命题为基本进行推理，而命题又包含个体和谓词。逻辑中的个体对应知识库中的实体对象，具有客观独立性，可以是具体一个或泛指一类，例如奥巴马、选民等；谓词则描述了个体的性质或个体间的关系。文献[1]针对已有一阶谓词逻辑推理方法中存在的推理效率低下等问题，提出了一种基于谓词变迁系统的图形推理法，定义了描述谓词间与/或关系的谓词，通过谓词图表示变迁系统，实现了反向的推理目标。实验结果表明：该方法推理效率较高，性能优越。描述逻辑是在命题逻辑与一阶谓词逻辑上发展而来，目的是

在表示能力与推理复杂度之间追求一种平衡。基于描述逻辑的知识库主要包括Tbox(terminology box)与ABox(assertion box)[2]。通过TBox与ABox，可将关于知识库中复杂的实体关系推理转化为一致性的检验问题，从而简化并实现推理[3]。

通过本体的概念层次进行推理时，其中概念主要是通过OWL(Web ontology language)本体语义进行描述的。OWL文档可以表示为一个具有树形结构的状态空间，这样一些对接结点的推理算法就能够较好地应用起来，例如文献[4]提出了基于RDF和PD*语义的正向推理算法，该算法以RDF蕴涵规则为前提，结合了sesame算法以及PD*的语义，是一个典型的迭代算法，它主要考虑结点与推理规则的前提是否有匹配，由于该算法的触发条件导致推理的时间复杂度较高，文献[5]提出了ORBO算法，该算法从结点出发考虑，判断推理规则中第一条推理关系的前提是否满足，不仅节约了时间，还降低了算法的时间复杂度。

2）基于图的推理

在基于图的推理方法中，文献[6]提出的pathconstraintrandom walk，path ranking等算法较为典型，主要是利用了关系路径中的蕴涵信息，通过图中两个实体间的多步路径来预测它们之间的语义关系。即从源节点开始，在图上根据路径建模算法进行游走，如果能够到达目标节点，则推测源节点和目标节点间存在联系。关系路径的建模方法研究工作尚处于初期，其中在关系路径的可靠性计算、关系路径的语义组合操作等方面，仍有很多工作需进一步探索并完成。
除上述两种类别的知识推理方法外，部分研究人员将研究重点转向跨知识库的推理方法研究，例如文献[7]提出的基于组合描述逻辑的Tableau算法，该方法主要利用概念间的相似性对不同知识库。

知识图谱开源库

Apache Jena（或简称Jena）是一个用于构建语义Web和关联数据应用程序的自由和开源的Java框架。 该框架由不同的API组成，用于处理RDF数据。

Jena是一个用于Java语义Web应用程序的API（应用程序编程接口）。它不是一个程序或工具，如果这是你正在寻找，我建议或许TopBraid Composer作为一个好的选择。因此，Jena的主要用途是帮助您编写处理RDF和OWL文档和描述的Java代码。

更多详细内容参见官网Apache Jena, 具体应用后续参考

知识图谱构建的典型应用

知识图谱为互联网上海量、异构、动态的大数据表达、组织、管理以及利用提供了一种更为有效的方式，使得网络的智能化水平更高，更加接近于人类的认知思维。目前，知识图谱已在智能搜索、深度问答、社交网络以及一些垂直行业中有所应用，成为支撑这些应用发展的动力源泉。

1 智能搜索

基于知识图谱的智能搜索是一种基于长尾的搜索，搜索引擎以知识卡片的形式将搜索结果展现出来。用户的查询请求将经过查询式语义理解与知识检索两个阶段：1) 查询式语义理解。知识图谱对查询式的语义分析主要包括：① 对查询请求文本进行分词、词性标注以及纠错；② 描述归一化，使其与知识库中的相关知识进行匹配[8]；③ 语境分析。在不同的语境下，用户查询式中的对象会有所差别，因此知识图谱需要结合用户当时的情感，将用户此时需要的答案及时反馈给用户；④ 查询扩展。明确了用户的查询意图以及相关概念后，需要加入当前语境下的相关概念进行扩展。2) 知识检索。经过查询式分析后的标准查询语句进入知识库检索引擎，引擎会在知识库中检索相应的实体以及与其在类别、关系、相关性等方面匹配度较高的实体[9]。通过对知识库的深层挖掘与提炼后，引擎将给出具有重要性排序的完整知识体系。

智能搜索引擎主要以3种形式展现知识：1) 集成的语义数据。例如当用户搜索梵高，搜索引擎将以知识卡片的形式给出梵高的详细生平，并配合以图片等信息；2) 直接给出用户查询问题的答案。例如当用户搜索“姚明的身高是多少？”，搜索引擎的结果是“226 cm”；3) 根据用户的查询给出推荐列表[7]等。

国外的搜索引擎以谷歌的Google Search、微软的Bing Search[10]最为典型。谷歌的知识图谱相继融入了维基百科、CIA世界概览等公共资源以及从其他网站搜集、整理的大量语义数据[11]，微软的BingSearch[10]和Facebook[11]、Twitter[12]等大型社交服务站点达成了合作协议，在用户个性化内容的搜集、定制化方面具有显著的优势。

国内的主流搜索引擎公司，如百度、搜狗等在近两年来相继将知识图谱的相关研究从概念转向产品应用。搜狗的知立方[13]是国内搜索引擎行业的第一款知识图谱产品，它通过整合互联网上的碎片化语义信息，对用户的搜索进行逻辑推荐与计算，并将最核心的知识反馈给用户。百度将知识图谱命名为知心[14]，主要致力于构建一个庞大的通用型知识网络，以图文并茂的形式展现知识的方方面面。

2 深度问答

问答系统是信息检索系统的一种高级形式，能够以准确简洁的自然语言为用户提供问题的解答。之所以说问答是一种高级形式的检索，是因为在问答系统中同样有查询式理解与知识检索这两个重要的过程，并且与智能搜索中相应过程中的相关细节是完全一致的。多数问答系统更倾向于将给定的问题分解为多个小的问题，然后逐一去知识库中抽取匹配的答案，并自动检测其在时间与空间上的吻合度等，最后将答案进行合并，以直观的方式展现给用户。

目前，很多问答平台都引入了知识图谱，例如华盛顿大学的Paralex系统[15]和苹果的智能语音助

手Siri[16]，都能够为用户提供回答、介绍等服务；亚马逊收购的自然语言助手Evi[17]，它授权了Nuance的语音识别技术，采用True Knowledge引擎进行开发，也可提供类似Siri的服务。国内百度公司研发的小度机器人[18]，天津聚问网络技术服务中心开发的大型在线问答系统OASK[19]，专门为门户、企业、媒体、教育等各类网站提供良好的交互式问答解决方案。

3 社交网络

社交网 站 Facebook 于2013 年推出了GraphSearch[20]产品，其核心技术就是通过知识图谱将人、地点、事情等联系在一起，并以直观的方式支持精确的自然语言查询，例如输入查询式："我朋友喜欢的餐厅""住在纽约并且喜欢篮球和中国电影的朋友"等，知识图谱会帮助用户在庞大的社交网络中找到与自己最具相关性的人、照片、地点和兴趣等。Graph Search提供的上述服务贴近个人的生活，满足了用户发现知识以及寻找最具相关性的人的需求。

垂直行业应用

下面将以金融、医疗、电商行业为例，说明知识图谱在上述行业中的典型应用。

1 金融行业

在金融行业中，反欺诈是一个重要的环节。它的难点在于如何将不同税务子系统中的数据整合在一起。通过知识图谱，一方面有利于组织相关的知识碎片，通过深入的语义分析与推理，可对信息内容的一致性充分验证，从而识别或提前发现欺诈行为；另一方面，知识图谱本身就是一种基于图结构的关系网络，基于这种图结构能够帮助人们更有效地分析复杂税务关系中存在的潜在风险[21]。在精准营销方面，知识图谱可通过链接的多个数据源，形成对用户或用户群体的完整知识体系描述，从而更好地去认识、理解、分析用户或用户群体的行为。例如，金融公司的市场经理用知识图谱去分析待销售用户群体之间的关系，去发现他们的共同爱好，从而更有针对性地对这类用户人群制定营销策略[21]。

2 医疗行业

耶鲁大学拥有全球最大的神经科学数据库Senselab[22]，然而，脑科学研究还需要综合从微观分子层面一直到宏观行为层面的各个层次的知识。因此，耶鲁大学的脑计划研究人员将不同层次的，与脑研究相关的数据进行检索、比较、分析、整合、建模、仿真，绘制出了描述脑结构的神经网络图谱，从而解决了当前神经科学所面临的海量数据问题，从微观基因到宏观行为，从多个层次上加深了人类对大脑的理解，达到了"认识大脑、保护大脑、创造大脑"的目标。

3 电商行业

电商网站的主要目的之一就是通过对商品的文字描述、图片展示、相关信息罗列等可视化的知识展现，为消费者提供最满意的购物服务与体验。通过知识图谱，可以提升电商平台的技术性、易用性、交互性等影响用户体验的因素[23]。

阿里巴巴是应用知识图谱的代表电商网站之一，它旗下的一淘网不仅包含了淘宝数亿的商品，更建立了商品间关联的信息以及从互联网抽取的相关信息，通过整合所有信息，形成了阿里巴巴知识库和产品库，构建了它自身的知识图谱[24]。当用户输入关键词查看商品时，知识图谱会为用户提供此次购物方面最相关的信息，包括整合后分类罗列的商品结果、使用建议、搭配等[24]。

除此之外，另外一些垂直行业也需要引入知识图谱，如教育科研行业、图书馆、证券业、生物医疗以及需要进行大数据分析的一些行业[25]。这些行业对整合性和关联性的资源需求迫切，知识图谱可以为其提供更加精确规范的行业数据以及丰富的表达，帮助用户更加便捷地获取行业知识.

4 司法行业

知识图谱在司法领域的运用悄然兴起,它帮助从业人员快速地在线检索相关的法务内容，从而提高法院审判工作质量和效率[26]。

参考文献

1. 描述逻辑. 描述逻辑基础知识[EB/OL]. (2014-02-24). http://www.2cto.com/database/201402/280927.html
2. LEE T W, LEWICKI M S, GIROLAMI M, et al. Blind source separation of more sources than mixtures using overcomplete representation[J]. Signal Processing Letters, 1999, 6(4): 87-90.
3. Ian Dickinson. Imp lementation experience with the DIG 1.1specification[EB/OL]. (2004-05-10). http://www.hpl.hp. com/semweb/publications.html.
4. 龚资. 基于OWL描述的本体推理研究[D]. 长春: 吉林大学, 2007.
5. LIU Shao-yuan, HSU K H, KUO Li-jing. A semantic service match approach based on wordnet and SWRL rules[C]//Proc of the 10th IEEE Int Conf on E-Business Engineering. Piscataway, NJ: IEEE, 2013: 419-422.
6. LAO N, MITCHELL T, COHEN W W. Random walk inference and learning in a large scale knowledge base[C]//Proc of EMNLP. Stroudsburg, PA: ACL, 2011:529-539.
7. 蒋勋, 徐绪堪. 面向知识服务的知识库逻辑结构模型[J].图书与情报, 2013(6): 23-31.
8. 王志, 夏士雄, 牛强. 本体知识库的自然语言查询重写研究[J]. 微电子学与计算机, 2009, 26(8): 137-139.

9. BLANCO R, CAMBAZOGLU B B, MIKE P, et al. Entity recommendation in web search[C]//Pro of the 12th International Semantic Web conference(ISWC). Berlin: Springer-Verlag, 2013: 33-48.
10. BRACHMAN R J. What IS-A is and isn't: an analysis of taxonomic links in semantic networks[J]. Computer; (United States), 1983, 10(1): 5-13.
11. Facebook. Facebook[EB/OL]. [2014-02-04]. https://www.facebook.com/.
12. Twitter. Twitter[EB/OL]. [2016-05-08]. https://twitter.com/.
13. 百度百科.搜狗知立方[EB/OL]. [2016-05-07]. http://baike.baidu.com/link?url=_J_2r2xYz0qSTwlYxqPZ00ZZuYyiA_kkZAohtC5EhmIzOjSwywKheEThY2gdXdzxS
14. Baidu. Zhi xin[EB/OL]. [2016-06-08].
15. Fader. Paralex[EB/OL]. [2016-05-08]. http://knowitall.cs.washington.edu/paralex.
16. 百度百科. Siri[EB/OL]. [2016-05-02]. http://baike.baidu.com/subview/6573497/7996501.htm.
17. 百度百科. Evi[EB/OL]. [2016-03-18]. http://baike.baidu.com/view/7574050.htm.
18. 百度. 度秘[EB/OL]. (2015-09-13). http://xiaodu.baidu.com/.
19. 百度百科. OASK 问答系统[EB/OL]. [2016-03-27]. http://baike.baidu.com/view/8206827.htm.
20. 百度百科. Graph search[EB/OL]. [2016-01-22]. http://baike.baidu.com/view/9966007.htm.
21. 李文哲. 互联网金融, 如何用知识图谱识别欺诈行为[EB/OL].
22. Senselab. Center for medical informatics at yale university school of medicine yale university school of medicine [EB/OL]. [2016-01-08]. http://ycmi.med.yale.edu/.
23. 田玲, 马丽仪. 基于用户体验的网站信息服务水平综合评价研究[J]. 生态经济, 2013(10): 160-162.
24. 一淘网.知识图谱[EB/OL]. (2014-12-12). https://www.aliyun.com/zixun/aggregation/13323.html.
25. 李涓子. 知识图谱：大数据语义链接的基石[EB/OL].(2015-02-20). http://www.cipsc.org.cn/kg2/.
26. 知识图谱技术在司法领域的应用：国双科技的探索与技术分享。http://mp.weixin.qq.com/s/aVEBf_VxkXpmx3Z3xUBtm

-END-

访问使用方法>>点击文章下方“**阅读原文**”访问专知网站。

专·知

微信号：Quan_Zhuanzhi

一站式AI知识服务
基于知识图谱的内容分发

长按识别二维码，关注了解更多。

阅读原文