《智能信息处理》课程考试

基于知识图谱的链接预测算法研究

方婉青

作业	分数[20]
得分	

2021 年 12 月 19 日

基于知识图谱的链接预测算法研究

方婉青[1)]

[1)](大连海事大学　信息科学与技术学院　大连　116026)

摘　要　随着互联网技术的发展，知识图谱技术为业界和学术界提供了一种更好的组织、管理和理解互联网中海量数据的有效方案。由于知识图谱的知识不完备，即图谱中存在缺失的实体或链接，导致知识图谱的使用存在巨大的限制，大大限制了知识图谱在用于检索和推理的准确性。补全知识图谱，完成知识图谱链接预测任务的研究成为知识图谱的核心任务之一。本文针对大规模开放领域知识图谱的链接预测问题展开研究。

关键词　知识图谱；链接预测；随机游走；

Research of Link Prediction based on Knowledge Graph

Wanqing Fang[1)]

[1)](Department of Information Science Technology, Dalian Maritime University, Dalian　116026)

Abstract　With the development of Internet technology, knowledge graph technology provides an effective solution for industry and academia to better organize, manage and understand the massive amounts of data in the Internet. Because the knowledge of the knowledge graph is incomplete, that is, there are missing entities or links in the graph, there are huge restrictions on the use of the knowledge graph, which greatly limits the accuracy of the knowledge graph for retrieval and reasoning. Completing the knowledge graph and completing the research of the knowledge graph link prediction task has become one of the core tasks of the knowledge graph. This paper examines the link prediction problem of large-scale open domain knowledge graphs.

Keywords　Knowledge Graph；Link Prediction；Random Walk；

1 绪论

近年来，随着信息技术的不断发展，人们的生活发生了巨大的改变。伴随着信息互联，用户信息、交互信息等各种各样的数据也随之产生。如何使用这些信息数据并进行知识挖掘与推理成为近年来人工智能领域的研究热点。为了实现更加智能快捷的搜索引擎，谷歌公司提出了知识图谱的概念。知识图谱技术的诞生为更好的组织、管理和理解互联网中的海量数据提供了一种非常有效的方案。因此，基于知识图谱的各种研究也成为了当前人工智能领域一大热点，并且受到了学术界和业界的广泛关注，吸引了一批又一批的研究人员投入到知识图谱相关领域的研究中。

知识图谱的链接预测任务或者说基于知识图谱的推理技术是知识图谱相关技术的重点研究问题。近年来，国内外各相关组织都在该研究方向投入巨大的时间与精力。在2020年的AAAI上有关图研究的论文接近总数的十分之一。由此可见，知识图谱链接预测技术的相关研究具有极大的研究价值，以及广泛的应用方向。

2 知识图谱

2012 年,谷歌正式提出知识图谱的概念。知识图谱(Knowledge Graph, KG)是一种具有有向图结构的知识库，其中，图的结点表示实体，边表示实体之间的语义关系，其基

本组成单位是“实体一关系一实体”三元组"。知识图谱具备良好的语义关联性及可解释性，便于实现逻辑推理及知识挖掘。2013年起，知识图谱开始在学术界和业界普及，并在语义搜索、智能问答、情报分析等场景中发挥重要作用。

知识图谱来源于各类结构化、半结构化或非结构化数据，其构建过程包括知识抽取、知识融合等核心步骤。然而，在构建过程中，由于数据来源不全面、知识抽取不准确等原因，所构建的图谱往往不具备良好的覆盖性。例如，在目前最大的开源知识库 Freebase 中，71%的人物缺少出生地信息，75%的人物缺少国籍信息，而其它不常见的关系或属性的覆盖率则更低 3。为了使知识图谱更加全面和完善，必须在现有的知识图谱的基础上，进一步挖掘缺失或隐含的知识，以实现知识图谱补全。

3 知识图谱链接预测

随着自然语言处理技术的飞速发展，知识抽取技术，即从大规模半结构化,非结构化文本数据中抽取实体和关系的技术，已经取得了长足的进步。然而知识抽取技术的准确率和覆盖率仍然有限，加之有限的数据来源，导致利用抽取出的知识构建的知识图谱不全面、不完善，由此引入了知识图谱补全的问题。

链接预测是实现知识图谱补全的主要手段。对于一个三元组(h,r,t)而言,本文讨论的链接预测是指给定头实体 h 和关系 r 预测尾实体 t，或者给定关系 r 和尾实体 t，预测头实体 h。通过链接预测技术,可以推理出实体之间的新的关联，即新的三元组（或称事实)，新事实通过质量评估后加入到知识图谱中，使得知识图谱更为丰富和完善。本文研究大规模开放领域知识图谱的链接预测算法，后续的知识质量评估不在本文研究范围内。

目前，知识图谱链接预测其主要实现方法可分为两种，基于符号的方法和基于统计的方法。两类推理模型具有各自的应用场景，并且具有互补性。基于符号的推理方法更多考虑确定性知识的推理，类似于专家系统，通过给定的逻辑规则设计推理机，对缺失的实体或关系进行推理。基于统计的推理是一种不确定性推理，通过统计规律对知识图谱中的缺失的边进行补全。实体关系学习方法是其中最常见的方法，也是最近几年知识图谱的一个比较热的研究方向，包括基于表示学习的方法和基于图特征的方法两种方式。

3.1 基于转移的链接预测算法

基于转移的链接预测算法是一种基于知识图谱表示学习的方法，具有简单、有效且计算效率高的优势，是本文的研究基础。基于转移的方法的核心思想为转移假设。依据转移假设，实体之间的关系可以用实体之间的转移来表示。在 TransE 算法采用的基本转移假设中，所有实体向量存在于一个共同的低维向量空间，而 TransH、TransR 等算法采用的复杂转移假设则将实体投射到关系特定的超平面或关系特定的空间中,计算投射后的实体之间的转移。

3.2 基于图特征的链接预测算法

基于图特征的链接预测算法利用知识图谱的图形结构特征来实现链接预测。该类算法具有简单直观、可解释性强等优点，但是其计算效率较低，应用于大规模知识图谱时的时间代价较高。因此,本文在将其融合到基于转移的方法中之前，首先研究图特征算法的效率问题。

图特征算法可以用于学习三元组之间的路径以及路径对于链接预测的重要性（或权重)。最具代表性的图特征算法是路径排序算法 PRA。PRA 通过在知识图谱上的随机游走发现路径特征，并通过训练逻辑回归模型实现链接预测。PRA 算法包括三个核心步骤，分别是:路径发现、特征计算和模型训练。

路径发现：每一条路径相当于一个“专家”，可以用于判断一个三元组成立的可能性。然而，通过随机游走发现的所有路径中不可避免地存在无效路径或噪声路径,这些路径无法进行有效的推理甚至可能误导推理。因此，PRA 算法通过路径有效性度量指标来进行路径特征选择。常用的路径有效性

度量包括 Hit 和 Precision。路径的 Hit 值表示通过该路径从查询的种子结点到达正确尾实体的查询的数目。路径的 Precision 表示通过该路径从查询的种子结点到达正确尾实体的平均游走概率。it 值和 Precision 值大于阈值的路径被保留。

特征计算：特征计算过程采用与路径发现阶段类似的随机游走算法。对于一个查询，通过上述公式计算从查询的种子节点 s 经由路径P_i到达查询的尾实体结点 e 的随机游走概率$h_{s,P_i}(e)$。$h_{s,P_i}(e)$即为实体对(s,e)对应路径 P 的路径特征值。其中，P 是通过路径发现得到的任一路径。假如给定一个路径集合P_1, P_2,..., P_n,，每一条路径均可得到一个路径特征值,所有路径特征构成三元组的路径特征向量。PRA 算法通过路径特征的线性组合来对三元组进行打分排序。

模型训练：给定一个关系 r 和一个三元组集合，若集合中的每一个三元组均为已标注数据，即标注了三元组是否成立，那么可以构造一个训练数据集。基于逻辑回归思想，定义目标函数。模型训练过程中通过随机梯度下降法等优化算法最大化目标函数,学习路径的权重。

4 知识图谱链接预测的评价指标

知识图谱链接预测常用的几个评价指标包括平均排名(MeanRank)、HITS@K 和平均倒数排名(Mean Reciprocal Rank, MRR)。链接预测算法的效率则通过算法运行时间来评价。

4.1 准确率评价指标

(1)Mean Rank

假如根据头实体和关系预测尾实体,实体集合中的所有实体构成候选实体集。对于测试集中的任一三元组，首先计算给定头实体、关系和任一候选实体构成的三元组的得分，得分越高，则该候选实体与给定头实体和关系匹配成正确三元组的可能性越大。然后将所有候选实体根据得分进行倒序排序，并根据测试集中的正确答案得到正确答案在所有候选实体中的排名。最后,对所有三元组的正确答案的排名求平均，得到平均排名，记为 Mean Rank。Mean Rank 值越小则正确答案的平均排名越靠前,预测准确率越高。

(2) Hits@K

假设根据头实体和关系预测尾实体,实体集合中的所有实体构成候选实体集。对于测试集中的任一三元组，首先计算给定头实体、关系和任一候选实体构成的三元组的得分，然后将所有候选实体根据得分进行倒序排序,并根据测试集中的正确答案得到正确答案在所有候选实体中的排名，若排名在前 K 则 Hit 值加一。最后，用 Hit 值除以测试集中所有三元组的个数，得到测试集中预测结果排名在前 K 的比例，记为 Hits@K。Hits@K 值越高则结果排名在前 K 的三元组越多，预测准确率越高。K 值通常取 1、5 或 10。与 Mean Rank 相同，Hits@K 也有 Raw 和 Filter 两种计算方式。根据关系和尾实体预测头实体时的 Hits@K 计算同上。

(3) Mean Reciprocal Rank (MRR)

假设给定头实体和关系（此处称为一个查询）预测尾实体，对于 PRA 算法推理出的候选答案集中的每一个实体,根据路径特征值与路径权重计算该实体与给定头实体和关系构成的三元组的得分，并根据得分进行倒序排序。而后根据测试集中该查询的正确答案集，得到第一个正确答案的排名并求倒数，记为 Reciprocal Rank。最后，对所有查询的 Reciprocal Rank 求平均，得到 MRR。MRR 越大，则正确答案的排名越靠前，链接预测越准确。

4.2 效率评价指标

算法运行时间（The Execution Time of the Algorithms）是指算法从开始运行到运行结束所花费的时间。本文采用算法运行时间（记为 Runtime）来衡量算法运行的效率。算法运行时间简单且直观，在同一种编程语言、同一种数据结构的前提下，使用算法运行时间衡量算法效率具有一定的合理性。然而，算法运行时间受到硬件性能、程序运行环境等因素的影响，因此，本文除算法运行时间外,也进行了算法时间复杂度的分析。

5 结语

本文详细介绍了与知识图谱链接预测相关的概念及技术,包括知识图谱、知识图谱链接预测等概念，基于转移的链接预测、基于图特征的链接预测等，以及如何对链接预测的准确率和效率进行评价，在此基础上可以进行一系列后续工作。

参 考 文 献

[1]唐宏,范森,唐帆,朱龙娇.融合知识图谱与注意力机制的推荐算法[J/OL].计算机工程与应用:1-13[2021-12-16].http://kns.cnki.net/kcms/detail/11.2127.tp.20211123.1655.002.html.

[2] 张厚源. 基于知识图谱的实体间链接预测方法研究[D]. 电子科技大学,2021.DOI:10.27005/d.cnki.gdzku.2021.001919.

[3] 刘科孟. 基于知识图谱的推荐系统的研究与实现[D]. 北京邮电大学,2021.DOI:10.26969/d.cnki.gbydu.2021.001835.

[4] 王建政. 知识图谱构建的方法研究与应用[D]. 电子科技大学,2021.DOI:10.27005/d.cnki.gdzku.2021.004528.

[5]黄志成.关于主动学习下的知识图谱补全研究[J].大众标准化,2021(12):63-65.