电 子 科 技 大 学

专业学位研究生学位论文开题报告表

攻读学位级别： □博士 硕士

培养方式： 全日制 □非全日制

专业学位类别及领域： 电子信息

学 院： 信息与软件工程学院

学 号：

姓 名：

论文题目：面向知识图谱的补全算法研究与实现

校内指导教师： 罗光春

校外指导教师： 刘楚雄

填表日期： 2021 年 12 月 27 日

电子科技大学研究生院

1. 学位论文研究内容

|  |  |  |
| --- | --- | --- |
| 课题类型 | | 应用基础研究 □应用研究 |
| 课题来源 | | □纵向 □横向 自拟 |
| 学  位  论  文  研  究  内  容 | 学位论文的研究目标、研究内容及拟解决的关键性问题（可续页）  **1.研究目标**  知识图谱通常是事实三元组（头实体、关系实体、尾实体）的集合，以结构化的方式表示人类知识。即使是最先进的知识图谱也是不完整的，图谱补全是一种根据知识图谱中已存在的实体去补全缺失事实的任务，不仅需要判断两个实体之间是否存在关系，而且还需要确定这是哪一种关系，是一种有前途的、广泛研究的、旨在解决知识图谱不完整的任务。  知识表示学习将知识库中的实体和关系表示为低维稠密的实体向量，学到的知识表示可以实现知识图谱补全任务，又被称为链接预测，可以构建大规模知识图谱，不断补充实体间的关系。从知识图谱表示模型的效果评价来看，知识补全成为衡量知识图谱表示模型效果的最广泛使用的任务。  本论文拟面向知识图谱表示学习方法实现图谱补全任务，利用已有三元组事实对节点关系联合学习，获得泛化模型，对预测的三元组进行评估，完成图谱补全，进一步地扩展到其他大型知识图谱上。  **2.研究内容**  **（1）节点和关系的表示学习**  实际生活中的图往往有大量的一对多，多对一这样的复杂结构，是一种更普遍和流行的图形式，而且每条边都有一个标签和与之相关的方向。本文拟结合卷积网络模型，提取节点与有向关系的特征，用于邻居节点的信息聚合，实现中心节点和关系的表示学习，为知识图谱补全奠定基础。  **（2）知识图谱补全**  在当前的知识补全研究中，较常用的做法是设计一个得分函数，用来计算三元组的可靠分数（或不可靠分数），在此基础上衡量三元组的可靠（或不可靠）程度，完成补充实体间关系的任务，实现图谱补全。通常而言，正确三元组比错误三元组获得更高的可靠分数，或者更低的不可靠分数。通过利用知识图谱中已存在的三元组拟合得分函数的方式学习实体和关系的表示，能够较好地获取实体和关系的潜在语义表示。因此，拟采用多尺度卷积，以完整三元组作为输入，评估可信度，构建知识图谱补全评估模型。  **3.拟解决的关键性问题**  **（1）有向关系缺少表示学习**  传统的基于频谱的图卷积网络（GCN）因为要求归一化图拉普拉斯算子为实对称正半定矩阵，以便于进行图的傅立叶变换，这也表明邻接矩阵必须是对称的，并且二维的邻接矩阵也将边限制为相同类型，即单一关系。但是这类方法无法表示学习有向图的有向关系，仅限于学习节点的表示形式。  **（2）三元组整体语义特征被忽略**  在实现知识图谱补全任务时，对预测的三元组可靠性评估需要考虑整个三元组内部的语义特征。现有研究将三元组进行分割，利用头实体和关系实体之间的特征提取和尾实体进行比对，忽略了头实体-关系实体-尾实体三者整体的语义特征联系，最终影响补全效果。 | |

1. 学位论文研究依据

|  |
| --- |
| 学位论文的选题依据和研究意义，国内外研究现状和发展态势；选题在理论研究或实际应用方面的意义和价值；主要参考文献，以及已有的工作积累和研究成果。（2000字）  **1.选题依据和研究意义**  知识图谱通常指存储了大量三元组的知识库，利用三元组（头实体，关系实体，尾实体）表示实体之间的关系。例如，（中国，首都，北京）表示中国的首都为北京。应用比较广泛的知识图谱包括Freebase[1]、DBpedia[2]、Wikidata[3]和YAGO[4]等。由于知识图谱中包含大量的人类知识，除了应用在搜索引擎任务，知识图谱已经成为众多领域的重要资源，如自动问答[5]、机器翻译[6]、词义消歧[7]和推荐系统[8]等任务。如何将知识图谱中的实体和关系等对象转化为向量形式，是知识图谱应用到各种机器学习等相关任务的重要基础。以独热表示（One-Hot）为例，为了区分不同对象，独热表示设置了一个对象总数大小的向量，其向量某一维度为1，其余位置为0。该方法虽然可以对不同对象加以区分，然而在该向量空间中，任意两个对象向量正交，无法体现表示对象之间的关系，且存在维度灾难、数据稀疏等问题[9]，因而在应用中存在较大局限。表示学习的目标是通过机器学习等方法将研究对象的语义信息表示为稠密低维实值向量[10]。与独热表示相比，表示学习方法学到的对象表示维度较低，可以快速计算实体间的语义相似度，这对于自然语言处理[11]和信息检索的很多任务有重要意义。  通过知识表示学习方法可以学到知识图谱中已有三元组的向量表示，然后用学到的表示进行图谱补全任务。知识图谱补全任务又称为知识图谱链接预测，一般是指在给定实体和关系的前提下预测出对应实体[12]，补全现有图谱缺失的关系。尽管知识图谱已经广泛地应用在许多实际任务中，然而由于人类知识的有限性和多变性,知识图谱仍然不完整。根据统计,即使是目前规模最大的知识图谱也存在不完整的问题。例如，广泛使用的Wikipedia中超过70%的人物对象没有种族信息[13]。这种规模庞大的知识图谱存在数量众多的实体、关系以及三元组信息，依靠人工修正的方式对知识图谱中的缺失信息进行补全过于费时，且耗费高昂。因此,如何采用自动化的方式补全缺失信息，近年来备受关注。由于知识图谱表示模型获得的实体和关系的向量表示是根据知识图谱中已存三元组得到，目前进行知识图谱补全任务主要是使用知识图谱表示模型。从知识图谱表示模型的效果评价[14-16]来看，图谱补全成为衡量知识图谱表示模型效果的最广泛使用的任务。  **2.国内外研究现状和发展态势**  近年来，研究者对知识图谱表示模型进行深入研究，提出学习实体和关系的分布式表示。在当前研究中，较常用的做法是设计一个得分函数，用来计算三元组的可靠分数（或不可靠分数），在此基础上衡量三元组的可靠（或不可靠）程度，完成图谱补全。通常而言，正确三元组比错误三元组获得更高的可靠分数，或者更低的不可靠分数，通过利用知识图谱中已存的三元组拟合得分函数的方式学习实体和关系的表示，能够较好地获取实体和关系的潜在语义表示。由于知识图谱中关系结构的复杂性（如在知识图谱中广泛存在的N-1、1-N、N-N等关系），研究者主要通过对得分函数进行改进设计以拟合知识图谱中复杂的三元组结构，因此得分函数是区分不同模型的主要依据。除此之外，负采样方法、训练方法、加入外部信息等因素也会对表示效果产生影响[17]。  知识图谱补全方法主要包含以下两部分的关键技术：其一是对知识图谱中的实体、关系等元素在向量空间构建它们的特征表示，这是知识图谱补全开展的数据基础，第二是确定得分函数来获取预测知识为真的概率，根据得分函数设计的不同，相关方法中包含基于距离的模型、基于语义匹配的模型和基于编解码结构评分的模型。  根据模型的内在结构以及得分函数对三元组真假的度量形式，将知识图谱补全方法主要分为基于翻译的方法、基于语义匹配的方法和基于神经网络的方法等。这里主要介绍基于神经网络的知识图谱补全模型。  基于神经网络的知识图谱补全模型在处理实体与关系之间的推理上也表现出了很好的效果，充分利用了神经网络对非线性复杂关系的建模能力，能够深入学习图谱结构特征和语义特征，实现对图谱缺失关系的有效预测。  基于翻译的模型多为浅层快速模型，缺乏对知识图谱深层语义特征的挖掘，很难适用于具有复杂图结构的大规模知识图谱。Dettmers等人提出的ConvE[18]是一个多层卷积模型，该模型以头实体和关系向量组成的矩阵作为输入，对该矩阵进行二维卷积操作再与尾实体做内积得到该三元组的分数。该方法得益于卷积网络的特性能够提取到丰富的特征，参数利用率高，但是只关注了嵌入向量不同维度的局部特征，忽略了三元组内部相同维度下的全局特征。为了改进这一问题，Nguyen等人提出ConvKB[19]模型，该模型以头实体、关系和尾实体向量组成的三列矩阵作为输入，再对该矩阵进行二维卷积操作，这样做不仅能够捕捉全局相同维度的关系特征，而且保持了三元组的平移特性。随后Vu等人结合胶囊网络提出CapsE[20]模型，目的是利用胶囊网络的特性对特征图重构，从而对三元组同一维度的特征进行深入建模，以获取三元组内部隐藏的语义信息。  2017年，Kipf等人提出了图卷积网络GCN[21]，首次将卷积操作用于图结构数据，该模型以结点的特征矩阵和邻接矩阵作为输入，在更新结点的向量表示时聚合了周围邻居结点的信息，但是该方法为每个邻居赋予了相同的权重，因此难以区分出邻居的重要性。Schlichtkrull等人首次提出用于建模关系数据的R-GCN[22]模型，可视为GCN模型在大规模关系数据的一种扩展。Velickovic等人提出图注意力网络GATs[23]，将注意力机制[24]引入图卷积网络，在注意力机制的指导下，不同的邻居节点被赋予不同的权重，通过堆叠注意力层使得中心结点能够根据权重有重点地聚合邻居结点的特征，丰富结点的向量表示。为了将GATs网络引申到知识图谱领域，Nathani[25]等人提出一种泛化后的图注意力网络模型用以知识表示，该方法以三元组的嵌入向量作为输入，强调了边的重要性，在注意力机制的指导下，不仅能够充分考虑到给定实体的邻居信息，且保持了知识图谱三元组的翻译特性。Wang等人[26]进一步提出图衰减注意力网络，为不同的关系路径分配不同的权重，并且考虑到关系的多样性，对每个关系向分别进行表示。  综上所述，实现知识图谱补全有多种模型，并且存在明显差异。得分函数、负采样和优化方法等模型结构，以及训练数据比例等因素对知识图谱补全任务上的效果存在显著影响。知识图谱补全在整个知识图谱理论与技术框架中占据着十分重要的地位，是知识图谱研究的一大重点和难点，在实际工程中也有非常广泛的应用场景。  **参考文献**  [1] Bollacker K, Cook R, Tufts P. Freebase: A Shared Database of Structured General Human Knowledge[C]//Proceedings of the 22nd AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2007: 1962-1963.  [2] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia A Crystallization Point for the Web of Data[J]. Journal of Web Semantics, 2009, 7(3): 154-165.  [3] WMF. Wikidata [EB/OL]. [2019-11-11]. https://www.wikidata.org/wiki/Wikidata:Main\_ Page.  [4] Suchanek F M, Kasneci G, Weikum G. YAGO: A Large Ontology from Wikipedia and WordNet[J]. Journal of Web Semantics, 2008, 6(3): 203-217.  [5] Zhang Y Z, Liu K, He S Z, et al. Question Answering over Knowledge Base with Neural Attention Combining Global Knowledge Information[OL]. arXiv Preprint, arXiv: 1606.00979.  [6] Yang B S, Mitchell T. Leveraging Knowledge Bases in LSTMs for Improving Machine Reading[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1436-1446.  [7] Almousa M，Benlamri R， Khoury R. A Novel Word Sense Disambiguation Approach Using WordNet Knowledge Graph [OL]. arXiv Preprint, arXiv: 2101.02875.  [8] 阮小芸,廖健斌,李祥,等.基于人才知识图谱推理的强化学习可解释推荐研究[J].数据分析与知识发现，2021,5(6): 36-50.  [9] Bellman R E. Dynamic Programming[M]. Dover Publications, Incorporated, 2003.  [10] 刘知远,孙茂松,林衍凯,等.知识表示学习研究进展[J].计算机研究与发展，2016, 53(2): 247-261.  [11] 余传明,王曼怡,林虹君,等.基于深度学习的词汇表示模型对比研究[J].数据分析与知识发现, 2020, 4(8): 28-40.  [12] 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.  [13] Ren F L, Li J C, Zhang H H, et al. Knowledge Graph Embedding with Atrous Convolution and Residual Learning [C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: ACM, 2020: 1532-1543.  [14] Li C, Peng X T, Zhang S H, et al. Modeling Relation Paths for Knowledge Base Completion via Joint Adversarial Training[J]. Knowledge-Based Systems, 2020, 201-202:105865.  [15] He L R, Liu B, Li G X, et al. Knowledge Base Completion by Variational Bayesian Neural Tensor Decomposition[J]. Cognitive Computation, 2018, 10(6):1075-1084.  [16] Wang H B, Jiang S C, Yu Z T. Modeling of Complex Internal Logic for Knowledge Base Completion[J]. Applied Intelligence, 2020, 50(10): 3336-3349.  [17] Ruffinelli D, Broscheit S, Gemulla R. You Can Teach an Old Dog New Tricks! On Training Knowledge Graph Embeddings[C]//Proceedings of the International Conference on Learning Representations. Addis Ababa: ICLR, 2020.  [18] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2d Knowledge Graph Embeddings[C]//Proceedings of the AAAI Conference on Artificial . Intelligence. 2018, 32(1): 1811-1818.  [19] Nguyen T D, Nguyen D Q, Phung D. A Novel Embedding Model for Knowledge Base Completion based on Convolutional Neural Network[C]//Proceedings of the Human Language Technologies Conference of the NAACL. 2018, 2: 327-333. DOI:10. 18653/v1/N18-2053.  [20] NGUYEN D Q, VU T, NGUYENT D, et al. A Capsule Network-based Embedding Model for Knowledge Graph Completion and Search Personalization[C]//Proceedings of the Human Language Technologies Conference of the NAACL.2019, 1:2180-2189.  [21] KIPF T N, WELLING M. Semi-supervised Classification with Graph Convolutional Networks [EB/OL]. 2016, arXiv preprint: 1609.02907.  [22] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling Relational Data with Graph Convolutional Networks[C]//European Semantic Web Conference. 2018:593-607. DOI:10.1007/978-3-319-93417-438.  [23] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph Attention Networks[EB/OL]. 2017, arXiv preprint:1710. 10903.  [24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need[C]//Proceedings of the NIPS Conference.2017: 6000-6010.  [25] NATHANI D, CHAUHAN J, SHARMA C，et al. Learning Attention-based Embeddings for Relation Prediction in Knowledge Graphs[C]//Proceedings of the ACL Conference.2019: 4710-4723.  [26] WANG R, LI B, HU S, et al. Knowledge Graph Embedding via Graph Attenuated Attention Networks[J]. IEEE Access,2020,8:5212-5224. DOI:10.1109/ACCESS.2019.2963367. |

1. 学位论文研究计划及预期目标

|  |
| --- |
| 1.拟采取的主要理论、研究方法、技术路线和实施方案（可续页）  **1.主要理论**  （1）知识表示学习  知识表示学习得到的分布式表示，可以快速计算实体间的语义相似度，这对于自然语言处理和信息检索的很多任务有重要意义；还可以补全已有知识图谱，构建大规模知识图谱，需要不断补充实体间的关系，利用知识表示学习模型，可以预测两个实体的关系；除此之外，还可以用于关系抽取，自动问答，实体链指等任务，展现出巨大的应用潜力。  （2）图卷积神经网络  图卷积神经网络将卷积神经网络推广到非欧几里得数据，属于一种重要的知识表示模型，现有的大多数图卷积神经网络方法都遵循消息传递神经网络框架来进行节点信息聚合。在实现图谱补全达到了较高的指标，能够快速挖掘图结构中的拓扑信息和复杂特征，为处理复杂图结构数据开辟了新空间。  **2.技术路线和实施方案**  提出一个知识图谱补全的encoder-decoder图卷积网络模型。encoder部分采用卷积网络方式学习知识图谱三元组节点和关系的表示，获得节点关系的更新；decoder部分以整个三元组作为输入，采用多尺度卷积特征提取三元组整体语义特征进行评估打分。总体框架如图1所示。    图1 模型总体框架  （1）节点关系表示学习方法：随机初始化知识图谱中的节点关系作为encoder部分输入，通过卷积提取邻居节点和关系的特征，由不同关系对应的转换矩阵映射到节点空间，再聚合所有邻居节点信息，得到中心节点向量更新。最后，利用线性变换的方式对关系向量更新。  （2）三元组整体语义特征获取方法：为了获取三元组整体语义特征，采用多尺度卷积神经网络。基本思想是在进行卷积操作时，不再使用单一卷积核，改用多个不同大小的卷积核，有利于挖掘三元组内部潜在的语义特征。以头实体、关系实体和尾实体向量组成的三列矩阵作为输入，使用不同大小的卷积核对输入矩阵进行卷积操作提取特征，得到的特征图再经过全连接层来计算得分，最后对不同尺度下的得分进行加权求和得到最终的分数，作为评估三元组可信度的依据，实现图谱补全。 |
| 2.研究计划可行性，研究条件落实情况，可能存在的问题及解决办法（可续页）  **1.研究计划可行性**  在理论实践方面，知识图谱补全的思路是学习节点和关系的表示，然后对三元组进行评价打分，将正样本和负样本区分，得分越接近正样本的三元组合理程度较高，即该三元组事实成立。节点关系的表示学习模型越来越多，还可以结合语言文本获取语义特征，尤其是以transformer和bert为代表的自然语言处理预训练模型，因此可以提供一定的理论支撑。  在算法设计方面，卷积神经网络能较好地获取节点关系特征，在实际运用中较为成熟，在特征提取方面取得了不错的效果，还可以考虑节点关系特征交互增强的方式，更好地提取特征，最后设置消融实验验证效果。  在实验与应用方面，现有几种公开的知识图谱数据集，如WN18RR、FB15k-237等，可以从中进行试验，检验效果。  **2.研究条件落实情况**  （1）硬件设施  个人计算机:1台，CPU 2.30GHz，内存8G，硬盘1T  服务器一台:1台，CPU E5-2660 v4 @ 2.00GHz，GPU Tesla K80 4张，内存40G，硬盘200G  （2）软件设施  操作系统：centos7.0，开发平台：pycharm，开发语言：python  **3.可能存在的问题及解决办法**  可能存在的问题：实验负样本集合生成，可能会出现测试集泄露的问题，负样本集合在数据集已经出现，造成实验数据不准确。  拟解决办法：采用随机替换正样本头尾实体的方法，删除负样本集合中已经在数据集中出现过的三元组。 |

|  |  |  |
| --- | --- | --- |
| **3.研究计划及预期成果** | | |
| 研  究  计  划 | 起止年月 | 完成内容 |
| 2022.1-2022.7 | 实现表示学习，完成encoder部分 |
| 2022.8-2023.1 | 实现图谱补全任务，完成decoder部分 |
| 2023.2-2023.3 | 学位论文撰写 |
| 2023.4-2023.5 | 学位论文完善及答辩 |
| 预  期  创  新  点  及  成  果  形  式 | 1.预期的创新点  （1）基于卷积网络的节点关系联合表示学习。实体和关系通过卷积操作，提取交互特征，并且对不同关系类型（正、反、自循环）使用不同的投影矩阵，聚合所有邻居节点和关系信息，实现中心节点更新，最后利用关系转换矩阵实现关系更新。  （2）基于多尺度卷积的三元组整体语义特征提取方式。将学习到的三元组表示组成三列矩阵作为输入，使用不同尺度的卷积核提取三元组整体特征，得到的feature maps经过全连接计算得分，实现图谱补全。  2.预期成果  硕士论文1篇 | |

1. 开题报告审查意见

|  |  |  |  |
| --- | --- | --- | --- |
| 1.导师对学位论文选题和论文计划可行性意见，是否同意开题：  校内导师（组）签字： 年 月 日  校外导师签字： 年 月 日 | | | |
| **2.开题报告考评组意见** | | | |
| 开题日期 |  | 开题地点 |  |
| 考评专家 |  | | |
| 考评成绩 | 合格 票 基本合格 票 不合格 票 | | |
| 结 论 | □通过 □原则通过 □不通过  **通过：**表决票均为合格  **原则通过：**表决票中有1票为基本合格或不合格，其余为合格和基本合格  **不通过：**表决票中有2票及以上为不合格 | | |
| 考评组对学位论文的选题、研究计划及方案实施的可行性的意见和建议： | | | |
| 考评组签名：  年 月 日 | | | |
| **3.学院意见：** | | | |
| 负责人签名： 年 月 日 | | | |