

心於至善

基于网络角色的知识图谱

缺失知识补全方法

王紫悦

东南大学

学校代码: 10286
分类号: TP393
密级: 公开
UDC: 004.9
学号: 184545



SOUTHEAST UNIVERSITY

东南大学 硕士学位论文

基于网络角色的知识图谱 缺失知识补全方法

研究生姓名: 王紫悦

导师姓名: 张祥 副教授

王小鹏 高工

申请学位类别 工程硕士 学位授予单位 东南大学

一级学科名称 计算机技术 论文答辩日期 2021年5月27日

二级学科名称 知识图谱 学位授予日期 2021年6月 日

答辩委员会主席 倪庆剑 评阅人 院盲



2021年6月2日

学校代码: 10286
分类号: TP393
密 级: 公开
U D C: 004.9
学 号: 184545



东南大学

硕士学位论文

基于网络角色的知识图谱 缺失知识补全方法

研究生姓名: 王紫悦

导师姓名: 张祥 副教授

王小鹏 高工

申请学位类别 工程硕士 学位授予单位 东南大学

一级学科名称 计算机技术 论文答辩日期 2021 年 5 月 27 日

二级学科名称 知识图谱 学位授予日期 2021 年 6 月 日

答辩委员会主席 倪庆剑 评 阅 人 院盲

2021 年 6 月 2 日

東南大學

硕士学位论文

基于网络角色的知识图谱 缺失知识补全方法

专业名称: 计算机技术

研究生姓名: 王紫悦

导师姓名: 张祥 副教授

王小鹏 高工

KNOWLEDGE GRAPH COMPLETION BASED ON NETWORK ROLE

A Thesis submitted to

Southeast University

For the Academic Degree of Master of Engineering

BY

Wang Ziyue

Supervised by:

Associate Prof. Zhang Xiang

and

Senior Engineer Wang Xiaopeng

School of Cyber Science and Engineering

Southeast University

2021/6/2

东南大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：_____日期：_____

东南大学学位论文使用授权声明

东南大学、中国科学技术信息研究所、国家图书馆、《中国学术期刊（光盘版）》电子杂志社有限公司、万方数据电子出版社、北京万方数据股份有限公司有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括以电子信息形式刊登）论文的全部内容或中、英文摘要等部分内容。论文的公布（包括以电子信息形式刊登）授权东南大学研究生院办理。

研究生签名：_____导师签名：_____日期：_____

摘 要

知识图谱作为人工智能技术的重要组成部分，被广泛应用于智能问答、智能搜索、个性化推荐等领域中。然而，知识图谱由于其动态增长的特性，始终存在着数据稀疏的问题，影响着人工智能模型的精度。知识图谱中知识的缺失问题亟待解决，知识图谱补全的概念得到了学术界和工业届的广泛关注。

现有的知识图谱补全方法大多依赖于知识图谱中实体间的关系，然而，真实知识图谱中部分实体由于较为罕见有较少的关系可达，这使得学习得到的向量表示质量较低，从而影响模型的精度。此外，由于处理非离散数据类型的属性值具有一定的挑战性，现有的先进的关系学习模型大多忽略了这些属性信息，在属性预测方面的研究较少。并且，现有的知识补全模型大都是基于原知识图谱中的三元组结构，这需要用户具有知识图谱的背景，才能迅速理解预测结果，忽视了可理解性的重要性。基于现有的工作背景，本文提出了一种基于网络角色的知识图谱缺失知识补全方法，来解决上述问题，主要研究内容如下：

(1) 提出一种基于网络角色的知识图谱实体嵌入方法。根据实体间不同层面的相似等效性，从同质性、属性相似性、结构相似性三个方面来进行实体角色发现，生成不同的实体路径，得到语义丰富的实体嵌入表示，解决了表示学习受知识图谱中关系稀疏影响较大的问题。

(2) 提出一种基于实体标签的知识补全方法。将画像标签结果作为预测模型的输入输出，用户通过少量的实体画像标签就可以快速理解预测结果，增强了结果的可理解性。同时将连续值属性离散化，利用模糊预测代替属性值的精确预测，属性预测的结果由确切的数值变为值域范围。

(3) 设计并实现基于网络角色的知识补全系统。该系统能够搜索查询目标实体，动态显示知识图谱中的实体画像标签补全的结果。同时，显示利用本研究的方法预测出的的标签补全结果的拟合度情况。

综上所述，本文研究了基于网络角色的知识图谱缺失知识补全方法：首先基于实体角色相似性生成实体嵌入表示，该实体嵌入表示用于生成实体画像标签。然后以实体标签作为知识补全模型的输入输出，进行标签预测任务，在真实的知识图谱数据集上进行实验，结果表明该方法的预测效果明显优于大多数现有模型。最后，设计并实现了知识图谱补全系统。

关键词： 知识图谱补全，网络角色，实体画像，标签预测

Abstract

As an important part of artificial intelligence technology, Knowledge Graphs (KGs) have been widely used in intelligent question-answering, intelligent search, personalized recommendation and other fields. However, due to the dynamic growth of knowledge graph, the existence of data sparsity affects the accuracy of artificial intelligence model. The problem of incompleteness of knowledge graph needs to be solved urgently. The concept of knowledge graph completion has been widely concerned in academia and industry.

Most of the existing knowledge graph completion methods rely on the relations between entities in the knowledge graph. However, some entities in the real knowledge graph are rare and have fewer relations, which leads to low quality of entity representation learned, thus affecting the accuracy of the model. In addition, due to the challenge of processing non-discrete attribute values, most of the existing advanced relational learning models ignore these attribute information, and there are few researches on attribute prediction. Moreover, existing knowledge completion models are based on the triples of the knowledge graph, which requires users who have the background of the knowledge graph to quickly understand the prediction results, ignoring the importance of comprehensibility. Based on the existing researches, a novel knowledge graph completion method based on network role is proposed to solve the above problems. The main content is as follows:

(1) A knowledge graph entity embedding method based on network role is proposed. According to the different similarities between entities, the entity role is discovered from three aspects of homogeneity, attributive similarity and structural similarity. Then different entity paths are generated to obtain representations of entities which have rich semantic information, solving the problem that the representation learning is negatively affected by the sparse relation in the knowledge graph.

(2) A novel knowledge graph completion method based on entity label is proposed. With entity labels as the input and output of the model, users can quickly understand the prediction results through a small number of entity labels, which enhances the comprehensibility of the results. At the same time, the non-discrete attribute value is discretized, and the fuzzy prediction is used to replace the exact prediction of the attribute value, so the result of attribute prediction is changed from the exact value to the range.

(3) Design and implement a knowledge graph completion system based on network role. The system can search the target entity and display the completion result of entity profiles in the knowledge graph dynamically. At the same time, the degree of fitting of the corresponding

label prediction results are shown.

In summary, a novel knowledge graph completion method based on network role is proposed: Firstly, the entity representations which are generated based on the entity role similarity, are used to generate the entity profiles. Then, with the entity labels as the input and output of knowledge graph completion model, new entity labels are predicted. Experimental results on real knowledge graph datasets show that the prediction performance of the proposed method is better than most existing models. Finally knowledge graph completion system is designed and implemented.

Keywords: Knowledge Graph Completion, Network Role, Entity Profiles, Label Prediction

插图目录

1.1 可理解性比较：(a) 基于三元组的补全方法 vs. (b) 基于网络角色的补全方法	2
2.1 知识图谱 RDF 描述示例	5
2.2 知识图谱技术架构	6
2.3 网络角色示例图	7
2.4 基于特征的网络角色发现方法框架	8
2.5 知识图谱补全框架	10
2.6 TransE 核心思想	11
2.7 CPL 流程框架	12
2.8 路径推理	13
3.1 基于网络角色的知识图谱缺失知识补全方法框架图	19
3.2 实体同质性示例	21
3.3 基于同质性策略生成的实体路径	21
3.4 实体属性相似性示例	23
3.5 基于属性相似性策略生成的实体路径	23
3.6 实体结构相似性示例	25
3.7 基于结构相似性策略生成的实体路径	25
3.8 实体画像示例图	27
3.9 实体画像方法流程图	28
3.10 基于实体标签的端到端模型示例	31
3.11 基于实体标签的知识图谱补全方法框架图	32
3.12 GCN 网络结构	33
3.13 DisMult 模型结构	34
4.1 实体画像结果对比：(a) Random 方法 vs. (b) HAS 方法	41
4.2 路径权重参数分析	48
4.3 向量维度参数分析	49
4.4 训练集比例敏感性分析	50
4.5 可理解性实验：实验组示例	52
4.6 可理解性实验：对照组示例	52
4.7 用户问卷所耗时间分布情况	53

4.8	用户问卷正确率分布情况	54
5.1	知识图谱补全系统用例图	55
5.2	知识图谱补全系统模块示意图	57
5.3	实体画像生成流程图	58
5.4	实体标签预测流程图	59
5.5	系统主界面演示	60
5.6	实体画像模块演示	60
5.7	模糊匹配查询实体	61
5.8	实体画像结果	61
5.9	<i>Beastie_Boys</i> 在 DBpedia 中的详细描述信息	62
5.10	实体标签预测模块演示	62
5.11	标签预测结果	63

表格目录

2.1	网络角色发现方法比较	10
2.2	基于 <code>embedding</code> 的知识图谱补全方法核心公式	14
2.3	知识图谱补全相关工作总结	15
3.1	属性连续值离散化示例	20
3.2	标签的分类和相关示例	29
4.1	数据集统计信息	38
4.2	各数据集标签统计信息	41
4.3	数据集 FB15k 和 FB15k-237 的实体画像评估结果	42
4.4	数据集 Band、University、Book 的实体画像评估结果	43
4.5	数据集 RadioStation 和 Actor 的实体画像评估结果	44
4.6	数据集 FB15k 和 FB15k-237 的实体标签预测结果	45
4.7	数据集 Band、University、Book 的实体标签预测结果	46
4.8	数据集 RadioStation 和 Actor 的实体标签预测结果	47
4.9	属性预处理消融实验结果	51
4.10	基于实体标签的知识图谱补全方法外部实验评估结果	53
5.1	实体画像生成模块主要接口函数	58
5.2	实体标签预测模块主要接口函数	59

算法目录

3.1	基于同质性策略生成的实体路径	22
3.2	基于属性相似性策略生成的实体路径	24
3.3	基于结构相似性策略生成的实体路径	26
3.4	NRKC 算法	35

术语与符号约定

\mathcal{G}	知识图谱
\mathcal{V}	知识图谱的节点集合
\mathcal{E}	知识图谱的边集合
\mathcal{T}	知识图谱的实体类型集合
E	知识图谱的实体集合
V_a	知识图谱的属性值集合
\mathcal{E}_r	知识图谱的关系集合
\mathcal{E}_a	知识图谱的属性集合
$P(v)$	以实体节点 v 为起点生成的路径
P^H	基于同质性生成的实体路径集合
P^A	基于属性相似性生成的实体路径集合
P^S	基于结构相似性生成的实体路径集合
$l_{property}, l_{value}$	标签属性, 标签属性值
\mathcal{P}	实体画像集合
\mathbb{L}	标签集合
\mathcal{B}	实体标签对组成的二元组集合

目录

摘 要	I
Abstract	III
插图目录	V
表格目录	VII
算法目录	IX
术语与符号约定	XI
第一章 绪论	1
1.1 研究背景与意义	1
1.2 研究目标与内容	2
1.3 论文结构与安排	3
第二章 相关工作	5
2.1 知识图谱	5
2.2 网络角色发现	6
2.2.1 基于图结构的方法	7
2.2.2 基于特征的方法	8
2.2.3 混合方法	9
2.2.4 各类网络角色发现方法比较	9
2.3 知识图谱补全	9
2.3.1 基于表示学习的方法	11
2.3.2 基于关系路径推理的方法	12
2.3.3 基于规则的方法	13
2.3.4 动态知识图谱补全方法	14
2.3.5 各类知识补全方法的比较	15
2.4 本章小结	16
第三章 基于网络角色的知识图谱缺失知识补全方法	17
3.1 问题描述	17

3.2	模型框架	18
3.3	属性预处理	19
3.4	实体角色发现	20
3.4.1	基于同质性的角色发现方法	20
3.4.2	基于属性相似性的角色发现方法	21
3.4.3	基于结构相似性的角色发现方法	24
3.4.4	路径混合	26
3.5	实体画像	27
3.5.1	构建标签池	28
3.5.2	标签度量	29
3.5.3	生成标签集	30
3.6	基于实体标签的知识补全方法	30
3.6.1	图卷积神经网络	32
3.6.2	实体标签预测	33
3.7	本章小结	35
第四章	实验设计和评估	37
4.1	数据集	37
4.2	基准方法和评估指标	39
4.2.1	基准方法	39
4.2.2	评估指标	40
4.3	实体画像结果评估	41
4.4	实体标签预测任务	43
4.4.1	实验结果及分析	44
4.4.2	超参数分析	47
4.5	连续值离散化	50
4.6	可理解性任务	51
4.7	本章小结	54
第五章	系统设计与实现	55
5.1	系统需求分析	55
5.1.1	功能需求	55
5.1.2	性能需求	56
5.2	概要设计	56
5.2.1	系统架构	56
5.2.2	系统设计	56
5.3	详细设计	57

5.3.1	实体画像生成	58
5.3.2	实体标签预测	59
5.4	系统演示	60
5.5	本章小结	63
第六章	总结与展望	65
6.1	工作总结	65
6.2	未来展望	66
致谢		67
参考文献		69
作者攻读硕士学位期间的研究成果		77

第一章 绪论

1.1 研究背景与意义

知识图谱 (Knowledge Graph) 作为人工智能技术的重要组成部分, 提供了一种更好地组织、管理和理解互联网海量信息的能力, 将互联网的信息表达成更接近于人类认知世界的形式。知识图谱最早是谷歌于 2012 年提出的概念, 其将人类知识通过链接关系以结构化形式呈现出来, 用于迅速描述物理世界中的概念及其相互关系, 引起了学术界和工业届广泛关注。知识图谱由实体、关系和语义描述组成。实体可以是真实世界中存在的对象, 也可以是抽象的概念; 关系则表示实体之间的关联; 实体及其关系的语义描述包含定义好的类型和属性。知识图谱给智能搜索带来了活力, 同时也在智能问答、大数据分析 & 决策中显示出强大威力, 已经成为智慧城市服务的基础设施。知识图谱与大数据和深度学习一起, 成为推动人工智能发展的核心驱动力之一。

知识图谱由于其动态增长的特性, 始终存在着数据稀疏问题。基于知识图谱的人工智能模型的精度往往受限于数据的质量, 当做实体相关性计算或是知识推理工作时, 准确率将大大降低。这极大地影响了人工智能系统的安全, 成为更加广泛应用知识图谱的障碍之一。因此, 知识图谱中知识的缺失问题亟待解决, 提出了知识图谱补全的概念。知识图谱补全的目的是预测出三元组中缺失的部分, 从而使知识图谱变得更加完整。

现有的知识补全方法主要分为基于表示学习的、基于关系路径推理的、基于规则的和动态知识图谱补全方法这四类。这些方法各有优劣, 适用于不同的应用场景。目前, 知识图谱补全方向仍存在以下亟待解决的问题:

(1) 关系稀疏问题。现有的知识图谱补全方法大多依赖于知识图谱中实体间的关系, 经典的 $\text{transE}^{[1]}$, $\text{transH}^{[2]}$ 等翻译模型利用空间向量的平移不变性, 将实体与关系映射到向量空间中, 使得头实体与关系的向量表示之和尽可能等于尾实体向量。 $\text{RESCAL}^{[3]}$ 作为张量分解模型代表性方法, 将知识图谱中的三元组集合表示为张量, 如果该三元组为正例, 则对应元素置为 1, 最终从中分解出实体和关系的向量表示。近年来, 实体和关系的表示学习获得了显著的成果。然而, 真实知识图谱中部分实体由于较为罕见有较少的关系可达, 这使得学习得到的向量表示质量较低, 从而影响模型的精度。

(2) 属性预测的低准确度问题。由于处理非离散数据类型的属性值具有一定的挑战性, 现有的先进的关系学习模型大多忽略了这些属性信息。目前在属性预测方面的研究较少, 已有的大多数方法在数值型属性预测上表现较差。例如, 要预测出 $\langle \text{Company}, \text{num_employees}, 1204 \rangle$ 中的属性值的精确值 “1204”, 对于现有的知识补全方法来说仍有一定的难度。 $\text{MT-KGNN}^{[4]}$ 利用图神经网络进行训练学习, 但预测连续值属性的结果仍有一定的偏差。

(3) 可理解性问题。现有的知识补全模型表现出较高的准确率，它们大都是基于原知识图谱中的三元组结构，却忽视了可理解性的重要性。在某些情况下，用户需要理解补全的结果和影响补全过程的因素，以知识图谱三元组作为输入输出的模型，需要用户具有知识图谱的背景，才能迅速理解预测结果。因此，补全方法应当对知识补全的过程提供一定的解释。

本研究采用一种基于网络角色的知识图谱缺失知识补全方法来解决上述问题。受网络角色发现方法的启发，从同质性、属性相似性、结构相似性三个方面来学习实体的向量表示，避免了实体的嵌入表示受知识图谱关系稀疏性影响的问题。对于属性预测的问题，将连续值属性离散化，属性预测的结果由一个精确的值变为一个模糊的值域范围，同时能够满足用户对属性预测的需求。在知识补全任务上，与已有的方法不同，本研究将实体画像标签作为输入输出，使结果更具可理解性，帮助快速理解。如图1.1所示，红色虚线代表预测结果，黑色实线代表原有的边，蓝色椭圆形节点表示实体，方框表示实体属性或标签。(a) 图表示以知识图谱中的三元组作为输入输出，用户需要阅读冗长且复杂的整张图，并且理解图中所有节点。而 (b) 图表示以实体画像的标签结果作为输入输出，用户通过少量的标签就可以快速理解预测结果。例如，(b) 图中只需要依靠标签 $\langle know, \langle people, age, [20, 30] \rangle \rangle$ ，就能直接理解预测的标签结果 $\langle age, [20, 30] \rangle$ 。

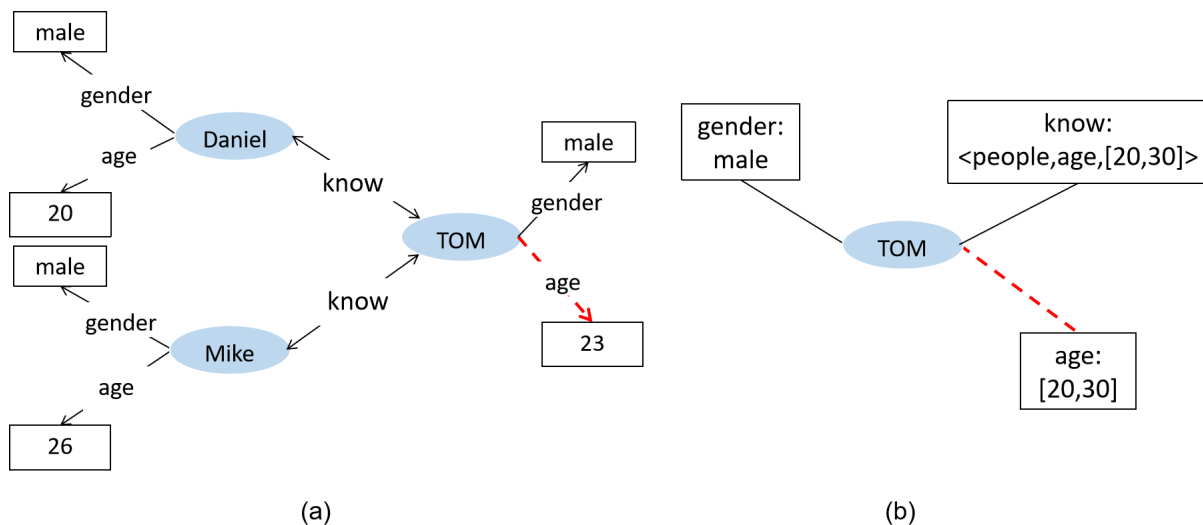


图 1.1: 可理解性比较: (a) 基于三元组的补全方法 vs. (b) 基于网络角色的补全方法

1.2 研究目标与内容

本课题的研究目标是根据知识图谱中实体间不同层面的相似等效性，挖掘不同的实体特征，生成实体画像，并将画像结果应用于知识补全任务中，提高可理解性。基于网络角色的知识图谱缺失知识补全方法的研究内容包括如下三点：

(1) 提出一种基于网络角色的知识图谱实体嵌入方法，用于解决知识图谱中关系稀疏问题。根据实体间不同层面的相似等效性，从同质性、属性相似性、结构相似性三个

方面来进行实体角色发现，生成不同的实体路径，利用 skip-gram 得到语义丰富的实体嵌入表示。

(2) 提出一种基于实体标签的知识补全方法，用于解决可理解性问题。将画像结果作为预测模型的输入输出，用户通过少量的实体画像标签就可以快速理解预测结果，增强了结果的可理解性，保证了人工智能安全。并且将连续值属性离散化，属性预测的结果不再是一个确切的数值，而是一个值域范围，同时能够满足用户对属性预测的需求。

(3) 设计并实现基于网络角色发现的知识补全系统。系统为网页形式，根据选择的数据集，能够搜索相应的实体，动态显示知识图谱中的实体画像标签补全的结果。同时，显示利用本研究的方法预测出的的标签补全结果的拟合度情况。

1.3 论文结构与安排

本文共分为六章阐述基于网络角色的知识图谱缺失知识补全方法的设计与实现，具体内容安排如下：

第一章，主要介绍本课题的研究背景、研究现状和研究内容。描述了目前的知识图谱不完整性问题和知识补全方法的研究现状，提出了本研究的知识补全方法，最后阐述了本文的研究目标和研究内容。

第二章，主要介绍与本研究相关的工作和理论，并阐述它们与本研究之间的关联，包括知识图谱的相关知识、网络角色发现以及现有的知识图谱补全研究方法。

第三章，主要介绍基于网络角色的知识图谱缺失知识补全方法的核心框架设计，包括问题描述、模型框架、实体角色发现、实体画像方法以及知识补全模型。利用一种基于网络角色的实体嵌入方法解决关系稀疏问题，使用画像标签来提高补全结果的可理解性。

第四章，主要介绍了实验数据集和评估方法，与现有的知识补全主流方法进行对比，分析优劣性。

第五章，主要介绍知识补全系统的需求分析、设计和实现，展示系统演示结果。

第六章，主要对本研究提出的方法进行总结，并结合研究中发现的问题，对未来进一步的研究工作给予展望。

第二章 相关工作

本章主要介绍了与本研究相关的工作和理论，并阐述它们与本研究之间的关联。首先介绍了知识图谱的相关知识和理论，其次说明了网络角色发现的含义和现有方法，最后介绍了现有的知识图谱补全研究方法，并分析其优劣性，对本研究给予一定的启发。

2.1 知识图谱

随着信息化时代的到来，网络数据多源异构，难于管理，需要人们从新的视角去探索新的知识互联方法，深入地体现人类认知的整体性和互联性。知识图谱 (knowledge graph, KGs) 以其强大的语义处理能力与开放互联能力，为新的知识互联方法奠定了扎实的基础。通过知识图谱能够将网络上的信息、数据以及链接关系聚集为知识，使信息资源更易于计算、理解以及评价。

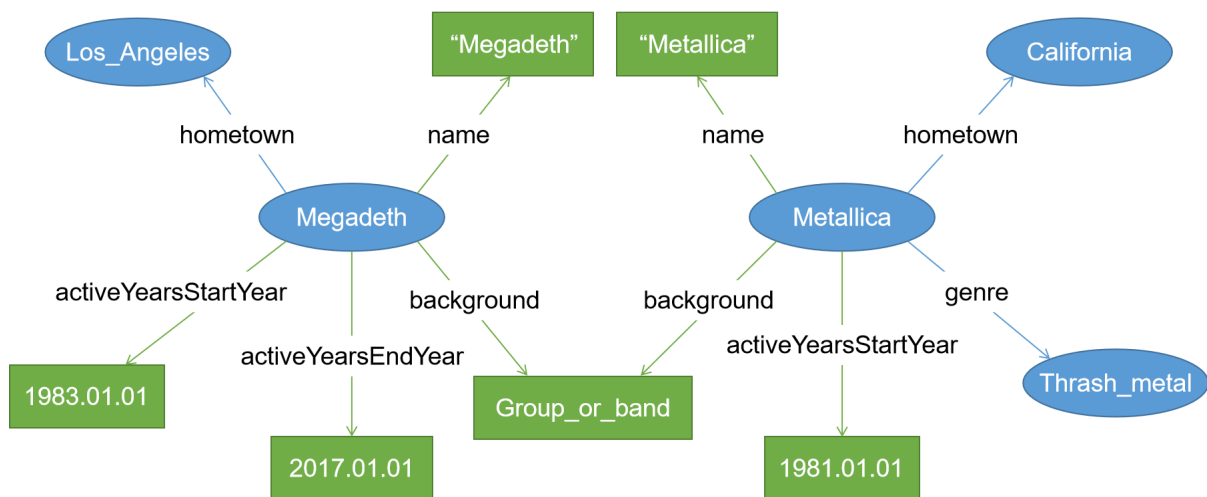


图 2.1: 知识图谱 RDF 描述示例

在维基百科中，知识图谱被定义为谷歌用于增强其搜索引擎功能的知识库。同时，知识图谱能够揭示实体之间的关系，并且对现实世界中的事物和其关系进行形式化的描述。它以“实体-关系-实体”和“实体-属性-属性值”的形式来存储事实信息，被广泛应用于实体搜索^[5]、数据集成^[6]等任务中。大部分的知识图谱使用 RDF (Resource Description Framework) 来描述事物，如图2.1所示，给出了 DBpedia 数据集中乐队类型的“Megadeth”和“Metallica”两个实体的描述信息。其中，蓝色圆形节点代表实体，蓝色直线代表实体间的关系，例如 $\langle \text{Megadeth}, \text{hometown}, \text{Los_Angeles} \rangle$ 表示乐队实体 Megadeth 的家乡在洛杉矶。绿色方形节点代表属性值，绿色直线代表实体的属性名，属性值又可以分为字符类型和数值类型。字符类型的属性值通常以文本的形式来

描述实体，例如 $\langle \text{Megadeth}, \text{background}, "Group_or_band" \rangle$ 表示 *Megadeth* 的背景是“乐队或组合”；数值类型的属性值通常以整数或浮点数的形式来提供离散化的实体信息，例如 $\langle \text{Megadeth}, \text{ActiveYearsStartYear}, 1983.01.01 \rangle$ 表示 *Megadeth* 开始活跃的年份是 1983 年。

知识图谱的技术架构如图 2.2 所示，主要包括信息抽取、知识获取、知识推理补全、质量评估四个部分。知识图谱构建从最原始的数据（包括结构化、半结构化、非结构化数据）出发，采用一系列自动或者半自动的技术手段进行信息的抽取。利用实体链接、关系链接、知识融合等步骤获取知识，并将其存入知识库的数据层和模式层。通过知识补全的方法，不断提高知识图谱的质量。随着人类的认知能力、知识储备和业务需求的不断增加，知识图谱也需要与时俱进，不断地迭代更新，扩展知识。因此，对于知识图谱的质量管理，知识补全环节至关重要。

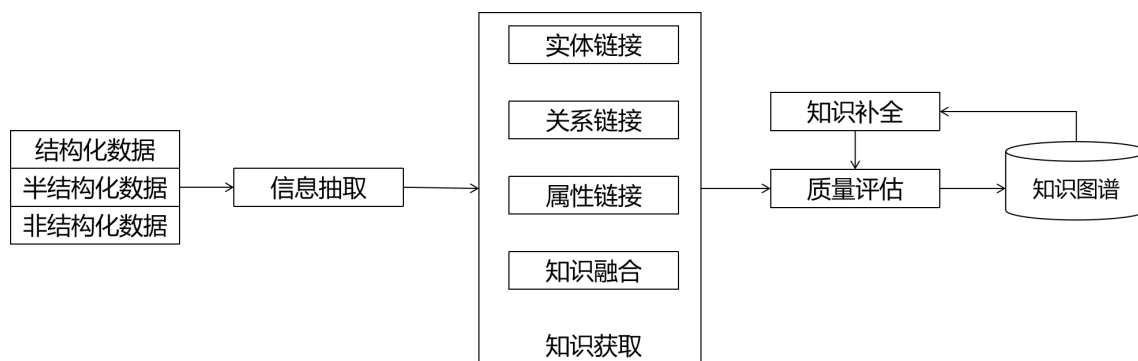


图 2.2: 知识图谱技术架构

随着网络资源的数量激增，大量数据被发布和共享，学术界与工业界的研究人员花费了大量的精力构建了高质量的大规模开放知识图谱。本研究使用的 DBpedia^[7] 是由德国莱比锡大学和曼海姆大学的科研人员创建的多领域多类型的综合型知识库，在链接开放数据（Linked Open Data, LOD）中处于最核心的地位。DBpedia 是从维基百科中抽取得到的结构化信息，以关联数据的形式发布到互联网上，为在线网络应用、社交网站以及其他知识库提供服务。截止至 2014 年年底，DBpedia 中的事实三元组数量已经超过了 30 亿条，被广泛地用于知识表示、知识获取和知识应用等任务中。

2.2 网络角色发现

网络角色的概念最早出现在社会科学中，其含义是个体倾向于在互动网络中扮演某种角色或承担某种职责，角色发现就是将在结构上具有相似特征的节点划分为同一角色的过程。例如，在一所大学里，每个人都可以被划分为教师、行政人员、工作人员或学生的角色。每个角色可以进一步划分为子角色：教师可以进一步划分为终身教职或非终身教职等。图 2.3 给出了网络角色的示例，在该网络中，具有相同形状的节点被分配了同样的角色，用统一的颜色标出。如果两个节点具有相似的结构模式，那么它们属于同一

个角色。近年来,研究人员发现角色不仅出现在其他类型的网络中,比如食物网^[8],世界贸易^[9],还可以帮助预测节点在其领域内的作用。例如,在蛋白质相互作用网络中,具有相似角色的蛋白质往往具有相似的代谢功能。因此,如果我们知道某种蛋白质的功能,我们可以预测所有其他具有类似作用的蛋白质也会具有类似的功能^[10]。

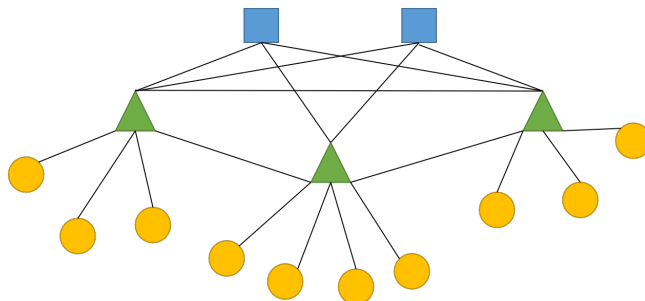


图 2.3: 网络角色示例图

知识图谱的本质是大规模的网络结构,许多知识图谱相关的研究汲取了网络角色发现领域的工作和技术,如实体中心度,社区探测,网络连通性以及节点相似性等。角色发现,就是基于某种相似关系将节点划分为同一种角色。它的关键问题在于如何定义角色的等价性,即知识图谱中的节点相似度,其研究方法主要分为基于图结构的方法、基于特征的方法和混合方法。基于图结构的角色发现方法是直接从图中计算出来,不包含任何特性。相反,基于特征的角色发现方法是通过将图形数据转换为新的特征表示来计算的。当然,也可以有混合的方法来利用两者的优势。

2.2.1 基于图结构的方法

基于图结构的方法中,将等价性划分为四类:

- (1) 结构等价性,即两个节点具有完全相同的邻居。
- (2) 自同构等价性。两个节点通过自同构映射能够互相交换位置,则它们是自同构等价的。
- (3) 正则等价性,即两个节点的直接邻居具有等价性。
- (4) 随机等价性。它认为节点与图中其它节点相连的概率分布是相同的,与角色无关。

基于这四类等价性,对图结构信息进行抽取和计算,所使用的算法模型可分为块模型和基于邻接矩阵的行列相似性方法。块模型作为在网络分析中应用最广泛的技术之一^[11],通过角色交互图来表示网络,其中节点代表角色,边代表角色间的联系。随机块模型^[12]就是基于随机等价性,由于其不对图结构作任何假设,在发现未知网络结构上具有较高的准确性。基于邻接矩阵的行列相似性的方法,利用行/列的相似性对节点进行聚类分析^[13;14]。

基于图结构的方法主要通过邻接矩阵来计算网络的结构特征,具有严苛的等价性要求,但在大规模的知识图谱中,矩阵计算耗时费力,无法得到很好的应用。

2.2.2 基于特征的方法

基于特征的角色发现方法，利用节点的结构特征更加泛化地定义了等价性，能够很好地应用于大型的数据集中。它将等价性定义为节点特征相似性，如果两个节点拥有相似的特征值，则它们属于同一种角色。

Rossi 总结了近年来的研究，提出了一般性的基于特征的角色发现方法的框架^[15]，该方法框架包括角色特征构建和角色发现两个步骤。如图2.4所示，对于给定的网络图结构，首先通过某种角色特征构建技术将图转换为基于图的特征的向量表示；然后，基于某一类的矩阵分解或聚类算法，将拥有相似特征的节点分配为相同的角色，在图中用相同的颜色标出。

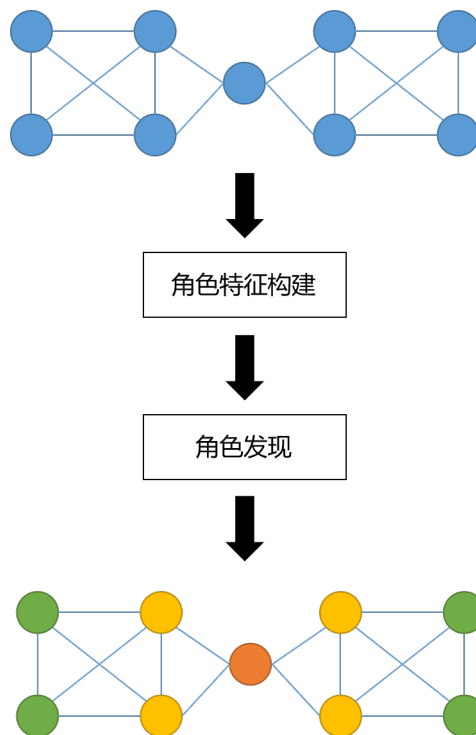


图 2.4: 基于特征的网络角色发现方法框架

在该框架中，角色特征构建和角色发现所采用的方法都是灵活多变的，可以选择不同的网络特征来构架图的结构特征，而角色发现步骤也可以选择矩阵分解或奇异值分解的方法来实现。该框架同时还具备可拓展性，能够依据不同的应用场景来进行调整，对应的角色在不同的场景下也具有不同的含义。基于此框架，Refex^[16] 利用节点的度中心性和 PageRank 值来初始化节点，递归生成特征值。RoIX^[17] 提出一种基于无监督的学习方法，来自动提取结构角色。在此基础上，Gilpin 等人^[18] 利用稀疏性和差异性约束，对半监督方法 GLRD 进行了拓展。

基于特征的角色发现方法提供了更大的灵活性来表示复杂的角色，在实际应用中尤为很重要。另一方面，基于特征的方法所提供的灵活性，使识别捕获角色所需的特征变得更加困难，该过程通常需要一些专家的指导或针对特定应用程序进行更多的调优。

2.2.3 混合方法

混合方法结合了基于图结构的方法和基于特征的方法两者的优势，通常在角色特征构造之前使用基于图的方法，或者在学习角色特征表示之后使用图结构方法。

第一类在角色特征构建之前使用基于图的方法的混合角色发现方法，通常应用于关系表示学习中^[19;20]。例如，我们可以使用块模型直接从图中提取角色，然后可以将其视为学习更复杂特征的“初始属性”。一旦初始属性被学习，就可以把它们连同图一起作为关系特征学习系统的输入。利用这些特征，使用一种自动学习角色的数量和分配的技术。由于块模型的限制^[12]，这类方法的主要缺点在于其可拓展性差。尽管如此，其他可拓展性强的基于图结构的方法仍使用这类混合角色发现方法。

第二类混合角色发现方法利用多个数据源作为一种归一化或影响角色分配阶段的方法。除了从图中学习特征以外，一些额外的数据源（例如，图和属性集）对角色发现也有所帮助。这些混合方法可以通过自适应张量分解方法^[21;22]或集合矩阵-张量分解(CMTF)方法^[23]来学习。与张量分解方法不同，集合分解方法可以通过融合多个异构数据源来学习角色。例如，给定一个商店类别矩阵、用户-商店-商品张量和社交网络矩阵。在这种上下文关系中，人们可能会认为用户角色应该受到商店的影响，同样也会受从那家商店购买的商品和商店的类别影响。集合矩阵-张量分解方法允许任意数量的矩阵或张量被分解，反之也允许它们直接影响已学习到的角色的定义。

2.2.4 各类网络角色发现方法比较

在网络角色发现方法中，这两类方法各有优劣，其代表工作和特点如表2.1所示。

传统的基于图结构的角色发现方法，大多只适用于规模相对较小的网络。尽管已经有研究使用降采样对广义块模型^[24]和随机块模型进行缩放，但在这一领域仍有很多工作要做。例如，进一步拓展适用于大规模网络的方法，使用快速推理步骤调整其他的方法。虽然，基于特征的角色发现方法通常比基于图结构的方法更加高效和灵活，但如何高效地学习空间特征仍然是一个挑战，仍需人工介入。

2.3 知识图谱补全

知识图谱补全的目标是预测出三元组中缺失的部分，从而使知识图谱更加完整。给定知识图谱 $G = (E, R, T)$ ，其中实体集 $E = \{e_1, e_2, \dots, e_m\}$ 、关系集 $R = \{r_1, r_2, \dots, r_n\}$ 以及三元组集 $T = \{(e_i, r_k, e_j) | e_i, e_j \in E, r_k \in R\}$ ，知识图谱补全的任务就是推理出缺失的三元组集 $T' = \{(e_i, r_k, e_j) | (e_i, r_k, e_j) \notin T\}$ 。根据三元组中具体的预测对象，知识图谱补全可以分成3个子任务：头实体预测、尾实体预测以及关系预测。现有的知识图谱补全研究方法主要分为基于表示学习的、基于关系路径推理的、基于规则的和动态知识图谱补全方法这四类。

针对知识图谱补全任务，解决问题框架如图2.5所示。

表 2.1: 网络角色发现方法比较

分类	理论依据	代表工作	特点
基于图结构的方法	结构等价性 自同构等价性 正则等价性 随机等价性	块模型 ^[11;12] 、 基于行列相似性方法 ^[13;14]	等价性要求苛刻， 矩阵计算耗时， 不适用于大规模网络
基于特征的方法	节点特征相似性	Rossi ^[15] 、ReFex ^[16] 、 RolX ^[17] 、GLRD ^[18]	极大的灵活性， 捕获特征困难
混合方法	结合图结构和特征方法	ATF ^[22] 、CMTF ^[23]	结合两者优势

(1) 获取数据：静态知识图谱中获取三元组，动态知识图谱中获取包含时序信息的四元组；

(2) 特征提取：通过表示学习、关系路径推理或者规则归纳来学习数据间的特征和联系；

(3) 预测：根据预测对象的不同，进行头实体预测、关系预测、尾实体预测和时间预测；

(4) 评估性能：对预测结果进行评估。

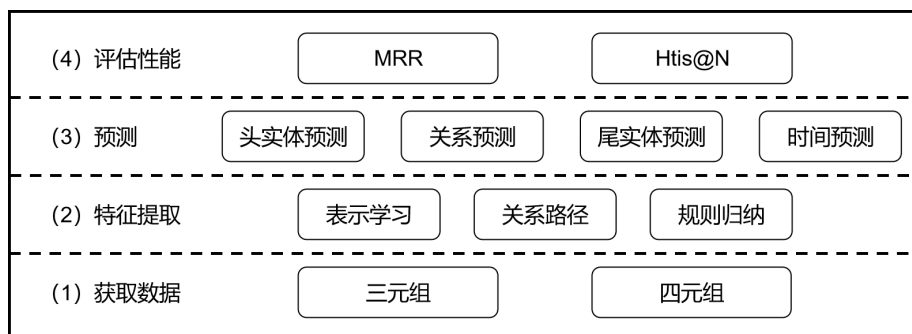


图 2.5: 知识图谱补全框架

2.3.1 基于表示学习的方法

早期的不少工作都是基于 **embedding** 的模型，比较经典的有 TransE^[1]、TransH^[2]、TransR^[25]。图2.6描述了 TransE 的核心思想，将三元组 (h, r, t) 中的头实体、关系、尾实体分别映射到向量空间中，认为它们之间存在着传递性的联系，即 $h + r \approx t$ ，通过训练学习实体和关系的嵌入表示。HolE^[26] 在 Trans 系列基础上，将组合运算符加入到实体的表示上。这些模型在小规模的知识图谱上表现良好，然而随着知识图谱规模的扩大，数据稀疏问题会加重，算法的效率也随之降低。

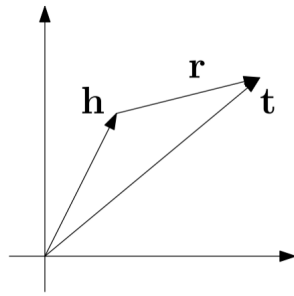


图 2.6: TransE 核心思想

这些经典的工作都是基于低维度的向量空间，若想要适应大规模的知识图谱，则需要更加复杂的特征空间，ComplEx^[27] 使用了复杂的向量表示来处理大量的二元关系。ProjE^[28] 提出了一种对实体和边同时进行表示学习的方法，避免了复杂性。IRNs^[29] 通过一个控制器实现共享内存，构建多跳关系模型。ANALOGY^[30] 利用实体间关系的类比推理，学习复杂关系的表示。

虽然这些工作降低了复杂性，但是没有很好地区分关系预测和实体预测，它们只能预测三元组中的头实体 $(?, r_k, e_j)$ 或尾实体 $(e_i, r_k, ?)$ ，无法对关系 r 进行预测。针对该问题，SENN^[31] 构建了一个神经网络框架，使实体和关系共享 **embedding**，并将知识图谱补全划分三个子任务——头实体预测、关系预测、尾实体预测。

现有的方法大多基于封闭世界假设，即知识图谱是固定的，ConMask^[32] 放宽了该假设，提出了开放世界的 KGC 任务。首先选择与给定关系相关的单词，从中提取实体的描述信息。然后从相关文本中，使用全卷积网络 (FCN)、语义平均化学习到单词的嵌入。该模型利用实体相关的文本描述信息来学习 **embedding**，避免了对知识图谱中已有关系的过度依赖，却更加依赖文本信息。

基于表示学习的知识图谱补全方法大多是利用神经网络训练学习，来获得三元组中语义特征，然后对候选集进行排序。这种方法仍停留在实体间直接关系的层面上，无法进行复杂关系的推理。

2.3.2 基于关系路径推理的方法

近年来，实体和关系的表示学习获得了显著的成果，但是它无法对复杂的多跳关系路径进行有效分析。比如说，对于三元组 $\langle Jiangsu, locatedIn, China \rangle$ 、 $\langle Nanjing, locatedIn, Jiangsu \rangle$ ，embedding 的方法无法推理出 $\langle Nanjing, locatedIn, China \rangle$ 。因此，研究者们转为探索如何利用图结构上的路径信息来实现知识图谱补全，即关系路径推理（Relation Path Inference）。DIVA^[33] 提出将多跳关系推理划分为两个子步骤——路径搜索和路径推理。

2.3.2.1 路径搜索（path finding）

路径搜索的目的是在两个给定的实体之间找到路径，但不能衡量其质量好坏。大多数的研究工作将强化学习引入到多跳关系推理中，将搜索实体对之间的路径看作马尔可夫决策过程（Markov decision process, MDP）。把 MDP 定义为元组 $\langle S, A, P \rangle$ ，其中 S 指连续状态空间，对应于知识图谱中的已知的节点， A 指所有可能的动作集合，对应于该节点的所有出边， P 指转移概率矩阵，而这在知识图谱中是确定的。

DeepPath^[34] 是第一个将强化学习应用到关系路径学习中的工作，提出了一种新的奖励函数来提高准确性、路径多样性和路径效率。它使用了全连接网络训练学习，通过 embedding 的方法在连续空间中对状态进行编码，并以关系空间作为其动作空间。与之相似的，MINERVA^[35] 将路径搜索作为一个序列优化问题，它排除了目标答案实体，并提供了更有能力的推理。与之前使用二元奖励函数的方法不同，Multi-Hop^[36] 提出了一种简单的奖励机制，并且在将部分出边过滤掉，提高了路径搜查效率。M-Walk^[37] 提出一种图游走的机制，利用 RNN 对历史路径编码，与 MCTS 结合，能够更加有效地生成路径。

CPL^[38] 提出协作策略学习来进行路径推理和事实提取，具体流程如图2.7所示。从语料库中提取相关的事实构成新的三元组，查找到新的路径，来丰富知识图谱。同时推理器基于新的三元组知识进行预测，并根据预测结果的正确与否给事实提取器反馈。

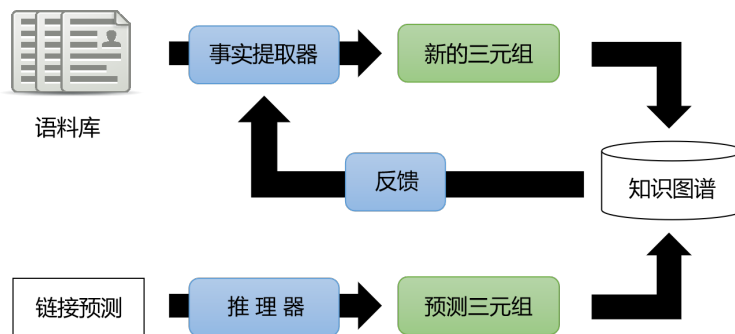


图 2.7: CPL 流程框架

2.3.2.2 路径推理 (path reasoning)

路径推理的目的是在给定路径，却不知道寻路过程的情况下，预测关系。如图2.8所示，路径推理通常的做法是将路径作为特征进行编码，通过多分类器进行未知关系的预测，由 r_1 、 r_2 、 r_3 之间的传递性推理出 e_1 与 e_4 之间的关系 \tilde{r} 。PTransE^[39] 在 TransE 简单高效的基础上，考虑到关系路径的表示。

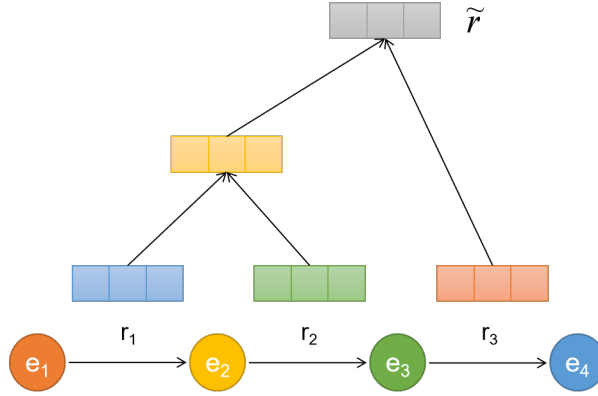


图 2.8: 路径推理

在关于路径推理的研究中，随机游走 (Random Walk) 得到了广泛应用。PRA^[40] 第一个使用随机游走来进行多跳推理，在路径约束下选择关系路径，然后进行最大似然分类。后来，Gardner 等^[41] 在此基础上进行了改进，通过引入文本内容，将向量空间相似性启发式方法引入到随机工作中，缓解了 PRA 的特征稀疏性问题。Neelakantan 等^[42] 将由 PRA 学习到的多跳路径作为输入，利用 RNN 推断其关系，但是它忽略了实体和实体类型的重要性，无法达到理想效果。Das 等^[43] 考虑了关系类型、实体和实体类型，并且引入 attention 机制来推理多跳路径。DIVA^[33] 将路径推理作为概率图模型中的一个推理问题，并从变分推理的角度来解决它。

基于知识图谱图结构的特性，研究者们利用关系路径推理的方法实现知识图谱补全。现有的工作使用的数据中关系的数目较少，当应用到大规模的知识图谱中时，这种方法受关系不完整性的影响较大。

2.3.3 基于规则的方法

知识图谱补全的另一个研究方向是逻辑规则学习，将规则定义为 $head \leftarrow body$ ，其中 $head$ 可以是关于某个实体的一个事实， $body$ 则是事实的集合。举个例子：

$$\langle f, fatherOf, c \rangle \Leftarrow \langle m, motherOf, c \rangle \wedge \langle m, marriedTo, f \rangle$$

由 m 是 c 的母亲、 m 和 f 是夫妻关系两个事实，可以推理出 f 是 c 的父亲。从知识图谱中挖掘出规则能够帮助补全知识图谱，发现可能存在的错误，并且让我们更好地理解知识。逻辑规则可以通过规则挖掘工具提取，类似 WARMER^[44]、AMIE^[45]。

基于规则的推理正确率高，具有可解释性，但是受大规模搜索范围的影响，效率较低；基于 **embedding** 的推理高效，但是它需要丰富的数据。因此，许多研究将逻辑规则引入到 **embedding** 中，各取所长。RUGE^[46] 提出一种迭代模型，同时学习标记三元组、未标记三元组和规则。IterE^[47] 提出迭代学习 **embedding** 和规则，既提高了知识图谱稀疏部分的 **embedding** 质量，又高效地学习了规则。

将神经符号主义 (Neural-Symbolic) 应用到规则推理中，也成为近年来研究者们关注的重点之一。Neural-Symbolic 可以认为是将人工智能中原本对立的连接主义和符号主义结合的一个新兴研究方向，从本质上来说，就是将现代数学中的分析学和代数学结合，分析学擅长处理数值、函数、逼近等问题，代数学擅长处理推演、抽象、结构等问题，这与基于规则的推理很相似。NTP^[48] 通过径向基核，在向量空间中进行微分计算，实现多跳推理。NeuralLP^[49] 使用一种神经网络框架，将 **attention** 机制和辅助记忆结合，使得梯度优化能够被应用于归纳逻辑编程中。基于马尔科夫逻辑网络能够利用具有一阶逻辑的知识，却在复杂图结构中难以推理的特性，pLogicNet^[50] 将高效的 **embedding** 方法与之结合，并且能够处理未知的逻辑规则。ExpressGNN^[51] 在 pLogicNet 的基础上微调，得到了更高效的逻辑推理。

基于规则推理的知识图谱补全方法具有极高的正确率，但是其受大规模搜索范围的影响，效率较低。因此，往往将符号主义和表示学习结合，克服了知识图谱稀疏性的问题，提高了表示学习的质量，使其具有可解释性。

2.3.4 动态知识图谱补全方法

动态知识图谱指将时间信息融入到原有的静态知识图谱中，将三元组拓展为包含时间的四元组。动态知识图谱补全任务主要包括链接预测、关系预测和时间预测。

基于 **embedding** 的方法将时序信息作为特征进行表示学习，与静态模型结合，实现链接预测。Trivedi^[52] 提出了一种新的深度网络体系结构来学习可以随时间动态非线性演化的实体嵌入，利用一个多变量点过程框架来模拟一个事实在连续时间内的发生。RE-NET^[53] 提出了一种新的自回归结构-递归事件网络用于多关系图的时间序列建模，它可以在未来的时间戳上执行顺序的全局结构推理，以预测新的事件。TA^[54] 将关系与时序关系中的数字组成序列 r_{temp} ，DE^[55] 使实体的表示既包含静态特征，又包含动态特征。如表2.2所示，这两种方法都是在 TransE 的基础上对嵌入表示做了拓展。

表 2.2: 基于 **embedding** 的知识图谱补全方法核心公式

TransE	TA-TransE	DE-TransE
$f = - \ h + r - t \ $	$f = - \ h + r_{temp} - t \ $	$f = - \ h_{DE} + r - t_{DE} \ $

基于规则推理的动态知识图谱补全方法,将时序信息作为逻辑推理中的一种数值约束条件,例如, $footballer(a) \wedge bdate(a, y) \wedge NC(y) \leftarrow dead(a), NC(y) = y < 1850$, 其中 a 是实体, y 是时间信息, 这条规则表示 1850 年之前出生的足球运动员都已经过世。Chekol^[56] 将马尔科夫逻辑网络与概率软逻辑相结合, 对不确定的动态知识图谱进行推理补全。RLvLR-Stream^[57] 考虑了时序上的最近路径规则, 并从知识图谱关系中学习规则结构来进行推理。

现实世界中的事件会改变实体的状态, 进而影响其相关的关系和属性, 因此实体状态改变的时间节点值得我们关注, 时间预测任务尤为重要。为了提高时间预测的质量, CTPs^[58] 将时间范围预测问题看作实体状态改变决策问题, 利用文本信息学习状态和状态改变向量。Know-evolve^[59] 利用深度进化知识网络研究实体的变化和变化关系, 多元时间点过程用于模拟事实的发生, 利用循环网络学习非线性时间变化的表示。

动态知识图谱补全作为一个新兴的研究方向, 现有的工作大多是单纯地将时间信息作为一种特征进行学习, 对于时间预测任务的研究方法较少, 在未来的研究工作中仍有很大提升空间。

2.3.5 各类知识补全方法的比较

在知识图谱补全工作上, 这四个研究方向各有优劣, 其代表工作如表2.3所示, 均表现出较高的准确率。基于表示学习的方法通常依赖于三元组的表示学习来捕捉语义特征, 但是它仍停留在独立关系的层面上, 在复杂推理问题中表现较差, 并且缺乏可解释性。基于知识图谱自身的图特性, 关系路径推理得到了广泛的研究, 但是在大型图谱中, 它受限于关系的稀疏性。基于规则的方法具有高准确性和可解释性, 克服了知识图谱的稀疏性问题, 但是在应用于大型知识图谱时, 效率较低。动态知识图谱中时序信息的特征学习与预测也是近年来的研究热点, 目前的相关研究较少, 有很大的提升空间。

表 2.3: 知识图谱补全相关工作总结

分类	方法	会议/期刊	模型
基于表示学习的方法	TransE ^[1]	NIPS,2013	Trans Model
	HolE ^[26]	AAAI,2016	Trans Model+circular correlation
	ComplEx ^[27]	ICML,2016	Bilinear Map+Hermitian product
	ANALOGY ^[30]	ICML,2017	Linear Map+ Analogical Inference
	ConMask ^[32]	AAAI,2018	FCN

路径推理	DeepPath ^[34]	EMNLP,2017	FCN
	MINERVA ^[35]	ICLR,2018	LSTM
	Multi-Hop ^[36]	EMNLP,2018	LSTM+reward shaping
	M-Walk ^[37]	NeurIPS,2018	GRU-RNN+FCN
	CPL ^[38]	EMNLP,2019	LSTM+PCNN-ATT
基于规则的方法	NeuralLP ^[49]	NIPS,2017	TensorLog+ATT+memory
	RUGE ^[46]	AAAI,2018	ComplEx+soft rules
	pLogicNet ^[50]	NeurIPS,2019	MLN+embedding
	IterE ^[47]	WWW,2019	Linear Map+Axiom Induction
动态知识补全	TA ^[54]	EMNLP,2018	LSTM
	DE ^[55]	AAAI,2020	Feature Discretization

2.4 本章小结

本章主要介绍了与本研究相关的工作和理论。首先定义了知识图谱的相关数据模型，而后介绍了网络角色发现的相关理论。最后，阐述了知识图谱补全的含义和相关的主流方法，分析了各种方法的优劣性，给本研究的工作带来一定的启发。接下来将对提出的基于网络角色的知识图谱缺失知识补全方法进行详细的介绍。

第三章 基于网络角色的知识图谱缺失知识补全方法

本章介绍了基于网络角色的知识图谱缺失知识补全方法。首先，在问题描述阶段对基于实体标签的知识补全任务和实体画像做出了详细描述；接着，阐述了模型的整体框架，针对知识图谱中关系稀疏的问题，提出了一种新的实体嵌入方法，并且将画像标签作为补全模型的输入输出，提高了模型的可理解性；最后，对本章做出总结。

3.1 问题描述

在知识图谱补全任务上，现有工作已经取得了一定的成功。传统的基于知识表示学习的方法仅仅对三元组本身建模，将实体和关系映射到向量空间中，尽可能拟合为三角闭合关系。这使得学习得到的嵌入表示包含的语义有限，无法处理复杂的多跳关系。近年来流行的基于关系路径的方法很好地克服了这一困难，但是受知识图谱中关系稀疏性影响较大。在知识图谱中部分实体由于较为罕见有较少的关系可达，能搜索到的路径有限，这使得学习得到的向量表示质量较低，从而影响模型的精度。

图2.1中体现了 DBpedia 数据集中关系稀疏的问题。其中蓝色节点表示实体，蓝色直线表示关系，绿色节点表示属性值，绿色直线代表属性。图中实体 *Megadeth* 和 *Metallica* 没有直接相连的关系路径，但是它们拥有相同的属性值 “*Group_or_band*”，在 *activeYearsStartYear* 属性上 “1983” 与 “1981” 值非常接近。两个实体在属性特征上有一定的相似性，而这种相似无法通过关系路径学习得到。与之类似的，一些实体在结构上也存在结构对等的特性，反映出实体的结构相似性。从属性相似和结构相似上搜索路径，学习得到的实体表示包含了丰富的语义信息，从而提高模型的精确度。

目前，知识图谱补全仍面临着一些挑战：

(1) 关系稀疏问题。现有方法很大程度上依赖于实体间的关系，知识图谱本身的关系稀疏性极大地影响了 *embedding* 的质量，从而影响预测模型的精确度。

(2) 属性预测的低准确度问题。知识图谱中也存在许多属性，有的是数值类型（人的年龄），有的是文本类型（自我评价），还有的是枚举类型（出生月份）。对于数值型属性，精确补全的方法带来了低准确度的问题。

(3) 可理解性问题。现有方法大多基于知识图谱中的三元组，对于预测补全的三元组结果，需要具有相关知识背景的研究者才能够理解，可理解性低。

针对上述问题，提出一种新的基于网络角色的知识图谱缺失知识补全方法，贡献如下：

(1) 提出一种基于网络角色的实体嵌入方法。区别于仅依赖于实体关系的现有工作，本研究引入网络角色发现的方法，考虑了结构相似性和属性相似性，大大提高了预测精

度。

(2) 提出一种属性连续值离散化的方法，用于解决精确补全带来的低准确度问题。将属性划分为两种类型：数值型和字符型。字符型属性，即属性值为字符类型；对于数值型属性，采用连续值离散化，将一个具体的属性数值表示为一个区间。这种模糊但细粒度的属性预测结果基本能够满足用户的需求，且结果能够被用户所理解。

(3) 提出一种基于实体标签的知识补全方法，用于解决可理解性问题。现有的方法在知识图谱补全工作上忽视了可理解性的重要性，本研究将画像标签作为模型的输入输出，帮助用户快速理解预测结果。

给出如下预先定义：

定义 3.1. 知识图谱 (*Knowledge Graph*) : 给定知识图谱 $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{T} \rangle$ ，其中 $\mathcal{V} = E \cup V_a$ 表示 \mathcal{G} 中的节点集合， E 指实体集， V_a 指属性值集合； $\mathcal{E} = \mathcal{E}_r \cup \mathcal{E}_a$ 表示边的集合， \mathcal{E}_r 指实体间的关系集合， \mathcal{E}_a 指实体的属性集合，连接实体和属性值； \mathcal{T} 表示知识图谱中出现的实体类型集合 $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$ 。

定义 3.2. 实体路径 (*Entity Path*) : 基于不同的网络角色发现方法，对知识图谱中的实体节点 $v \in E$ 进行随机游走，生成相应的路径，定义为 $P(v) = \{v \rightarrow v_1 \rightarrow v_2 \rightarrow \dots \rightarrow v_n\}$ 。例如，基于属性相似性的网络角色发现方法生成的实体路径为 $P_A = \bigcup_{v \in E} P_A(v)$ 。

定义 3.3. 标签与标签集 (*Label and Label Set*) : 给定 \mathcal{G} 中的一个类型 t ，标签集 \mathbb{L}_t 表示描述类型 t 的特征的有限标签集。 \mathbb{L}_t 中的每个标签是一个三元组： $l = \langle t, l_{property}, l_{value} \rangle$ 。 $l_{property}$ 指标签属性，它可以是实体的属性 (*attribute*) 或关系 (*relation*)，而 l_{value} 指标签属性值。

定义 3.4. 基于实体标签的知识图谱补全任务 (*Knowledge Graph Completion Based on Entity Labels*) : 给定知识图谱 \mathcal{G} 下的实体画像标签集合 $\mathcal{P} = \langle E, \mathbb{L}, \mathcal{B} \rangle$ ，其中实体集 $E = \{e_1, e_2, \dots, e_m\}$ 、标签集 $\mathbb{L} = \{l_1, l_2, \dots, l_n\}$ 以及二元组集 $\mathcal{B} = \{(e_i, l_j) \mid e_i \in E, l_j \in \mathbb{L}\}$ ，知识图谱补全即预测出实体缺失的标签 $\mathcal{B}' = \{(e_i, l_j) \mid (e_i, l_j) \notin \mathcal{B}\}$ 。

3.2 模型框架

本研究采用一种基于网络角色的方法来解决知识图谱中知识缺失的问题，整体框架如图3.1所示。

首先，对输入的知识图谱进行预处理，对于数值型属性进行连续值离散化，将具体的值表示为区间的形式。然后采用网络角色发现的方法，基于不同的相似性，挖掘不同的实体特征，构造路径。本研究采用三种不同角度的角色发现方法，分别是基于同质性、属性相似性和结构相似性的角色发现方法。根据不同的角色发现方法，对实体节点进行随机游走，生成对应的路径。然后，对生成的不同路径序列按一定比例混合，得到包含丰富语义的实体嵌入表示，解决关系稀疏问题。

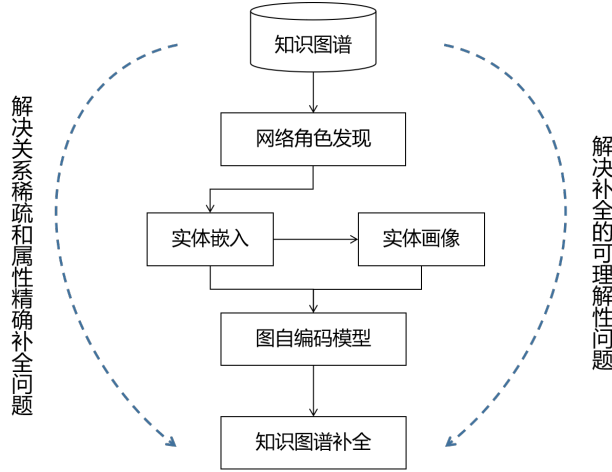


图 3.1: 基于网络角色的知识图谱缺失知识补全方法框架图

实体角色发现方法得到的嵌入表示可作为图自编码模型中实体节点的初始化表示，亦可用于生成实体画像，概括性描述实体信息。将画像标签结果作为图自编码模型的输入输出，极大地提高了可理解性。迭代训练网络，预测缺失的关系标签和属性标签，最终实现知识图谱的补全。

3.3 属性预处理

知识图谱中，实体包含关系信息和属性信息，对应表示为关系三元组和属性三元组。关系三元组表现为 $\langle e_h, r, e_t \rangle$ 的形式，表示两个实体之间的联系。属性三元组表现为 $\langle e, a, v_a \rangle$ 的形式，表示实体自身的属性信息。属性值又可以分为数值类型和字符类型。字符类型的属性值以文本的形式来描述实体，而数值类型的属性值以整数或浮点数的形式来体现实体的信息。其中，数值型的属性更能够体现实体间的相似性，例如年龄“18”的实体，与年龄“46”的实体相比，和年龄“20”的实体更相似。

将连续数值离散化的方法，是机器学习中常见的预处理数据的方法之一。离散化的特征相对于连续型的特征，更加接近知识层面的表达，易于理解。在网络角色发现中，离散化的方法能够更好地找到实体在属性层面上的相似性。在知识补全任务中，离散化的特征简化了预测目标，由预测具体的值变为预测一个区间。这种模糊但细粒度的属性预测结果基本能够满足用户的需求，且结果能够被用户所理解。

本研究设置了一些简单的离散化规则来找到一些特定值的合适区间，例如使用五年的周期作为各种年份的间隔。对于其他类型的数值，等宽和等频^[60]是比较简单的离散化方法，也是常用的离散化方法。然而，这些方法的不足之处在于没有考虑数据的分布。因此，本研究采用了一种基于局部密度的离散化算法，该算法来源于^[61]。其主要思想是找出属性值的密度区间，确保区间中间的密度高，边界附近的密度低。属性值排序后，密度值呈现多峰现象，密度分布的每个峰值表示两个区间之间的边界。

表3.1给出了不同离散化方法的示例。对于乐队实体 *Megadeth* 的活跃年份，选择以

表 3.1: 属性连续值离散化示例

离散化方法	示例
特定值人工设定	$\langle \text{Megadeth}, \text{ActiveYearsStartYear}, 1983.01.01 \rangle$
	\Downarrow $\langle \text{Megadeth}, \text{ActiveYearsStartYear}, [1980, 1985) \rangle$
基于密度的方法	$\langle \text{Acadia_University}, \text{numberOfStudents}, 4782 \rangle$
	\Downarrow $\langle \text{Acadia_University}, \text{numberOfStudents}, [3302.0, 5630.0) \rangle$

5 年为间隔进行划分。而对于学校 *Acadia_University* 的学生数量，在知识图谱中学校实体的学生数量分布大多在千人以上，选择基于密度的划分方法更为合适。

3.4 实体角色发现

为了解决关系稀疏问题，本研究利用实体角色发现来提高实体嵌入的质量。受到基于路径的嵌入模型的启发，如 DeepWalk^[62] 和 Node2Vec^[63]，它们使用 skip-gram^[64] 训练节点的游走序列，生成节点的向量。然而，DeepWalk 只考虑同质性，Node2Vec 考虑了结构相似性，但是对于给定的路径，进行编码的是基于哪种相似性仍不明确。本研究考虑了不同层次的实体相似性：同质性（**H**omophily）、属性相似性（**A**ttributive similarity）和结构相似性（**S**tructural similarity）。

本研究将知识图谱作为输入，对每个实体进行路径查找操作，得到从该实体开始的一组路径。基于不同层次的实体相似性策略，会得到三种类型的路径：(1) H 策略用来找到代表同质性的 H 路径；(2) A 策略用来找到代表属性相似的 A 路径；(3) S 策略用来找到代表结构相似的 S 路径。针对不同的实体相似性，利用随机游走生成不同的实体路径，并进行相应的路径混合，进而更全面的刻画实体的特征。最终对实体路径建模，生成实体嵌入。

3.4.1 基于同质性的角色发现方法

同质性（**H**omophily）指实体之间由直接链接关系相连，图3.2给出了 DBpedia 中实体同质性的示例。圆形节点代表实体，根据类型的不同，被标注为不同的颜色，例如 *Megadeth* 和 *Metallica* 都属于“**B**and”类型，而 *Dave_Mustaine* 和 *Mike_Muir* 属于“**M**usicalArtist”类型。方形节点代表属性值，绿色直线代表属性名。蓝色直线表示实体间的关系，表明了相邻实体节点之间存在着同质性，例如 *Megadeth* 和 *Metallica* 由关系

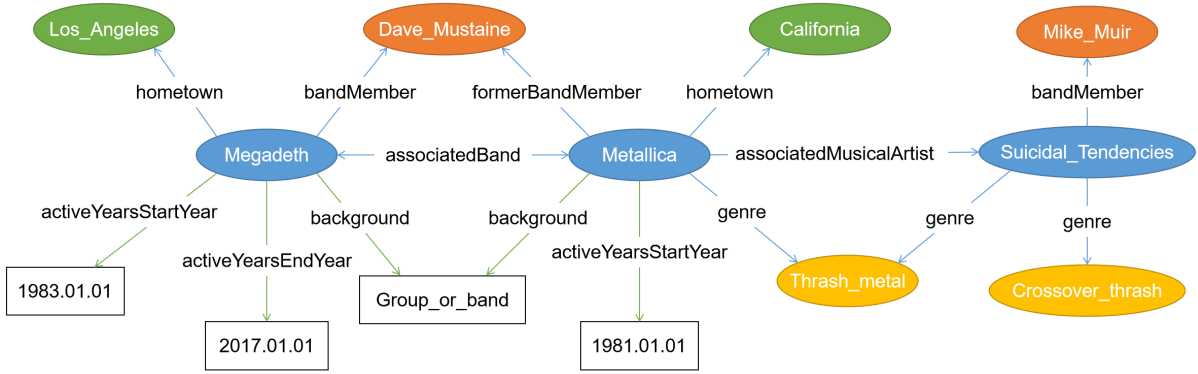


图 3.2: 实体同质性示例

associatedBand 相连, 存在着同质性。基于实体同质性的 H 策略, 在知识图谱中利用深度优先搜索 (Deep First Search, DFS) 生成多条 H 路径。在示例中, $\{Megadeth, Metallica, Suicidal_Tendencies, Mike_Muir\}$ 就是一条以节点 *Megadeth* 为起点, 基于同质性策略找到的游走序列长度为 4 的 H 路径。

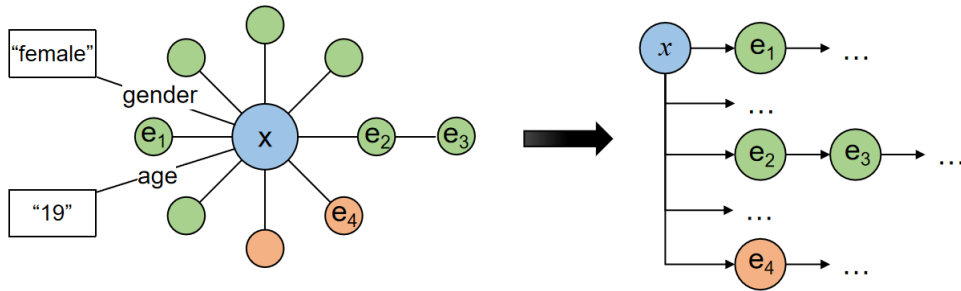


图 3.3: 基于同质性策略生成的实体路径

图3.3展示了基于 H 策略生成的实体 x 的 H 路径。图的左半部分表示输入的知识图谱部分, 其中圆形节点是实体, 白色矩形是属性值信息, 边是实体间的关系或连接实体和属性值的属性名。根据实体类型的不同, 实体被标注为不同的颜色。图的右半部分展示了从实体 x 出发, 依据实体间直接相连的关系, 通过深度优先搜索生成的 H 路径。选择实体节点 x 作为当前开始节点, 随机选取 x 的直接邻居 e_1 作为后继节点, 再以 e_1 为当前开始节点随机寻找其后继节点, 直到路径深度为 l 。算法3.1描述了基于同质性策略生成 H 路径 P^H 的过程。

3.4.2 基于属性相似性的角色发现方法

属性相似性指同类型下的实体在属性值上形成的相似性, 根据此来构建实体的隐含关联路径, 路径中相邻的节点便是同类型下属性相似的实体。图3.4给出了 DBpedia 中实体属性相似性的示例, *Barnard_College* 和 *Acadia_University* 为 “University” 类型下

算法 3.1 基于同质性策略生成的实体路径**Input:** $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ n : number of walks l : length of walks**Output:** P^H : set of node sequences based on Homophily

```

1:  $P^H = \emptyset$ 
2: for each  $e \in \mathcal{V}$  do
3:    $n' = n$ 
4:   while  $n' > 0$  do
5:      $walk = \emptyset$ 
6:     add  $e$  to  $walk$ 
7:      $current\_node = e$ 
8:      $l' = l - 1$ 
9:     while  $l' > 0$  do
10:       $neighbors = current\_node.getNeighbors()$ 
11:       $successor\_node = random.choice(neighbors)$ 
12:      add  $successor\_node$  to  $walk$ 
13:       $current\_node = successor\_node$ 
14:       $l' --$ 
15:    end while
16:     $n' --$ 
17:     $P^H.append(walk)$ 
18:  end while
19: end for
20: return  $P^H$ 

```

的两个实体。绿色方框代表属性值，绿色直线代表属性名。数值类型的属性可作为区分实体的重要指标，具有自身的语义信息，例如，在 *numberOfUndergraduateStudents* 属性下“2360”与“11549”更类似于“3753”。而字符类型的属性则不能基于距离去度量这种实体属性的区别。

图3.5给出了基于属性相似性的 A 策略寻找实体的 A 路径的示例。从 x 开始，基于属性相似性的 A 策略尝试查找知识图谱中与 x 属性值最相似的同类型的后继实体。在图3.5中，上半部分为输入的知识图谱，下半部分为在属性空间找到的 A 路径，实体 x 与 y 、 z 都是相同的类型。可以看出， z 与 y 相比，和 x 更相似，因为 z 的性别与 x 相同，并且和 y 相比， z 的年龄更接近 x 。所以，基于属性相似性的策略，选择 z 作为从 x 出发的路径的后继节点。

然而，在一个大规模的知识图谱中，将最近的邻居节点作为下一跳在计算上是不可

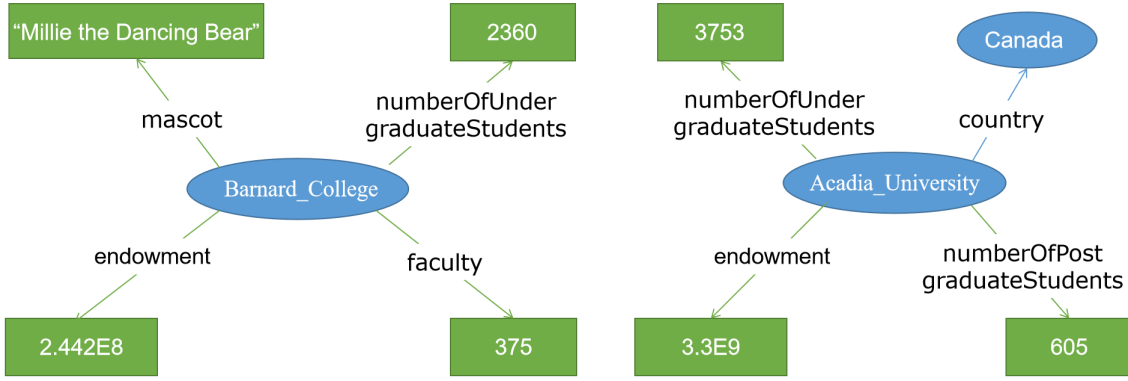


图 3.4: 实体属性相似性示例

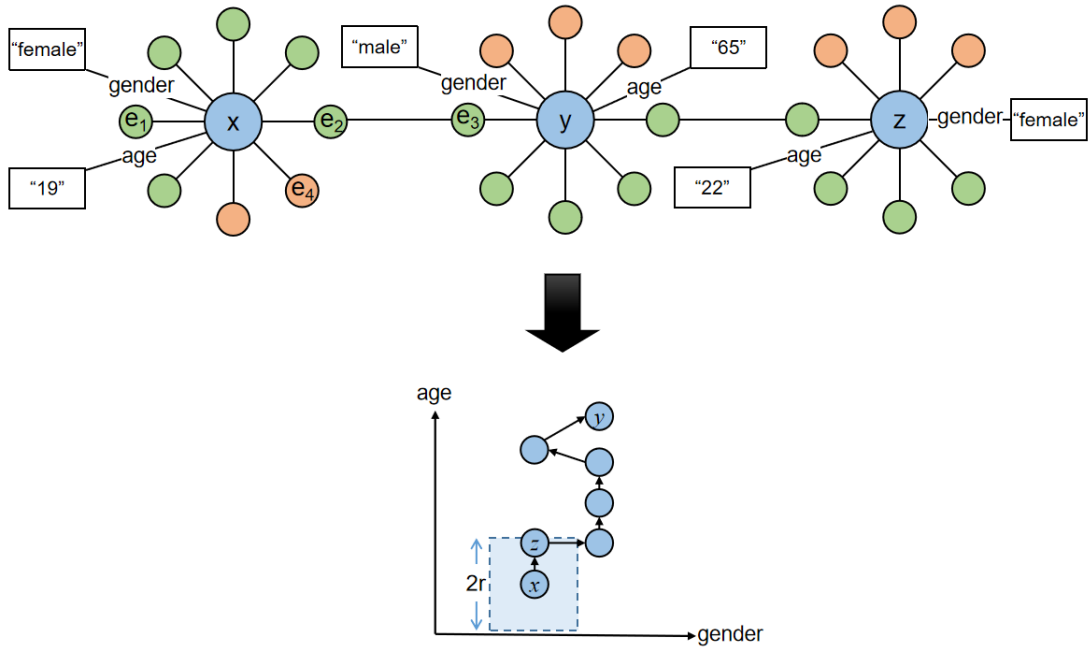


图 3.5: 基于属性相似性策略生成的实体路径

行的，更不用说所有实体的多跳路径集合。本研究首先将相同类型的实体嵌入到一个属性空间。对于每个类型 t ，其属性空间有 $|\mathcal{A}_t|$ 维， $\mathcal{A}_t \subseteq \mathcal{A}$ 表示 t 的属性数。在示例中， x 、 y 和 z 被嵌入到一个 2D 空间中 (年龄和性别)。对每个维度进行归一化处理后，多维空间中以 x 为中心的超立方体刻画出 x 的邻居。我们将超立方体的边长定义为 $2r$ ，落在区域内的实体被认为是 x 的近邻。在 A 路径中，随机选择其中一个邻居邻作为 x 的后继实体。继续迭代，直到为 x 生成一个固定长度的 A 路径。根据属性空间中相邻实体间的平均间隔，估计 $2r$ 的初始设置。超参可以将超立方体放大或缩小，使超立方体中邻居的数量接近原始知识图谱中直接邻居的平均数量。

算法 3.2 描述了基于属性相似性策略生成 A 路径 P^A 的过程。其中， A 指 t 类型下所有实体的属性构成的特征矩阵， A_{ij} 表示实体 i 在属性 j 上的属性值。对矩阵 A 归一化后，在属性空间下进行实体的随机游走。

算法 3.2 基于属性相似性策略生成的实体路径**Input:** $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{T} \rangle$ A : attributes of entities r : size of neighbor field n : number of walks l : length of walks**Output:** P^A : set of node sequences based on Attributive similarity

```

1:  $P^A = \emptyset$ 
2: for each  $t \in \mathcal{T}$  do
3:    $n' = n$ 
4:   for each  $e \in \mathcal{V}$  do
5:     while  $n' > 0$  do
6:        $walk = \emptyset$ 
7:       add  $e$  to  $walk$ 
8:        $current\_node = e$ 
9:        $l' = l - 1$ 
10:      while  $l' > 0$  do
11:         $neighbors = current\_node.getAttNeighbors(A, r)$ 
12:         $successor\_node = random.choice(neighbors)$ 
13:        add  $successor\_node$  to  $walk$ 
14:         $current\_node = successor\_node$ 
15:         $l' --$ 
16:      end while
17:       $n' --$ 
18:       $P^A.append(walk)$ 
19:    end while
20:  end for
21: end for
22: return  $P^A$ .

```

3.4.3 基于结构相似性的角色发现方法

结构相似性通常反映在实体的局部结构中。如图3.6所示，图中的实体节点根据其实体类型的不同被标注为不同的颜色。实体 *Metallica* 和 *Suicidal_Tendencies* 在 *bandMember*、*genre* 和 *hometown* 等属性上是比较相似的，所连接的实体类型和类型个数十分相近，在知识图谱中扮演的角色相似。

图3.7给出了基于结构相似性的 S 策略寻找实体的 S 路径的示例，上半部分为输入

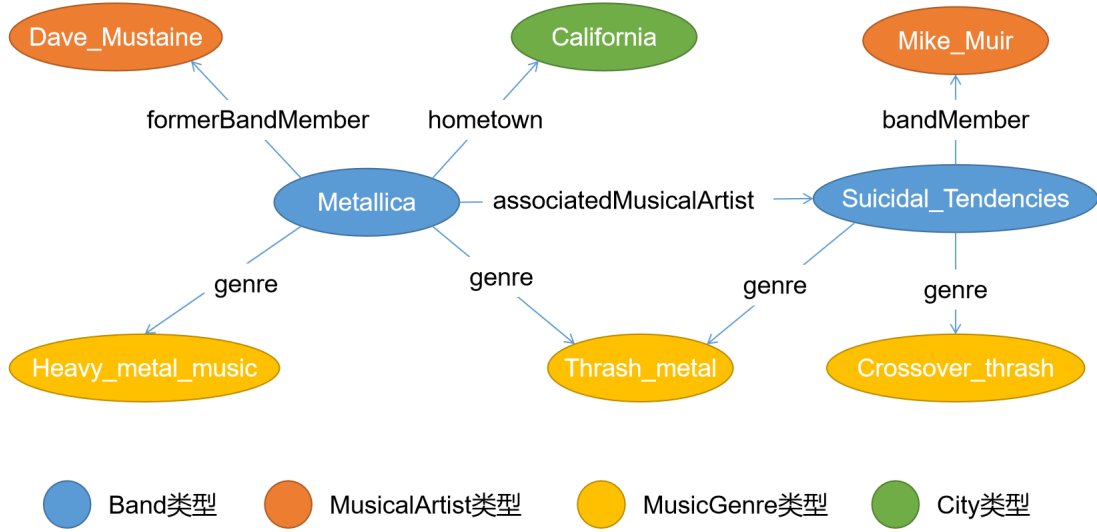


图 3.6: 实体结构相似性示例

的知识图谱，下半部分为在结构空间中找到 S 路径。与 A 策略相似， S 策略通过在结构空间中实体的嵌入来查找 S 路径。给定一个类型 t ，它的结构空间有 $|\mathcal{T}|$ 维，其中 $|\mathcal{T}|$ 是知识图谱中的类型数。某个维度 t' 中的实体的组成部分是其类型为 t' 的直接邻居的数量。图3.7的下半部分中，横轴表示邻居数量，其类型被标注为橙色，纵轴表示其类型被标注为绿色的邻居数量。 x 、 y 、 z 的坐标分别是 $(2, 6)$ 、 $(3, 5)$ 、 $(3, 4)$ ，则 y 与 z 相比，和 x 更相似。接下来的路径查找步骤与 A 策略中相似。

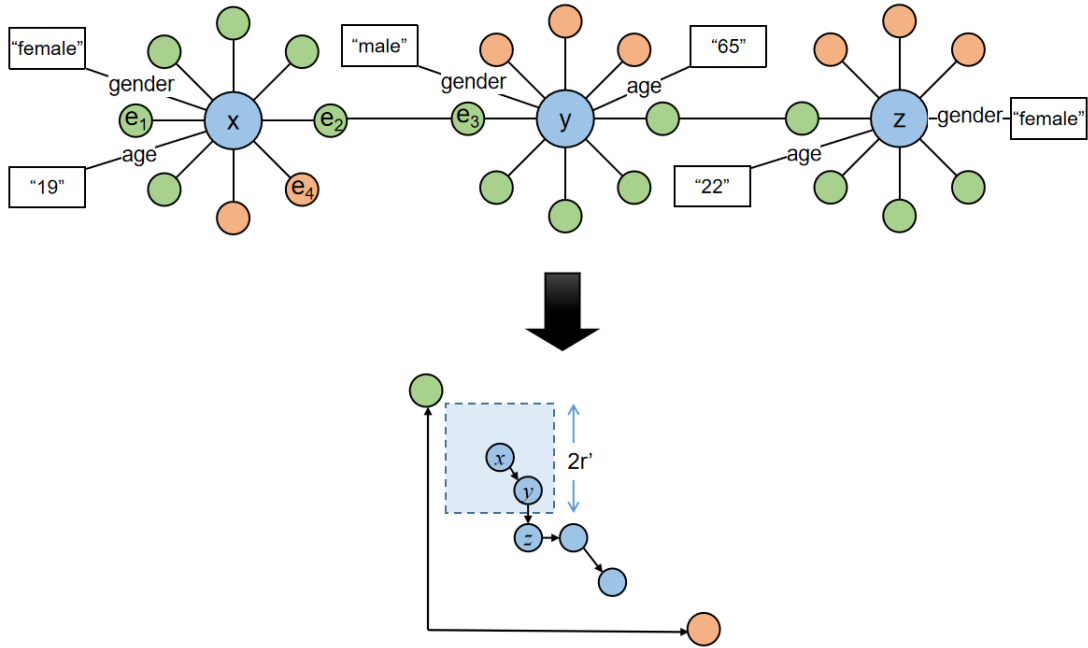


图 3.7: 基于结构相似性策略生成的实体路径

算法3.3描述了基于结构相似性策略生成 S 路径 P^S 的过程。其中， S 指 t 类型下实

体的邻居类型构成的特征矩阵, S_{ij} 表示实体 i 在类型 j 上拥有的直接邻居数量。对矩阵 S 归一化后, 将实体映射到结构空间中, 进行随机游走。

算法 3.3 基于结构相似性策略生成的实体路径

Input: $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{T} \rangle$

S : types of neighbors

r' : size of neighbor field

n : number of walks

l : length of walks

Output: P^S : set of node sequences based on Structural similarity

```

1:  $P^S = \emptyset$ 
2: for each  $t \in \mathcal{T}$  do
3:    $n' = n$ 
4:   for each  $e \in \mathcal{V}$  do
5:     while  $n' > 0$  do
6:        $walk = \emptyset$ 
7:       add  $e$  to  $walk$ 
8:        $current\_node = e$ 
9:        $l' = l - 1$ 
10:      while  $l' > 0$  do
11:         $neighbors = current\_node.getstruNeighbors(S, r')$ 
12:         $successor\_node = random.choice(neighbors)$ 
13:        add  $successor\_node$  to  $walk$ 
14:         $current\_node = successor\_node$ 
15:         $l' --$ 
16:      end while
17:       $n' --$ 
18:       $P^S.append(walk)$ 
19:    end while
20:  end for
21: end for
22: return  $P^S$ .

```

3.4.4 路径混合

基于三种不同的实体相似性, 得到了实体的 3 种随机游走路径集合: P^H 、 P^A 、 P^S , 分别代表 H 路径、A 路径和 S 路径, 刻画了实体不同方面的特征, 这些路径将采样到

最终特征集中。如公式3.1所示， P 是最终的特征集， λ_H 、 λ_A 、 λ_S 为路径采样的比例参数。一种均匀的路径采样策略为 $\lambda_H : \lambda_A : \lambda_S = 1 : 1 : 1$ 。偏置抽样是一种不平衡加权方案的策略，当 $\lambda_H : \lambda_A : \lambda_S = 1 : 0 : 0$ ，就等价于 DeepWalk，仅关注于实体间的同质性。基于不同的应用场景，对于三种实体相似性的侧重也将相应的调整。在路径混合之后，我们遵循 DeepWalk 的 skip-gram 学习过程，这里省略了细节。

$$P = \lambda_H P^H + \lambda_A P^A + \lambda_S P^S \quad (3.1)$$

基于网络角色的实体嵌入方法是可扩展的，除了 P^H 、 P^A 和 P^S 外，也可以依据实体的其他特征构造出更多的实体路径，进一步的丰富实体的采样路径。

3.5 实体画像

近年来，知识图谱呈现动态增长的趋势，在各个领域中都有广泛的应用。然而，知识图谱的容量和结构复杂性大大降低了识别和比较实体的效率，在实际应用中遇到了阻碍。实体画像方法的思想是将实体与其他实体进行比较，帮助用户了解其唯一性和独特性。Li 等人^[65] 引入了实体画像的方法来描述真实世界的实体，这些实体能够以不同的方式用相同的信息来描述。Rybak 等人^[66] 提出了利用实体画像标签来识别每个专家的专长，他们认为实体标签具有动态特性，这意味着实体画像结果会随时间而改变。

实体画像工作的目标是自动挖掘出实体的标签特征，能够体现实体相对于其他实体的唯一性的“全局”信息——区分性，而不再仅限于实体本身的特征信息。在图3.8中给出了实体的画像标签示例。图中的实体是定义在 DBpedia 中的一个乐队实体 *Beastie Boys*。该实体的 top-5 标签从知识图谱中提取，并用绿色标注出标签的区分度，其中“ $\neq 80\%$ ”表示该实体在该特征上与其他 80% 的乐队不同；“ $>60\%$ ”或“ $<95\%$ ”表示该实体在该特征上与其他乐队相比，具有较大或较小的值。

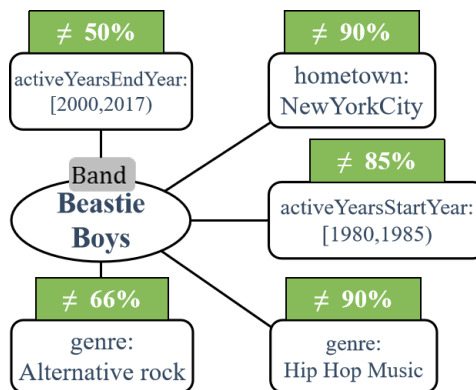


图 3.8: 实体画像示例图

上一章中提到的基于网络角色发现的实体嵌入方法，能够用来度量实体的相似性，将其应用到实体画像任务中，来描述知识图谱中的实体，可以提取出能够体现实体独特

特征的结构化标签。实体画像方法流程如图3.9。首先，给定一个知识图谱作为输入，所有可能的标签都会被自动枚举到一个标签池中。之后，每个标签将由上一节中提到的基于网络角色发现的实体嵌入方法来度量其区分性，衡量正例和负例之间的差异性。只有具有区分性标签被留在标签集中，我们使用重排序来减少标签空间中的冗余。最后，给每个实体贴上对应的标签，生成实体画像。

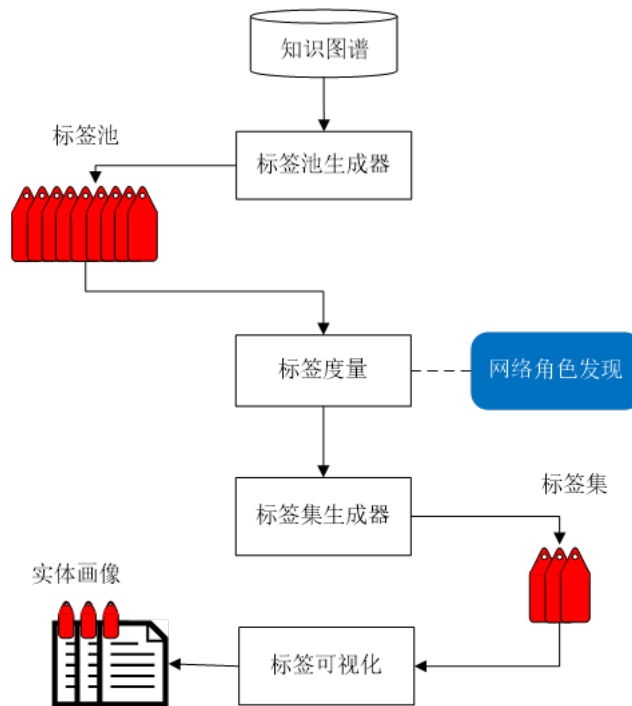


图 3.9: 实体画像方法流程图

3.5.1 构建标签池

实体画像的核心思想是为每个实体类型构造一个标签集，标签能够体现实体间的区分性。在知识图谱中，实体的特征在结构上是异构的，一部分特征是属性型，来描述实体的属性值，而另一部分特征是关系型，表示了实体间的连接。本研究根据特征结构的不同，对标签进行分类，如表3.2所示。

AIL(Attributive-interval labels) 和 AVL(Attributive-value labels) 是属性型标签，AIL 表示属性区间标签，数值型的属性经过预处理后，具体的某一属性值落在了某一个区间内。例如，标签 $\langle Film, rating, [8.0, 9.0] \rangle$ 描述了一部高分电影，通常这是值得观看的。AVL 表示属性值标签，针对的是字符型属性，其中标签的值不是一个区间，而是具体的，比如 $\langle People, gender, "female" \rangle$ 。而 RAL(Relational-attributive labels) 和 REL(Relational-entity labels) 是关系型标签，RAL 指关系-属性标签，它表示某类实体与其他具有具体属性的实体之间存在联系。例如， $\langle Director, directorOf, \langle Film, rating, [8.0, 9.0] \rangle \rangle$ 描述了拍了高分电影的导演。REL 指某类实体与一个具体的实体有联系，例如 iPhone、

表 3.2: 标签的分类和相关示例

标签	示例
AIL	$\langle \text{Film}, \text{rating}, [8.0, 9.0] \rangle$
AVL	$\langle \text{People}, \text{gender}, \text{"female"} \rangle$
REL	$\langle \text{Product}, \text{producedBy}, \text{Apple} \rangle$
RAL	$\langle \text{Director}, \text{directorOf}, \langle \text{Film}, \text{rating}, [8.0, 9.0] \rangle \rangle$

iPad 和其他苹果产品都是带有这种标签的实体: $\langle \text{Product}, \text{producedBy}, \text{Apple} \rangle$ 。

在没有先验知识的情况下, 通过自动化生成标签的方式, 从知识图谱中暴力枚举出所有标签。通过枚举所有属性和属性值的组合, 或关系和实体的组合, 直接生成候选标签。AIL 标签将属性的连续值生成为包含该值的一个更广泛的区间, 使其更加具有代表性。例如, 给定一个三元组 $\langle \text{ForrestGump}, \text{rating}, 8.3 \rangle$, 仅仅是简单地生成候选标签 $\langle \text{Film}, \text{rating}, 8.3 \rangle$ 是毫无意义的, 因为这个标签过于特殊, 几乎无法代表其他电影的特征。因此, $\langle \text{Film}, \text{rating}, [8.0, 9.0] \rangle$ 优于 $\langle \text{Film}, \text{rating}, 8.3 \rangle$, 前者更能代表高分的电影。

3.5.2 标签度量

在生成标签池之后, 所有的标签需要进一步的评估, 确保实体最终的标签结果具有区分性, 在正例和负例之间有明显的界限。将正例定义为符合该标签的实体, 负例为不符合该标签的实体。

一个好的标签能够将一组相似的实体和其他不同的实体区分开。标签度量的本质是计算实体的相似度, 对于一个区分性的标签, 正例之间是相似的, 而跟负例之间是不同的。研究者们提出了许多方法在图中计算相似度, 比如 Katz 相似度^[67]、SimRank^[68] 和 P-Rank^[69]。他们的主要思想是若两个实体的邻居是相似的, 则它们结构相似。因此, 计算两个实体之间的相似度就变成沿着邻域迭代传播相似度的问题。然而, 这些方法存在两个不可避免的问题: (1) 对于大规模的知识图谱, 基于路径的相似性度量在计算上是不可行的; (2) 基于路径的方法是基于同质性假设的, 即实体会与其相似实体相连。但在关系稀疏的大规模知识图谱中, 这两个问题无法得到很好地解决。

基于网络角色发现的实体嵌入方法从多个角度来刻画实体间的相似性, 将其应用到实体画像中, 能够很好地解决以上问题。使用公式 3.2 来度量标签, 将 $d(l)$ 定义为 l 的区分度, 对于类型 t 的一个标签 l , E_t^l 表示正例集, 而 $E_t^{\bar{l}}$ 表示负例集, $\text{sim}(i, j)$ 表示实体 i 和 j 之间的相似性, 基于网络角色发现的实体嵌入结果用于度量实体间的相似性, 计算实体嵌入的 \cos 距离。第一项是正例集合之间平均相似度, 第二项是正负例集合之间

平均相似度，该差值越大，则表明标签 l 质量越好。

$$d(l) = \frac{\sum_{i,j \in E_t^l} \text{sim}_{i,j}(i,j)}{|E_t^l|^2} - \frac{\sum_{i \in E_t^l, j \in E_t^{\bar{l}}} \text{sim}_{i,j}(i,j)}{|E_t^l| |E_t^{\bar{l}}|} \quad (3.2)$$

3.5.3 生成标签集

在将标签添加到最终的标签集之前，需要进一步的评估。一个好的标签应该满足两个要求：（1）给标签集带来的冗余少；（2）提高了标签集的完整性。第一个要求是优先处理与已有标签不同的标签，第二个要求倾向于与已有标签互补的标签。

本研究提出一种重排序方法，来生成最终的标签集。如公式3.3所示， $f(l_i)$ 指候选标签 l_i 的得分函数。给定一个在候选标签集 \mathbb{L}_c^t 中但不在最终标签集 \mathbb{L}^t 中的候选标签 l_i ， $d(l_i)$ 是 l_i 的区分度得分， $\text{reward}(l_i, \mathbb{L}^t)$ 是 l_i 对知识图谱中的正例实体覆盖率的提高所做的潜在贡献， penalty 是 l_i 对 \mathbb{L}^t 中冗余增加的潜在影响。公式3.4和3.5中定义了 reward 和 penalty ， δ 是偏置因子， E_t 是类型 t 的实体集合， $E_t^{l_i}$ 代表标签 l_i 的正例实体。最后，对候选标签进行排序，并将其逐个加入到标签集中。

$$f(l_i) = \arg \max_{l_i \in \mathbb{L}_c^t} [d(l_i) + \delta \cdot \text{reward}(l_i, \mathbb{L}^t) - (1 - \delta) \text{penalty}(l_i, \mathbb{L}^t)] \quad (3.3)$$

$$\text{reward}(l_i, \mathbb{L}^t) = \frac{|\bigcap_{l_j \in (\mathbb{L}^t \cup \{l_i\})} E_t^{l_j}|}{|E_t|} \quad (3.4)$$

$$\text{penalty}(l_i, \mathbb{L}^t) = \frac{\sum_{l_j \in \mathbb{L}^t} |E_t^{l_i} \cap E_t^{l_j}|}{|\mathbb{L}^t| \times |E_t|} \quad (3.5)$$

最后，将遍历标签集，以查找某个实体是否与某些标签匹配，为实体贴上相应的画像标签描述。

3.6 基于实体标签的知识补全方法

现有的知识补全模型或从表示学习角度对实体和关系建模，或从大量数据中归纳规则进行推理，但它们都是基于知识图谱中的三元组结构。以三元组作为输入输出的知识补全模型，需要用户具有知识图谱的背景，从冗长而复杂的 RDF 描述中提取到有效的信息，才能迅速理解预测结果。而在某些情况下，用户需要快速理解补全的结果和影响补全过程的因素。因此，补全方法应当对知识补全的过程提供一定的解释。与以往的知识图谱补全工作不同，本研究采用的是基于实体标签的知识图谱补全方法，输入的不再是原知识图谱，而是实体的标签集。实体画像的标签结果本身就包含了能够体现实体相对于其他实体的唯一性的特征信息，以简洁精炼的形式使用户能够快速理解实体。如图1.1给出的示例，用户通过少量的标签就可以快速理解预测结果。本研究提出的基于网络角色的知识补全方法（**Network Role based Knowledge Completion, NRKC**）使用基

于实体标签的图自编码模型，知识补全中原本的链接预测任务也转变为对于实体标签的预测。

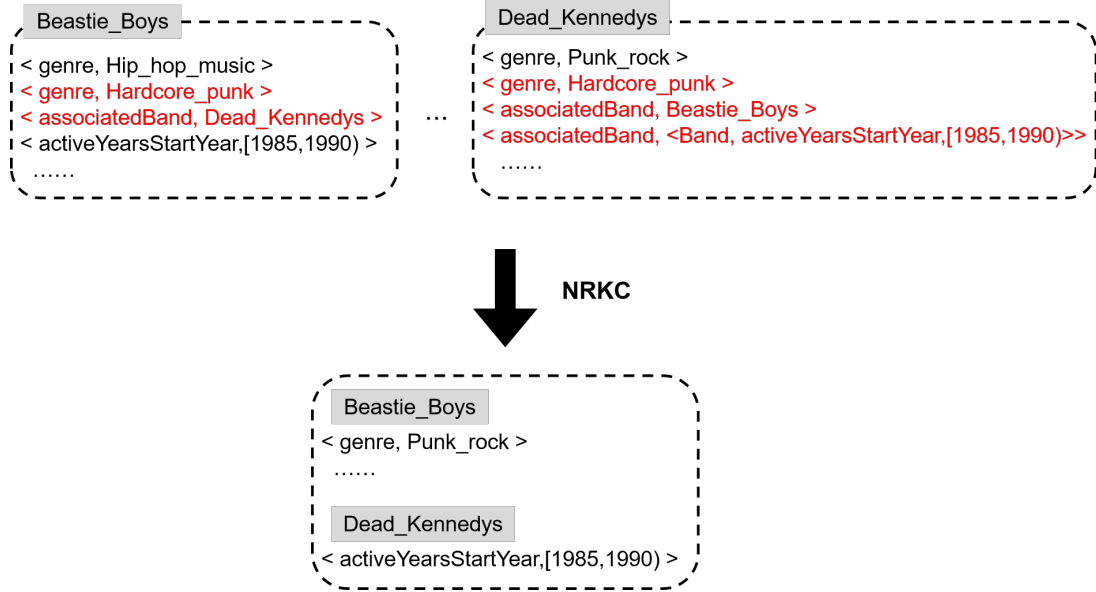


图 3.10: 基于实体标签的端到端模型示例

基于实体标签的知识图谱补全方法采用以实体标签作为输入，以预测的缺失标签作为输出的端到端模型，大大减少了工程的复杂度。端到端模型将原本需要多步骤或多模块解决的任务使用单个模型来建模解决，是深度学习中常见的模型之一。图3.10给出了模型输入输出的示例，输入的是 **Band** 类型下所有实体的标签集，通过本研究的端到端模型 NRKC 的训练学习，输出的是预测出的新标签。图中实体 *Beastie_Boys* 与 *Dead_Kennedys* 彼此关联，音乐类型都为 *Hardcore_punk*，具有较高的相似度。仅由 *Dead_Kennedys* 的标签 “< associatedBand, < Band, activeYearsStartYear, [1985, 1990) >>”，就可以明显推测出缺失标签 “< activeYearsStartYear, [1985, 1990) >”。传统方法使用多步骤或多模块来解决一个复杂任务的时候，一个明显的弊端就是由于各个部分的训练目标不一致，导致某个部分的目标函数可能与系统的宏观目标出现偏差，使得最终训练出来的系统很难达到最优的性能。同样由于误差的累积，前一个步骤产生的偏差可能影响后续步骤的计算。而端到端模型仅使用一个模型、一个目标函数，规避了多模块模型固有的缺陷。

传统的深度学习端到端模型无法在图中获取节点合适的嵌入表示，而合适的节点嵌入表示能够很好地应用于下游任务，如链接预测等等。Thomas^[70] 提出用图自编码器（**Graph Auto-Encoder**）来学习图中的节点表示，包括 **encoder** 和 **decoder** 两个部分。**encoder** 阶段中，训练模型学习分布，得到节点的 **embedding**。图卷积神经网络能够很好地捕捉图中节点的特征信息，常被用于学习节点的特征表示，作为 **encoder** 阶段的模型。**decoder** 阶段采用内积（**inner-product**）或计算图中两点之间存在的概率来重构原始的图，最终选择合适的损失函数使得重构的图与原始的图结构尽可能的相似。

在图自编码模型中, **encoder** 把图中的每个节点 $v_i \in \mathcal{V}$ 映射到一个真值向量 $\vec{v}_i \in \mathbb{R}^d$, **decoder** 则是根据节点的表示重建图的边, 即通过函数 $f(\cdot)$ 来对实体标签对进行评分。在对图中的节点和边建模方面, 已有许多表示学习方法, 例如 TransE^[1] 将节点和边进行线性映射, NTN^[71] 融合了线性表示和双线性表示来对图建模。Yang^[72] 通过实验发现, 在一般化的表示框架下, 选择不同的节点以及边的表示方法, 在链接预测任务上简单的双线性模型取得了最好的效果。

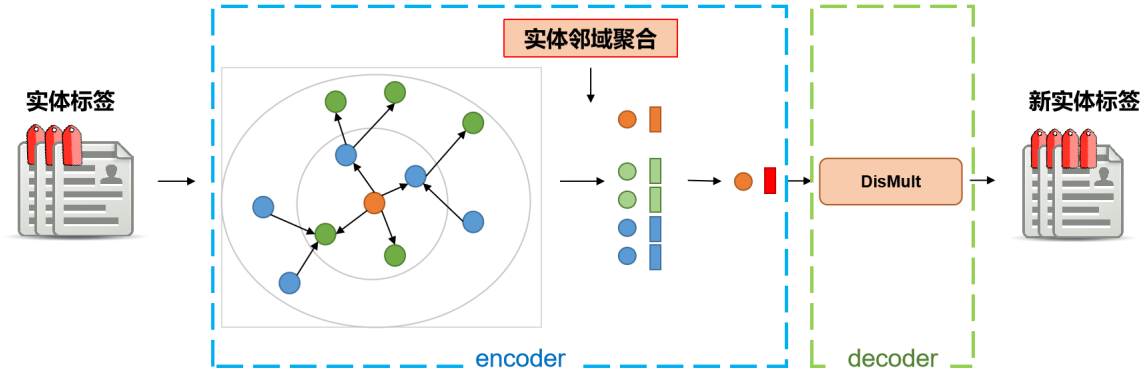


图 3.11: 基于实体标签的知识图谱补全方法框架图

受此启发, 本研究采用一种图自编码模型 NRKC 来解决知识图谱中知识缺失的问题, 选择图卷积神经网络作为 **encoder**, 以 DisMult 作为 **decoder**, 整体框架如图3.11所示。将实体标签作为图卷积网络的输入, 对于一个独立的实体节点 (橘色), 它的邻居节点包括属性标签节点 (绿色) 和关系标签节点 (蓝色), 将邻居节点表示与自身表示结合, 生成新的节点表示。通过图卷积神经网络得到节点的新表示后, 由于双线性模型在链接预测任务中优秀的表现, 本研究使用一个简单有效的双线性模型 DisMult^[72] 作为 **decoder**, 对缺失的标签进行预测。

3.6.1 图卷积神经网络

神经网络作为人工智能领域主流方法之一, 在语音识别、图像处理等应用上展现出了优越的性能。例如, 卷积神经网络利用其平移不变性能够很好地捕捉图像的特征。然而真实世界中, 许多重要的数据集都是以图或者网络的形式存在的, 比如社交网络, 知识图谱, 蛋白质相互作用网, 世界贸易网等等。对于这些不规则的数据对象, 普通的卷积网络难以选取固定的卷积核来适应整个图的不规则性, 效果不尽如人意。在过去几年里, 研究者们尝试将图神经网络应用在任意图结构数据中^[73;74], 并且在一些领域已经达到了非常好的效果。由于图卷积网络能够很好地处理数据多样的拓扑结构, 在知识图谱领域也得到了广泛应用。

本研究将图卷积神经网络模型作为 **encoder** 来捕捉实体标签中的结构相似性。在图卷积网络中, 对于一个独立的节点 v , $N(v)$ 表示与 v 直接相连 (出边) 的节点集合。公式3.6定义了隐藏层的特征传递变换, 用于计算由 v_i 表示的节点的向前传递更新, 将邻

居节点表示与自身表示结合，生成新的节点表示：

$$h_i^{(l+1)} = \sigma(h_i^{(l)}w_0^{(l)} + \sum_{j \in N(v_i)} \frac{1}{c_i} h_j^{(l)}w^{(l)} + b^{(l)}) \quad (3.6)$$

其中 $h_i^{(l)}$ 表示节点 v_i 在第 l 层的特征， $w^{(l)}$ ， $b^{(l)}$ 分别表示第 l 层的权重和截距。 c_i 表示归一化因子， $\sigma(\cdot)$ 表示激活函数，比如 $ReLU(\cdot) = \max(0, \cdot)$ 。

经过研究^[75;76]说明，使用2层或3层的图卷积网络模型就可以得到很好的效果。在本研究中使用2层图卷积网络结构，如图3.12所示。网络的输入是由实体和标签构成的整张图结构，输出是图中节点和边的特征表示。隐藏层的层数代表节点特征能够传输的最远距离，在隐藏层中，每个节点仅从其邻居那里获取特征，在图中用橘色标出，将聚合的邻域表示与节点自身表示相加作为新的节点表示。每个节点收集信息的过程是独立进行的，对所有节点来说都是在同一时间进行。2层图卷积网络结构，即在一层的基础上再叠加一层，重复收集信息的过程。因此，最终节点收集到的信息还包含了2跳邻居的信息。

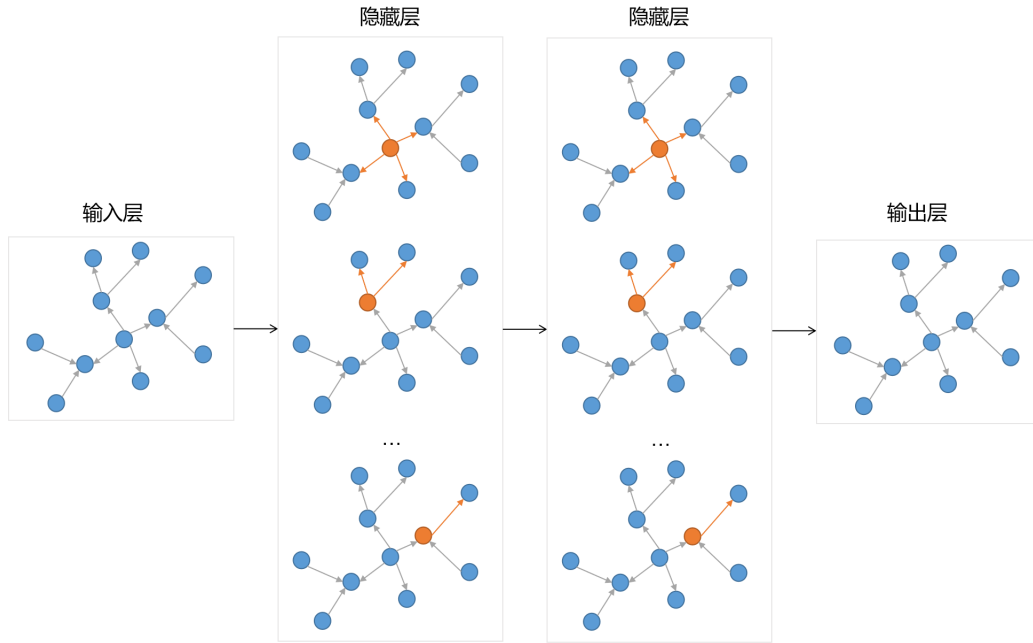


图 3.12: GCN 网络结构

3.6.2 实体标签预测

基于三元组的知识图谱补全方法通常以链接预测任务作为子任务，即预测出 $(?, r, e_t)$ 或 $(e_h, r, ?)$ 。在本研究的基于实体标签的知识补全模型中，输入的不再是原知识图谱，而是实体的标签集。知识补全中的链接预测任务也转变为对于实体标签的预测，预测出新的标签对 (e, l) 。

在图自编码模型中，**encoder** 部分使用图卷积网络模型将每个实体和标签映射为一个具体的向量表示，**decoder** 部分使用在标准预测任务中表现出优秀性能的 **DisMult** 模型，对图中的边重新建模。**DisMult** 模型对知识图谱中的三元组进行建模，将图中节点之间的边表示为对角矩阵的形式，并将得分函数定义为公式3.7。其中， v_h, v_t 分别表示头实体和尾实体， M_r 为关系矩阵。

$$f(v_h, v_t) = v_h^T \text{diag}(M_r) v_t \quad (3.7)$$

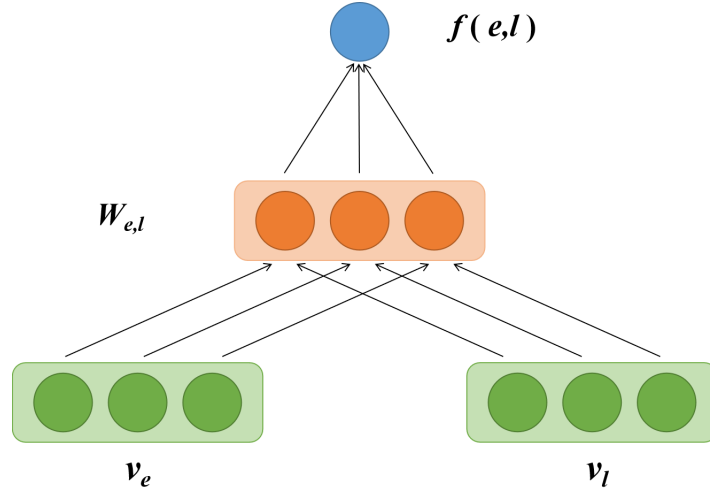


图 3.13: DisMult 模型结构

DisMult 模型依赖于 **encoder** 部分得到的节点表示，来对整张图重新建模。在标签预测任务中，对于候选标签对 (e, l) ，得分函数定义为公式3.8，图3.13给出了 **DisMult** 模型的核心思想， v_e, v_l 分别是实体 e 和标签 l 的向量表示， $W_{e,l}$ 表示实体和标签之间的边，是一个对角矩阵。

$$f(e, l) = v_e^T W_{e,l} v_l \quad (3.8)$$

算法3.4描述了知识补全模型算法的整体流程。给定实体和其对应的标签集，利用实体角色发现方法得到的实体嵌入初始化实体节点，令实体本身包含了丰富的语义信息。 $N[i]$ 表示实体 e 的 i 阶邻居节点，即经历了 i 跳后找到的邻居节点，特别地， $N[0]$ 为实体 e 本身。 $e[h]$ 表示实体经过 h 层 GCN 的嵌入表示， $v_{N_e[h]}$ 表示实体的 h 阶邻域表示，此时 h 阶邻居的嵌入表示包含了它们 $h-1$ 阶邻居的信息。 $\mathcal{F}(E, \mathbb{L})$ 表示实体和标签对的得分集合。

本研究采用随机污染正例标签对中的实体或标签来构建负样本。如公式3.9所示，使用交叉熵损失使得正例得分高于负例：

$$\mathcal{L} = -\frac{1}{|E|} \sum_{(e,l,y) \in Y} y \log \sigma(f(e, l)) + (1 - y) \log(1 - \sigma(f(e, l))) \quad (3.9)$$

算法 3.4 NRKC 算法

Input: $\mathcal{P} = \langle E, \mathbb{L}, \mathcal{B} \rangle$; V_e :representations of entities based on network role; H :depth of layer; $w, b, c, f(\cdot)$:trainable parameters

Output: $\mathcal{F}(E, \mathbb{L})$:prediction function

```

1: for  $e \in E$  do
2:    $\{N[i]\}_{i=0}^H \leftarrow GetNeighbor(e)$ 
3:    $e[0] \leftarrow V_e, \forall e \in N[0]$ 
4:   for  $h = 1, \dots, H$  do
5:     for  $e \in N[h]$  do
6:        $e[h] \leftarrow agg(e[h-1], v_{N_e[h]})$ 
7:     end for
8:   end for
9:    $\mathbf{e} \leftarrow e[H]$ 
10:  计算得分函数  $f(e, l)$ 
11:  更新参数
12: end for
13: return  $\mathcal{F}(E, \mathbb{L})$ .

```

其中 Y 表示正例和负例的集合, σ 表示 *sigmoid* 函数, y 是正负例标识, $y = 1$ 时表示正例, 而 $y = 0$ 时表示负例。

3.7 本章小结

本章详细介绍了基于网络角色的知识图谱缺失知识补全方法。利用实体角色发现的方法, 更加全面的捕获实体的特征, 提高了实体嵌入的质量, 并且应用于实体画像和知识补全任务。实体画像体现了实体的唯一性特征, 将实体画像标签作为知识补全模型的输入输出, 从一定程度上提高了模型的可理解性。

第四章 实验设计和评估

为了验证基于网络角色发现的知识图谱缺失知识补全方法的有效性，本章将从多个层面在数据集上进行实验，实验环境如下：

(1) 操作系统：Windows10 64 位，CentOS 7.5.184

(2) 硬件配置：CPU 1.8GHz、内存 32G

(3) 软件配置：JDK 1.8、Python 3.6

实验相关源码目前已发布在：<https://github.com/wds-seu/Ziyue-Wang-Networ-k-Role-based-Knowledge-Completion>.

4.1 数据集

在知识图谱补全任务中，常使用 FB15k^[1] 和 FB15k-237^[77] 作为数据集进行实验。考虑到本研究进行的标签预测任务中包含了对于属性标签的预测，选择了信息丰富的 DBpedia 数据集，具体描述如下：

(1) FB15k: FB15k 数据是从 Freebase¹抽取到的一系列三元组。Freebase 是一个巨大的事实数据库，其中的实体和关系都有类别，目前其公共 API 已不再提供，但可以根据 Freebase 中实体 id 找到 wikidata 中对应的实体映射。FB15k 作为其子集，描述了同义集之间的三元关系，包含大量对称/反对称关系和逆关系。

(2) FB15k-237: FB15k-237 为 FB15k 的子集。FB15k 作为早期的知识补全任务的数据集，存在一定的测试集泄露问题，其中的逆关系三元组预测，基于简单的规则推理就能达到先进模型的效果。因此，FB15k-237 将 FB15K 中的逆关系删除，仅包含对称/反对称关系和组合关系。

(3) DBpedia: DBpedia 是一个多领域多类型的综合型知识库，被广泛应用于关系学习任务。本研究使用的是 2016 年版本的 DBpedia 数据集²，包含了超过 500 万个的实体和 3800 万个事实三元组。由于 DBpedia 规模过大，本实验中挑选了具有丰富属性的特定领域子集，具体细节如下：

- **Band:** 提取了 DBpedia 中 Band 类型下的相关实体，并选择了与该领域紧密关联的关系，例如音乐类型 (Genre)、乐队成员 (BandMember)、唱片公司 (RecordLabel) 等等。属性中既有字符型属性，如乐队名称 (name)，也有数值型属性，如活跃时间的开始年份 (activeYearsStartYear)。

¹Freebase: <http://www.freebase.com>

²DBpedia: <https://wiki.dbpedia.org/downloads-2016-10>

- **University:** 提取了 DBpedia 中 University 类型下的相关实体，并选择了与该领域紧密关联的关系，例如所在城市（city）、所属联盟（affiliation）、学校校长（chancellor）等等。该类型下包含了丰富的属性，既有字符型属性，如学校颜色（officialSchoolColour）、座右铭（motto），也有数值型属性，如本科生人数（numberOfUndergraduateStudents）、学校规模（facultySize）、捐款数额（endowment）。
- **Book:** 提取了 DBpedia 中 Book 类型下的相关实体，并选择了与该领域紧密关联的关系，例如作者（author）、语言（language）、后续工作（subsequentWork）等等。属性中既有字符型属性，如出版商（publisher）、国际标准图书编号（isbn），也有数值型属性，如书页数（numberOfPages）。
- **RadioStation:** 提取了 DBpedia 中 RadioStation 类型下的相关实体，并选择了与该领域紧密关联的关系，例如所在城市（city）、语言（language）、广播区域（broadcastArea）。字符型属性有标语（slogan）、许可证（licensee）等，也有数值型属性，如高度（heightAboveAverage Terrain）、频率（frequency）。
- **Actor:** 提取了 DBpedia 中 Actor 类型下的相关实体，并选择了与该领域紧密关联的关系，例如国籍（nationality）、出生地（birthPlace）、代表作（knownFor）。属性中既有字符型属性，如姓名（name）、别名（alias），也有数值型属性，如出生年份（birthYear）、活跃开始年份（activeYearsStartYear）。

表 4.1: 数据集统计信息

Datasets	# Entities	# Relations	# Rel_ triples	# Attributes		# Attr_ triples	
				str	num	str	num
FB15k	14,951	1,345	483,142	/	/	/	/
FB15k-237	14,541	237	272,115	/	/	/	/
Band	12,588	16	100,846	4	2	2,679	1,805
University	3,901	19	7,741	6	7	4,068	1,822
Book	7,395	16	18,455	7	1	13,172	1,657
RadioStation	13,318	13	40,835	10	2	36,438	6,829
Actor	7,478	33	10,536	3	3	5,670	3,630

表4.1描述了各数据集的详细信息，“# Entities”表示实体数量，“# Relations”表示关系数量，“# Rel_ triples”表示关系三元组数量，“# Attributes”表示属性数量，“# Attr_ triples”表示属性三元组数量。其中“str”指字符型属性，“num”指数值型属性。FB15k

和 FB15k-237 中仅包含关系三元组，DBpedia 的相关子集中包含了丰富的属性信息和关系信息。DBpedia 的子集中，Band 和 RadioStation 数据量较大，Band 关系信息最丰富，RadioStation 属性信息更丰富。University、Book 和 Actor 数据集相对较小，三者属性类别的分布上有着明显的倾向性，Actor、University 中有丰富的数值属性，字符型属性和数值型属性分布较平均，而 Book 中只有一个数值型属性，字符型属性丰富。

4.2 基准方法和评估指标

4.2.1 基准方法

依据不同的评测任务，本研究从知识补全和实体画像两个任务出发，选取了以下方法作为 baseline 进行比较，从不同的角度来说明基于网络角色发现的知识图谱缺失知识补全方法的有效性。选取的基准方法具体描述如下：

(1) 知识补全模型：

TransE^[1]：该方法基于空间向量的平移不变性，将实体与关系映射到向量空间中，通过训练学习使得头实体与关系的向量表示之和尽可能等于尾实体向量，即 $h + r \approx t$ 。

DisMult^[72]：该方法将关系表示为对角矩阵的形式，使得实体和关系的信息可以进行深层次交互。

ComplEx^[27]：在 DisMult 的基础上，将嵌入表示拓展到复数空间中，能够解决更加复杂的关系类型。

NRKC*：在本研究介绍的图自编码模型中使用 GCN 作为 encoder，仍使用 DisMult 作为 decoder，随机初始化节点。GCN 的训练学习过程利用了实体的同质性相似性，捕获实体的邻域信息。

A-NRKC：在 NRKC* 的基础上，利用基于属性相似性的网络角色发现方法学习到的实体嵌入表示，作为实体的初始化表示。

S-NRKC：在 NRKC* 的基础上，利用基于结构相似性的网络角色发现方法学习到的实体嵌入表示，作为实体的初始化表示。

NRKC：在基于网络角色发现的实体嵌入方法中，按一定比例将 P^H 、 P^A 、 P^S 混合，生成包含多种相似性信息的实体嵌入，作为 NRKC 中实体的初始化表示。

(2) 实体画像：实体画像任务的关键在于对于标签的度量选取，选择以下方法进行比较：

Random：在所有候选标签中随机选择标签。

TF-IDF：通过使用 TF-IDF 度量标签的重要性，生成标签集。

Rule：使用一个简单的启发式规则来度量标签的区分性。给定一个实体类型 t 和与 t 相关候选标签 l ，我们将 E_t 定义为类型 t 下所有实体的集合，而 $E_t^l \subseteq E_t$ 表示与标签 l 相关的正例实体的集合。定义在公式 4.1 中的 $support(l)$ 表示正例的比率，假定具有低 $support$ 值的标签往往是不具备代表性的，而高 $support$ 值代表无区分性。则得分函数

如4.2公式所示， γ 为区分性阈值。

$$support(l) = \frac{|E_t^l|}{|E_t|} \quad (4.1)$$

$$f(l) = -\log|support(l) - \gamma| \quad (4.2)$$

DeepWalk (H): 仅使用基于同质性的 H 策略生成的 P^H 路径。

A: 仅使用基于属性相似性的 A 策略生成的 P^A 路径。

S: 仅使用基于结构相似性的 S 策略生成的 P^S 路径。

HAS: 按照 1 : 1 : 1 的比例将 P^H 、 P^A 、 P^S 三种路径混合，学习实体嵌入表示。

4.2.2 评估指标

在标签预测任务中，通过得分函数来计算每个候选标签与实体的得分，按照降序对其进行排列。本研究使用在知识补全工作中常用的两个评估指标：**Mean Reciprocal Rank (MRR)** 和 **Hits@N**，具体描述如下：

(1) MRR

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \quad (4.3)$$

如公式4.3所示，其中 Q 为测试集， $|Q|$ 表示测试集中标签对个数， $rank_i$ 表示在第 i 个 $\langle e, ? \rangle$ 预测中，第一个正确答案的排名。若第一个正确答案排在第 n 位，则 MRR 得分就是 $\frac{1}{n}$ 。如果没有正确答案，则得分为 0。

(2) Hits@N

$$Hits@N = \frac{N_{acc}}{N} \quad (4.4)$$

如公式4.4所示，其中 N_{acc} 表示按得分排名的前 N 个标签中正确答案的数量。

在实体画像任务中，选择 **mean Average Precision (MAP)** 和 **F-Measure** 作为评估指标，具体表述如下：

(1) MAP

MAP，即 AP 的平均值，通过计算 Precision-Recall 曲线下面积得到 AP 值。精度 (Precision) 和召回率 (Recall) 的定义如公式4.5和4.6所示：

$$Precision = \frac{TP}{TP + FP} \quad (4.5)$$

$$recall = \frac{TP}{TP + FN} \quad (4.6)$$

其中 TP 指将正例预测为真，FP 指将负例预测为真，FN 指将正例预测为假。

(2) F-Measure

F 值定义为精度和召回率的调和平均，如公式4.7所示，F 值较高时说明方法有效，结果符合预期。

$$F - Measure = \frac{2Recall * precision}{Recall + Precision} \quad (4.7)$$

4.3 实体画像结果评估

本研究在各个数据集上进行实体画像标签评估的实验，来验证基于网络角色发现的实体嵌入方法的有效性。

表 4.2: 各数据集标签统计信息

数据集	AVL	AIL	REL	RAL
FB15k	/	/	133,909	/
FB15k-237	/	/	59,734	/
Band	545,923	510	21,471	66,488
University	11,538	21	10,071	18,005
Book	24,293	3	13,795	26,597
RadioStation	18,150	12	2,136	27,952
Actor	2,213	17	2,959	175,829

利用自动化枚举的方式构建每个数据集的标签池，每个标签的统计信息如表4.2所示。FB15k 和 FB15k-237 中仅包含 REL 标签，其中 FB15k 中关系标签最丰富。Band 数据集中信息丰富，虽然其数值型属性较少，但是它的实体数量最多，因此属性标签在三个数据集中最多。University 数据集中，尽管其数值型属性更为丰富，但是其实体数量较少，在数值属性上分布较为集中。因此，关系标签和属性标签分布较为平均。Book 数据集中数值型属性很少，导致属性区间标签极少，相较之下，关系标签较丰富。RadioStation 数据集字符型属性较多，Actor 数据集属性标签和关系标签分布较平均，但由于与它相关联的实体拥有丰富的属性，因此它的 RAL 标签最多。

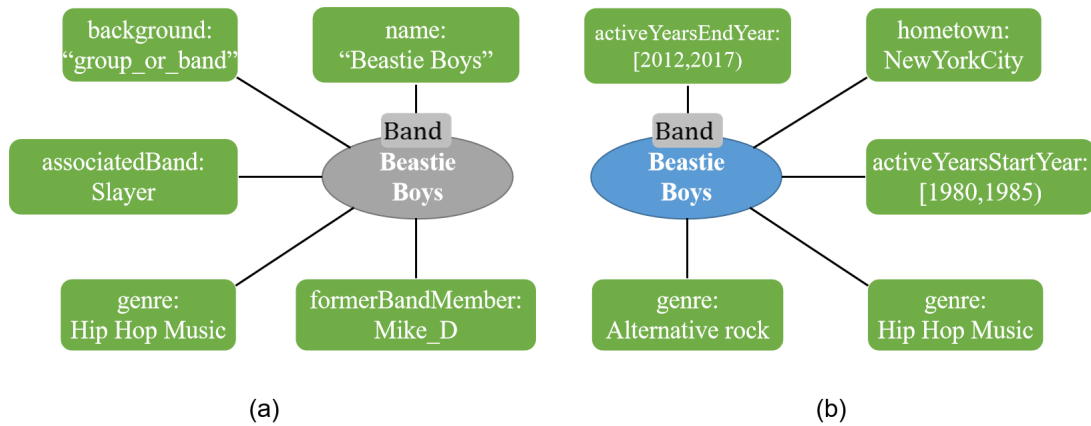


图 4.1: 实体画像结果对比：(a) Random 方法 vs. (b) HAS 方法

图4.1给出了基于 Random 方法的实体画像结果和基于网络角色发现的实体嵌入方

法的实体画像结果的对比。(a) 图为 Random 方法得到的结果, (b) 图为本研究的 HAS 方法得到的实体标签。从图中能够看出, 基于本研究的嵌入方法得到的实体画像标签结果能够鲜明地体现 *Band* 实体区别于其他乐队实体的特点, 而 Random 方法得到的实体画像标签结果, 如 $\langle name: "BeastieBoys" \rangle$ 是其独有的特征, $\langle background: "group_or_band" \rangle$ 是所有乐队实体都具有的普遍性特征, 不具有区分性。

我们邀请了 5 位知识图谱专家阅读五个子集, 并手工构建每种类型实体的基准标签集。为了减少人工工作, 基准标签集中的标签是简化的标签, 只包含标签的属性部分。专家们在区分独特性标签上意见基本一, top-5 和 top-10 标签组的平均一致性分别为 2.87 和 6.09。

表 4.3: 数据集 FB15k 和 FB15k-237 的实体画像评估结果

模型	FB15k				FB15k-237			
	MAP@		F-Measure@		MAP@		F-Measure@	
	5	10	5	10	5	10	5	10
Random	0.085	0.137	0.154	0.197	0.056	0.139	0.083	0.167
TF-IDF	0.109	0.212	0.187	0.268	0.106	0.251	0.182	0.274
Rule	0.172	0.231	0.193	0.317	0.139	0.238	0.206	0.298
DeepWalk(H)	0.251	0.392	0.280	0.434	0.193	0.421	0.268	0.461
S	0.173	0.243	0.180	0.298	0.121	0.311	0.189	0.326
HS	0.248	0.402	0.284	0.411	0.201	0.449	0.298	0.487

在表4.3中, 通过计算 MAP 和 F-measure 来比较在 FB15k 和 FB15k-237 数据集上基于网络角色的实体嵌入方法和其他方法的优劣。在每个指标上, 我们比较了基于网络角色的方法生成的 top-5/10 标签与标准答案之间的一致性。由于两个数据集中都没有属性信息, 所以不对 A 策略进行评估, HAS 模型等价于 HS 模型。在 FB15k 中包含了大量的关系信息, DeepWalk (H 策略) 表现最优, HS 模型仅次于它。而在 FB15k-237 中删除了大量的逆关系之后, HS 表现最优, 得到的实体画像结果更能体现实体的特征。

在表4.4、4.5中, 比较了在 DBpedia 子集上基于网络角色的实体嵌入方法和其他方法的评估结果。实验表明, 基于网络角色的一系列策略方法都优于 baselines。从评估结果可以看出, 仅对候选标签进行启发式过滤不足以生成高质量的标签集。在 Band 中, 与 A 策略和 S 策略相比, H 策略在识别区别性标签方面具有更好的性能, 这是因为 Band 中有大量关系信息。A 策略在 University、RadioStation 和 Actor 数据集上体现出它的性能提升效果, 这是因为这三个数据集中属性标签更加丰富, 而在实体类型丰富的 Book 数据上, S 策略明显优于 A 策略, H 策略最优。

表 4.4: 数据集 Band、University、Book 的实体画像评估结果

数据集	方法	MAP@5	MAP@10	F-M@5	F-M@10
Band	Random	0.105	0.167	0.142	0.281
	TF-IDF	0.109	0.197	0.167	0.296
	Rule	0.118	0.241	0.242	0.408
	DeepWalk(H)	0.166	0.273	0.278	0.410
	A	0.115	0.201	0.212	0.331
	S	0.098	0.165	0.237	0.339
	HAS	0.165	0.298	0.317	0.524
University	Random	0.027	0.114	0.038	0.141
	TF-IDF	0.033	0.049	0.159	0.208
	Rule	0.050	0.048	0.172	0.213
	DeepWalk(H)	0.241	0.410	0.269	0.429
	A	0.207	0.323	0.232	0.337
	S	0.198	0.275	0.197	0.301
	HAS	0.283	0.417	0.311	0.475
Book	Random	0.036	0.133	0.034	0.158
	TF-IDF	0.076	0.190	0.081	0.240
	Rule	0.045	0.183	0.062	0.209
	DeepWalk(H)	0.253	0.440	0.283	0.549
	A	0.117	0.224	0.112	0.231
	S	0.210	0.321	0.237	0.403
	HAS	0.281	0.456	0.263	0.552

显然，从实验数据可以看出对于实体嵌入模型来说，H、A、S 三种策略的选择需要一个有偏加权方案，针对不同的应用场景，动态地调整权重。

4.4 实体标签预测任务

实体标签预测的目标是根据给定的标签集，预测出新的标签对 (e, l) 。在不同的数据集上进行实验，通过与其它知识补全模型的比较，分析各个模型的特点，验证基于网络角色发现的知识图谱缺失知识补全方法的有效性。

表 4.5: 数据集 RadioStation 和 Actor 的实体画像评估结果

数据集	方法	MAP@5	MAP@10	F-M@5	F-M@10
RadioStation	Random	0.116	0.132	0.078	0.143
	TF-IDF	0.135	0.206	0.111	0.250
	Rule	0.128	0.231	0.097	0.313
	DeepWalk(H)	0.301	0.392	0.343	0.553
	A	0.227	0.295	0.264	0.440
	S	0.207	0.301	0.249	0.428
	HAS	0.301	0.447	0.361	0.632
Actor	Random	0.021	0.053	0.023	0.158
	TF-IDF	0.036	0.076	0.046	0.103
	Rule	0.051	0.092	0.096	0.178
	DeepWalk(H)	0.126	0.181	0.233	0.398
	A	0.098	0.163	0.209	0.363
	S	0.078	0.147	0.169	0.294
	HAS	0.134	0.283	0.231	0.420

4.4.1 实验结果及分析

在实验中，每个实体和标签被表示为一个 200 维的向量，负采样率设置为 10。与 NRKC 相关的方法中使用 Adam 优化算法，学习率为 0.01。其他 baseline 的参数设置遵从 ComplEx^[27] 中的数据。对于数据丰富的 FB15k 和 FB15k-237 数据集，在进行实体角色发现的实体嵌入学习时，设置节点采样次数为 100，路径长度为 8，邻域窗口大小为 5。而在 DBpedia 的子集中，设置采样次数为 10，路径长度为 8，邻域窗口大小为 5。

在 FB15k 和 FB15k-237 的实验中，给出了两种测试结果 Raw 和 Filtered。由于在测试时通过替换得到的标签对并不一定就是负例，可能恰巧是正例，例如 $\langle Obama, \langle presidentOf, USA \rangle \rangle$ 被替换为 $\langle Trump, \langle presidentOf, USA \rangle \rangle$ ，排名高也是正确的。因此测试时在替换后需要检查一下新三元组是否出现在训练集中，这就是 Filtered 训练方法。

数据集 FB15k 和 FB15k-237 中，仅包含了实体的关系标签和实体类型信息。因此无法学习到基于属性相似性的实体嵌入表示，即 A-NRKC 等价于 NRKC*。此时本研究的 NRKC 仅包含了同质性和结构相似性信息，等价于 S-NRKC。实验结果数据如表 4.6 所示，用粗体标出了最好的实验结果。

由于 FB15k 数据集中包含了大量的逆关系三元组，存在一定的测试集泄露问题。大

表 4.6: 数据集 FB15k 和 FB15k-237 的实体标签预测结果

模型	FB15k					FB15k-237				
	MRR		Hits@			MRR		Hits@		
	Raw	Filtered	1	3	10	Raw	Filtered	1	3	10
TransE	0.221	0.380	0.231	0.472	0.641	0.144	0.233	0.147	0.263	0.398
DisMult	0.248	0.634	0.522	0.718	0.814	0.100	0.191	0.148	0.247	0.403
ComplEx	0.242	0.692	0.599	0.759	0.840	0.109	0.201	0.112	0.213	0.388
NRKC*	0.251	0.651	0.533	0.726	0.802	0.152	0.248	0.151	0.264	0.415
NRKC	0.273	0.683	0.576	0.767	0.842	0.161	0.287	0.162	0.277	0.458

部分模型都可以达到比较好的预测效果。除了逆关系之外，对称/反对称关系在 FB15k 中也占有很大比重。TransE 模型的预测效果明显弱于其他方法，是因为它无法对对称和反对称关系三元组建模，不能很好地处理这样的预测问题。ComplEx 模型作为 DisMult 的拓展方法，能够很好地对对称/反对称关系建模，在 MRR 和 Hits@1 指标上略优于 NRKC。NRKC* 与其他 baseline 相比，没有仅仅局限于实体标签对本身，考虑到了实体邻域信息的重要性，实验效果达到了平均水平。NRKC 与 NRKC* 相比，还加入了实体结构相似性信息，在一定程度上丰富了实体表示的语义信息，对实验结果有着积极的作用，预测效果优于大部分 baselines。

去除了逆关系三元组之后，与 FB15k 数据集相比，FB15k-237 数据集上的关系更加稀疏，使得仅依赖于关系的表示学习方法在标签预测任务上效果大大下降。NRKC 模型中融合了 S 策略，不仅考虑到了实体同质性，还考虑了实体结构相似性，预测效果明显优于其他方法，解决了现有表示学习方法受关系稀疏性影响较大的问题。

数据集 DBpedia 的 Band、University、Book、RadioStation 和 Actor 子集包含了丰富的属性信息，能够从属性相似性方面来学习实体的嵌入表示。表 4.7、4.8 给出了各个模型在 DBpedia 子集上的实验结果，并用粗体标出了最好的评估得分。

Band 数据集中关系三元组丰富，其中有许多一对多/多对多的复杂关系，例如 $\langle \text{Megadeth}, \text{associatedBand}, \text{Metallica} \rangle$ 和 $\langle \text{Megadeth}, \text{associatedBand}, \text{Mass_Mental} \rangle$ ，这使得实验结果中 TransE 模型表现最差。Band 中包含的数值型属性较少，而字符型属性要么是具有普遍性的标签，如 $\langle \text{background}, \text{"group_or_band"} \rangle$ ，要么是具有唯一性的属性，如乐队名称 (name)。因此，基于属性相似性的 A 策略对于 NRKC* 性能的提升有限。相反，基于结构相似性的 S 策略对于模型的性能提升明显，在最终的 NRKC 模型中 S 路径所占比重较大，在下一小节中将具体分析。

Book 数据集在实体类型上更加丰富，数值型属性极少，只有 “numberOfPage”，

表 4.7: 数据集 Band、University、Book 的实体标签预测结果

数据集	模型	MRR	Hits@1	Hits@3	Hits@10
Band	TransE	0.197	0.194	0.238	0.341
	DisMult	0.233	0.249	0.319	0.406
	ComplEx	0.250	0.248	0.289	0.428
	NRKC*	0.276	0.210	0.313	0.429
	A-NRKC	0.277	0.224	0.312	0.431
	S-NRKC	0.298	0.225	0.337	0.439
	NRKC	0.301	0.226	0.344	0.440
University	TransE	0.140	0.074	0.128	0.197
	DisMult	0.205	0.131	0.233	0.316
	ComplEx	0.251	0.168	0.245	0.315
	NRKC*	0.203	0.167	0.209	0.373
	A-NRKC	0.291	0.257	0.296	0.455
	S-NRKC	0.272	0.203	0.244	0.406
	NRKC	0.389	0.258	0.393	0.519
Book	TransE	0.097	0.054	0.100	0.144
	DisMult	0.134	0.091	0.147	0.215
	ComplEx	0.135	0.102	0.145	0.222
	NRKC*	0.165	0.118	0.183	0.254
	A-NRKC	0.165	0.119	0.189	0.256
	S-NRKC	0.197	0.156	0.201	0.318
	NRKC	0.212	0.173	0.223	0.340

且每个 Book 实体在 *numberOfPage* 上的区间分布相对集中, 区分性不大, 因此 A 策略的提升效果不大, S 策略提升效果显著。Book 数据集在关系信息和属性信息的数据分布倾向性上与 Band 数据集相似, 因此各个模型在其上的预测表现与 Band 数据集类似。

University、RadioStation 和 Actor 数据集较其他两个数据集而言, 属性更丰富一些。University 中包含了丰富的数值型属性, 因此 A 策略在该数据集上的提升效果更加明显。RadioStation 中关系信息和属性信息分布相对平均, 包含的数值型属性最丰富, A 策略的效果优于 S 策略。表 4.8 结果显示, 在 Actor 数据集上关系信息更加复杂, Complex 擅长将节点映射到高维空间中来处理这种复杂关系, 表现效果最好。NRKC* 较 TransE 和 DisMult 这类基础模型, 有更好的预测效果, Actor 数据集上字符属性和数值属性分布较

均衡，A 策略较 S 策略而言对模型的性能提升帮助更大。总体来看，综合了属性相似性和结构相似性的 NRKC 模型在数据集上的预测效果最优。

表 4.8: 数据集 RadioStation 和 Actor 的实体标签预测结果

数据集	模型	MRR	Hits@1	Hits@3	Hits@10
RadioStation	TransE	0.174	0.163	0.211	0.286
	DisMult	0.198	0.192	0.244	0.311
	ComplEx	0.197	0.203	0.281	0.327
	NRKC*	0.201	0.196	0.273	0.310
	A-NRKC	0.267	0.218	0.293	0.379
	S-NRKC	0.226	0.199	0.284	0.361
	NRKC	0.283	0.232	0.302	0.407
Actor	TransE	0.103	0.082	0.134	0.167
	DisMult	0.175	0.165	0.169	0.238
	ComplEx	0.246	0.188	0.241	0.321
	NRKC*	0.217	0.141	0.198	0.263
	A-NRKC	0.263	0.154	0.231	0.318
	S-NRKC	0.241	0.150	0.255	0.332
	NRKC	0.289	0.177	0.261	0.383

纵观各模型在不同的数据集上的标签预测表现，NRKC 模型明显优于其他模型。区别于选取的 baselines，NRKC 模型关注到邻域信息的重要性，并从属性相似性和结构相似性学习实体的表示，解决了关系稀疏的问题。

4.4.2 超参数分析

基于网络角色的知识图谱缺失知识补全方法中最关键的超参数就是用于生成实体嵌入的路径混合比例。由于图卷积神经网络是根据实体间的直接链接寻找邻域，本身就包含了实体的同质性，故在路径混合时仅考虑基于属性相似性的 A 路径和基于结构相似性的 S 路径。在生成实体嵌入时，嵌入维度对于最终的实体表示质量也有极大的影响。在知识补全模型的训练过程中，训练集数据的大小也对最终的预测效果产生影响。本小节将就 λ_A 、 λ_S 、嵌入维度 d 和训练集比例 T_R 进行敏感性分析，探究其对实体标签预测任务的影响。

4.4.2.1 路径混合比例参数敏感性分析

图4.2给出了在 DBpedia 的五个子集下，不同的路径混合比例对实体标签预测结果的 MRR 值的影响。(a)、(b)、(c) 分别、(d) 和 (e) 对应于数据集 Band、University、Book、RadioStation 和 Actor 的分析结果，横坐标为 λ_A ，纵坐标为 MRR 值，当 λ_A 为 0.2 时，则对应的 λ_S 为 0.8。当 λ_A 为 0 时，即使用 S-NRKC 模型。当 λ_A 为 1 时，就是使用 A-NRKC 模型。

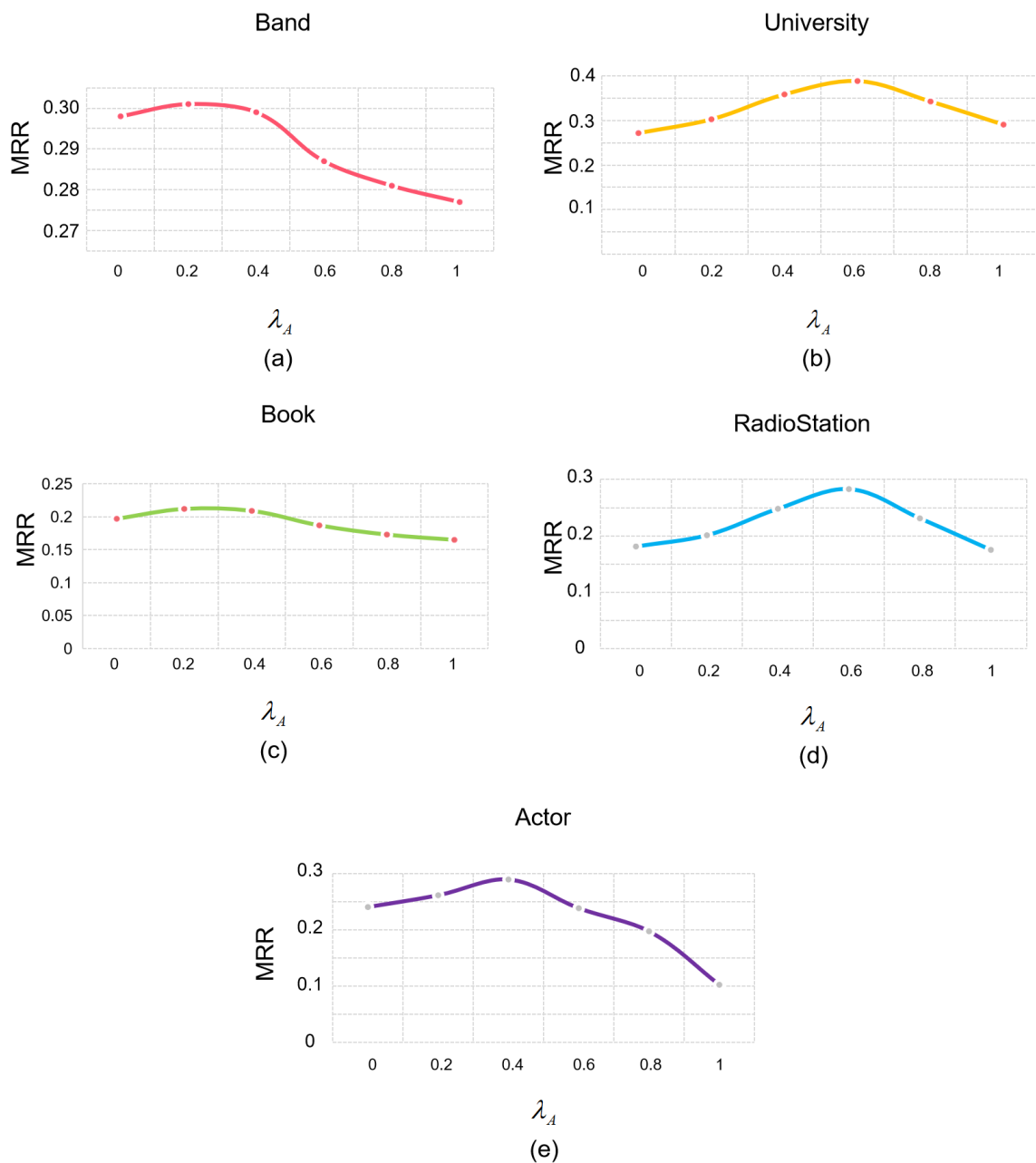


图 4.2: 路径权重参数分析

在 Band 数据集中，随着 λ_A 比重的增加，MRR 值有略微提升，当 $\lambda_A = 0.2$ 时达到

最高数值。但随着比重的继续增加，MRR 值下降较快，这是因为 Band 中属性较少，对于实体的嵌入表示学习帮助不大。University 数据集中属性丰富，随着 λ_A 比重的增加，MRR 值有大幅度提升，在 $\lambda_A = 0.6$ 处达到最高值。Book 数据集中实体类型丰富，S 路径对于预测效果有极大的帮助，随着 λ_A 比重的增加，MRR 值逐渐下降，在 $\lambda_A = 0.2$ 处达到最高值。RadioStation 数据集中虽然包含的数值型属性较少，但各个实体在属性上呈现的差异明显，能够很好地学习到实体间的相似性，因此 A 路径对模型有很大的帮助，在 $\lambda_A = 0.6$ 处达到最高值。Actor 数据集里实体类型丰富，同时实体在数值型属性上呈现明显的区分性，A 路径和 S 路径都发挥了重要的作用，在 $\lambda_A = 0.4$ 处 MRR 值达到最高。

通过对路径权重参数的实验分析，可以看出对于本研究的知识补全模型来说，A 路径和 S 路径的比重需要根据不同特点的数据集动态地调整。显然，对于属性丰富的数据集，A 路径能够使模型性能得到明显提升，而对于实体类型丰富的数据集，S 路径能够起到很大的帮助。

4.4.2.2 向量维度敏感性分析

图4.3给出了在 DBpedia 数据集的 5 个子集上向量维度对于预测结果的影响，横坐标为向量维度，取 50-250 维，间隔为 50，纵坐标为 MRR 值。

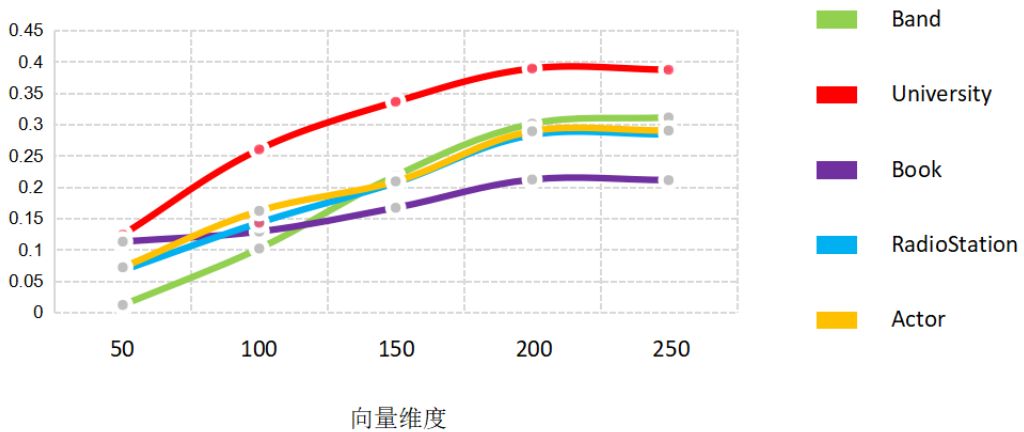


图 4.3: 向量维度参数分析

从图中可以看出，向量维度越高，实体表示包含的特征越多，MRR 值越高。当向量维度到达 200 维之后，MRR 值不再有太大的起伏，趋于稳定。因此，对于实体的表示一般取高维度的 200 较为合适。

4.4.2.3 训练集比例敏感性分析

图4.4给出了在 FB15k、FB15k-237、Band、University 和 Actor 五个各具特点的数据集上训练集比例对于预测结果的影响，横坐标为训练集比例 T_R ，纵坐标为 MRR 值。前三个数据集中关系信息丰富，后两个数据集属性信息更丰富。

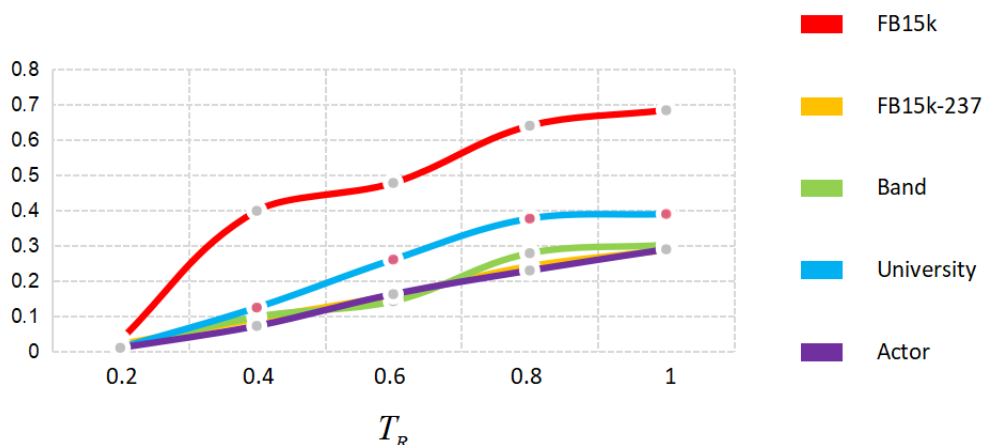


图 4.4: 训练集比例敏感性分析

从图中可以看出，随着训练数据的增多，MRR 值越高，模型的预测效果越好。在 University 数据集中，当 $T_R = 0.8$ 时，MRR 趋于稳定，这是因为该数据集中关系和属性较简单，容易达到较好的效果。而对于关系较为复杂的其它四个数据集来说，增加适当的训练数据能使模型达到更好的预测效果。

4.5 连续值离散化

在以往的工作中^[4]，都是直接对数值型属性的具体值进行预测。显然，MRR 和 Hits@N 指标不适用于连续值属性的预测，因此常常使用方差和标准差来衡量连续值属性预测的结果。利用连续值离散化的预处理方法，虽然降低了属性预测的精确性，但提高了预测结果的准确度，基本满足用户对属性预测的需求。为了验证属性预处理步骤的有效性，选择在包含了数值型属性的 DBpedia 数据集的 5 个子集上进行消融实验，实验组是将连续值属性离散化后的知识图谱，对照组是未处理的原始知识图谱。表4.9给出了各个模型在预处理前后的标签预测结果。

在表中，用粗体标出了预测效果最好的得分，从表中的结果可以看出，将连续值属性离散化的预处理方法使得每个模型的预测结果都有一定的提升。对于 University、Actor 数据集来说，该预处理方法提升模型性能的效果更加明显，这是因为这两个数据集中数值型属性最丰富。实验数据证明，连续值离散化的预处理方式具有高适用性，提高了预测结果的准确度。

表 4.9: 属性预处理消融实验结果

	模型	实验组				对照组			
		MRR	Hits@			MRR	Hits@		
			1	3	10		1	3	10
Band	TransE	0.197	0.194	0.238	0.341	0.163	0.162	0.201	0.298
	DisMult	0.233	0.249	0.319	0.406	0.201	0.221	0.256	0.387
	ComplEx	0.250	0.248	0.289	0.428	0.231	0.240	0.276	0.390
	NRKC	0.301	0.226	0.334	0.440	0.287	0.226	0.331	0.412
University	TransE	0.140	0.074	0.128	0.197	0.122	0.071	0.128	0.185
	DisMult	0.205	0.131	0.233	0.316	0.184	0.122	0.212	0.287
	ComplEx	0.251	0.168	0.245	0.315	0.239	0.162	0.238	0.299
	NRKC	0.389	0.257	0.393	0.519	0.361	0.255	0.376	0.508
Book	TransE	0.097	0.054	0.100	0.144	0.089	0.053	0.098	0.139
	DisMult	0.134	0.091	0.147	0.215	0.129	0.090	0.142	0.207
	ComplEx	0.135	0.102	0.145	0.222	0.131	0.092	0.139	0.208
	NRKC	0.212	0.173	0.223	0.340	0.204	0.169	0.200	0.335
RadioStation	TransE	0.174	0.163	0.211	0.286	0.171	0.163	0.198	0.271
	DisMult	0.198	0.192	0.244	0.311	0.193	0.191	0.237	0.303
	ComplEx	0.197	0.203	0.281	0.327	0.183	0.192	0.276	0.310
	NRKC	0.283	0.192	0.302	0.407	0.278	0.192	0.291	0.392
Actor	TransE	0.103	0.082	0.134	0.167	0.092	0.081	0.119	0.152
	DisMult	0.175	0.165	0.169	0.238	0.172	0.165	0.167	0.211
	ComplEx	0.246	0.188	0.241	0.321	0.241	0.182	0.238	0.317
	NRKC	0.289	0.177	0.261	0.383	0.278	0.165	0.260	0.376

4.6 可理解性任务

为了验证基于实体标签的知识图谱补全方法的有效性,本研究设计了外部实验来进行评估。在该实验中,我们手工构建了 20 个选择题,需要选择与该实体匹配的标签。每个选择题包含 4 个选项标签,这 4 个标签的标签名相同,但标签值不同,在每个问题中有且只有一个标签与该实体匹配。十位用户被邀请来参与这个实验,他们被要求准确且快速地选择出实体标签。将用户分为实验组和对照组,其中提供给实验组的每个用户实

体的画像标签结果，给对照组的用户只提供原始的三元组描述。若是用户根据实体标签能够快速选出正确答案，从侧面说明基于实体标签的预测结果符合逻辑，能够快速被理解。

Q : Which is the *activeYearsStartYear* of *Megadeth* ?
A.[1980, 1985) B.[1985, 1990) C.[1945, 1950) D.[2000, 2005)

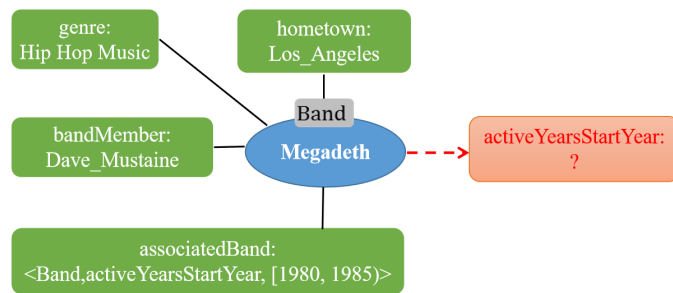


图 4.5: 可理解性实验：实验组示例

图4.5给出了本实验设计的实验组知识补全问卷示例。为实验组的用户提供了乐队实体 *Megadeth* 的画像标签结果，包括家乡 (*hometown*)、音乐类型 (*genre*)、乐队成员 (*bandMember*) 等信息，在图中用红色标出了需要推断的活跃开始年份的属性值。根据其中关键的乐队实体 *Megadeth* 标签 $\langle associatedBand : \langle Band, activeYearsStartYear, [1980, 1985) \rangle \rangle$ 能够直观快速地推断出它的活跃开始年份在 $[1980, 1985)$ ，正确答案为 A。

Q : Which is the *activeYearsStartYear* of *Megadeth* ?
A.[1980, 1985) B.[1985, 1990) C.[1945, 1950) D.[2000, 2005)

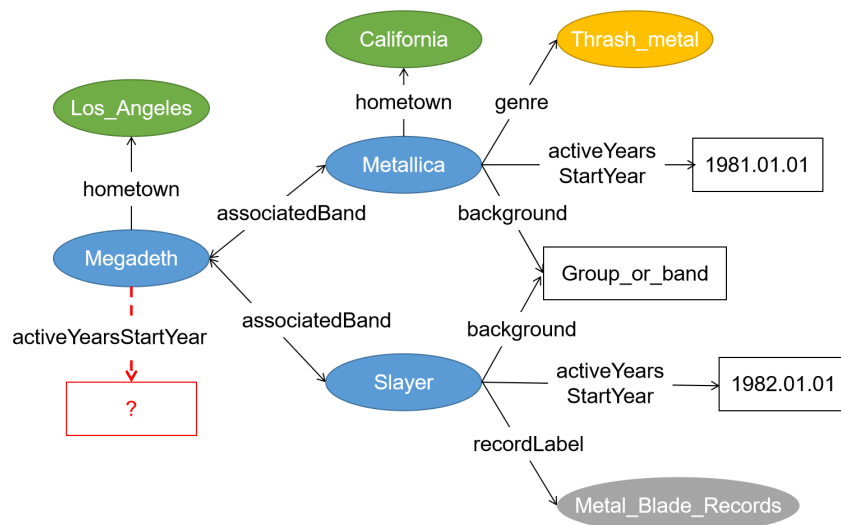


图 4.6: 可理解性实验：对照组示例

图4.6给出了本实验设计的对照组知识补全问卷示例。对照组的用户需要根据知识

图谱中的原始三元组描述，回答与实验组相同的问题，即乐队实体 *Megadeth* 的活跃开始年份的属性值。从图中能够看出，与实体标签相比，三元组的表示形式冗余且复杂，需要花费大量时间和精力来找到正确答案。

最终，本实验统计了用户在问卷中的平均准确率和他们在每个问题上花费的平均时间，具体的实验结果如表4.10所示，实验组中正确率最高能达到 95%，而对照组中最高只有 85%。在表中用粗体标出了实验结果平均值中效果最好的指标值，实验组的平均正确率比对照组高出 10%，所花费的平均时间却是对照组的一半。

表 4.10: 基于实体标签的知识图谱补全方法外部实验评估结果

	结果					平均值
正确率（实验组,%）	95	90	95	85	80	89
正确率（对照组,%）	75	80	85	70	85	79
时间（实验组,min）	2.35	3.14	3.26	3.40	3.12	3.05
时间（对照组,min）	5.58	7.32	7.11	6.32	7.46	6.756

图4.7、4.8展示了用户问卷结果的所耗时间和正确率的分布情况。图4.7中横坐标表示时间，纵坐标表示在所在组别中占的比例。在时间分布上，实验组和对照组之间的区分性很明显，实验组所耗时间大多集中在 3-4 分钟，而对照组大多集中在 7-8 分钟，几乎是实验组的两倍。图4.8中横坐标表示正确率，纵坐标表示在所在组别中占的比例。能够看出实验组中分布较均匀，大多数集中在 90%-95%，对照组中最高在 80%-85%。

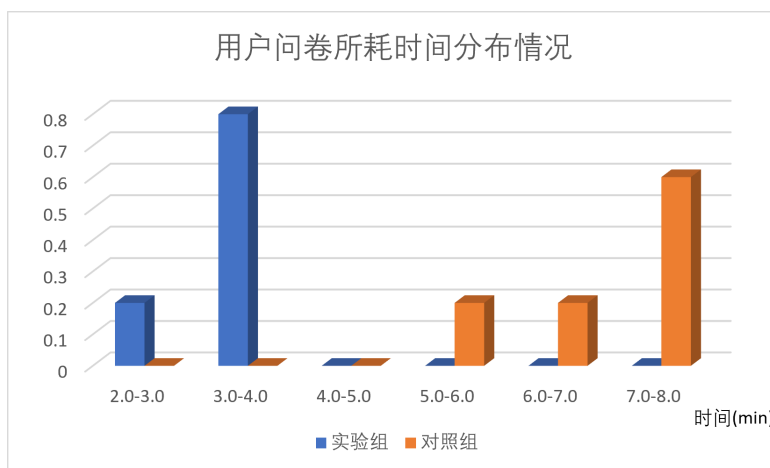


图 4.7: 用户问卷所耗时间分布情况

实验显示，实验组的用户在问卷中表现出更高的准确率，花费的时间明显缩短。与冗长而复杂的三元组描述相比，实体画像的标签结果以简洁的方式表达了丰富的信息，这极大地帮助用户快速理解知识补全的结果。

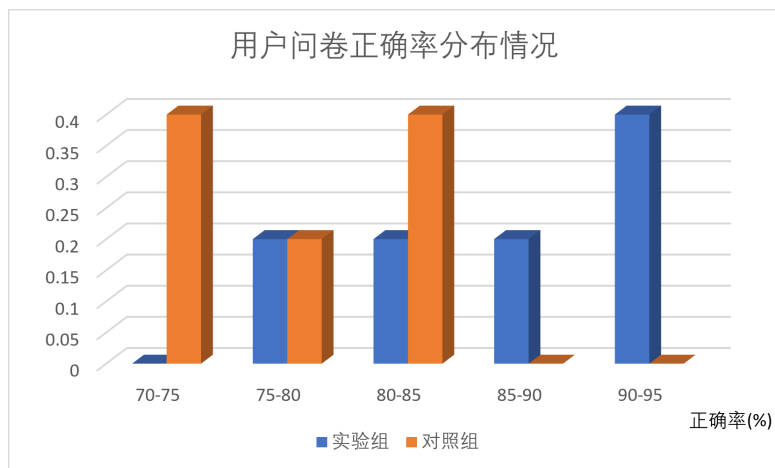


图 4.8: 用户问卷正确率分布情况

实体画像标签本身的特点就是为实体贴上具有区分性的标签，使用户快速理解实体。通过外部评估实验，结果表明与传统的基于三元组的知识补全方法相比，基于实体标签的知识图谱缺失知识补全方法在一定程度上提高了实体标签预测结果的可理解性。

4.7 本章小结

本章主要是通过多个层面的实验来验证基于网络角色发现的知识图谱缺失知识补全方法的有效性，并给出基于不同的实体相似性生成的实体嵌入的适用性分析。在上述的实验中，基于网络角色发现的实体嵌入方法能够很好地解决知识图谱中关系稀疏问题，提高画像标签质量和知识补全模型的精度。将属性连续值离散化的预处理方法，使得实体更易于理解，并且提高预测的精度。以实体画像标签作为补全模型的目标，在一定程度上提高了补全结果的可理解性，提高了人工智能模型的安全性。

第五章 系统设计与实现

本章介绍了基于网络角色的知识图谱补全系统的设计和实现，该系统能够根据输入的知识图谱数据内容，预测出潜在的知识，并以实体画像标签的形式直观的展示给用户，帮助用户快速理解预测结果。

5.1 系统需求分析

结合知识补全的概念含义和相关领域人员的实际需求，本节将从功能需求和性能需求两方面分析系统。

5.1.1 功能需求

目前知识图谱相关的可视化工具，均是展示知识图谱的 **RDF** 描述信息，这使得必须是知识图谱相关领域的人员花费一定的时间才能够从中获取有价值的信息。本研究的知识补全系统以实体画像标签的形式展示知识补全的预测结果，为用户更加直观的理解实体和预测结果提供了可能性。整个系统的功能需求如图5.1所示：

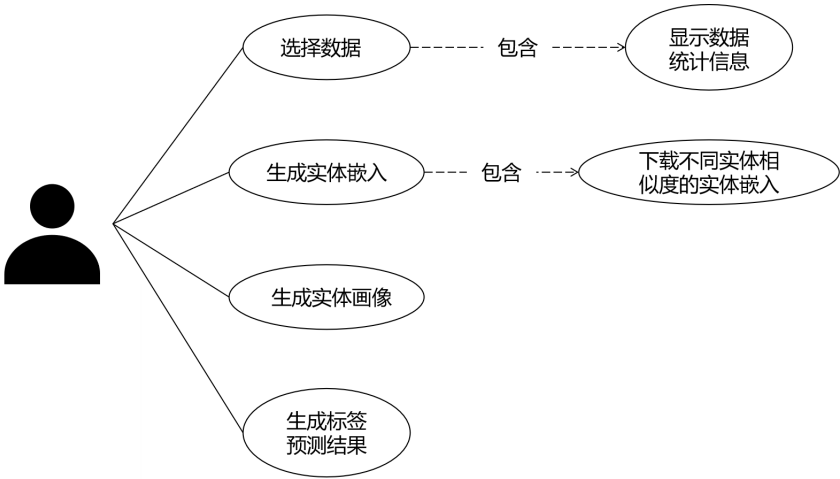


图 5.1: 知识图谱补全系统用例图

(1) 选择知识图谱数据：从导航栏中选择需要进行预测的数据集，本系统提供了 **DBpedia** 数据集下的相关子集，供于预测任务。

(2) 生成基于不同实体角色发现的实体嵌入：根据选择的数据集，利用不同的基于实体角色发现的实体嵌入方法，得到包含不同的实体相似性的多种实体嵌入表示。本系统可生成基于同质性、基于属性相似性和基于结构相似性的实体嵌入，并且可下载查看其具体内容。

(3) 可视化生成实体画像标签结果：将由基于实体角色发现的实体嵌入生成的实体画像展示在网页前端，使用户能够快速理解该实体。

(4) 生成标签预测结果：根据选择数据集，以实体标签作为模型的输入，展示出标签预测的结果，并显示其相关的实体画像，帮助用户快速理解预测结果。

5.1.2 性能需求

除了保证功能需求，基于网络角色的知识补全系统在性能上应基本满足：

(1) 准确性：保证从知识图谱中得到的补全预测结果有一定的准确性，能够帮助用户进行后续的决策等任务。

(2) 实用性：系统界面清晰明了，操作简洁，能够友好地交互。由于从数据读取到展示预测结果需要多个步骤和服务，无法保证快速响应，但应尽量减少系统的服务响应时间。

(3) 可扩展性：在未来的工作中，能够根据预测补全结果进行后续任务，如预测结果的可溯源分析等等。

5.2 概要设计

本系统基于 Python 的 Django 框架,采用 MTV 模式,即模型(Model)、模板(Template)和视图(View),数据库选择 SQLite 进行存储,前端使用 Bootstrap 开源工具包,构建起整个系统设计。

5.2.1 系统架构

本系统采用 Django 框架中典型的 MTV 架构：

(1) 模型 (Model)：采用 SQLite 数据库对生成的实体标签、预测标签结果进行存储。用户每一次发送生成实体画像或标签预测结果请求，系统就会从数据库中读取相应内容，显示到网页上。

(2) 模板 (Template)：以网页的形式展示可视化结果。前端采用 Bootstrap 框架，结合动态图形可视化的 JavaScript 程序库 D3.js 展示实体标签及预测结果。

(3) 视图 (View)：利用 JavaScript，与 Django 的 url 分发器结合完成页面与后台的交互，后台的数据处理由 python 完成。

5.2.2 系统设计

知识图谱补全系统生成预测结果的模块如图5.2所示，系统主要分为两大模块：实体画像标签生成模块和知识补全预测结果生成模块。

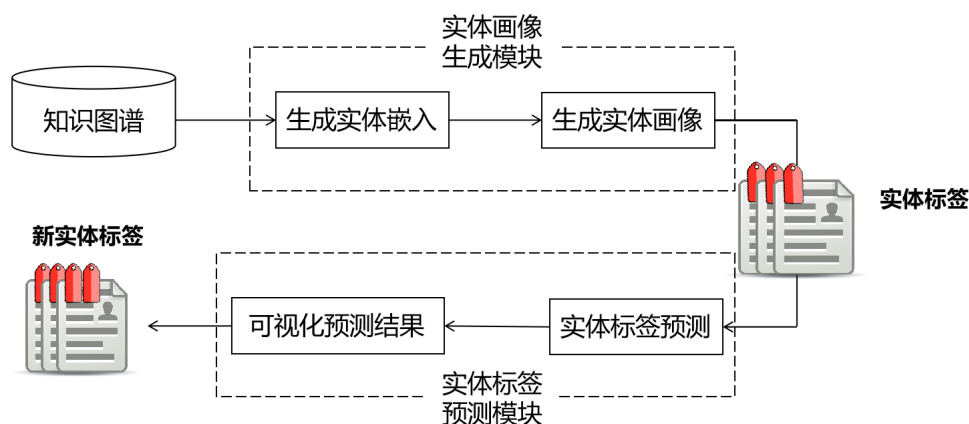


图 5.2: 知识图谱补全系统模块示意图

(1) 实体画像生成模块：该模块主要是利用基于网络角色发现的实体嵌入方法，通过计算区分度为每个实体生成对应的画像标签。其中，实体标签区分度计算的核心是利用本研究的 H、A、S 策略学习嵌入表示，得到的是实体嵌入包含多个层面的相似度信息。

在显示实体画像部分，通过实体名的模糊匹配，来搜索知识图谱中目标实体的画像标签结果，在数据库中存储与目标实体相关的关键字。这些关键字都来自于实体在知识图谱中的关系谓语 $\langle property : label \rangle$ 和 $\langle property : description \rangle$ 。例如，对于乐队实体 *Beastie_Boys*，它的关键字为 *Beastie_Boys*、*Boys* 和 *Beastie*。最终使用 D3.js 框架将生成的实体画像展示到网页前端，代码示例如下：

```

函数功能：根据输入的实体名称进行模糊查询画像结果
def search_keywords(localname):
    sql = "SELECT entity, keywords
          FROM entity_keywords
          WHERE entity_keywords MATCH '%" + localname + "%' LIMIT 15"
    return execute_query(sql)

```

(2) 实体标签预测模块：该模块主要是利用本研究的图自编码模型对实体标签进行预测。输入的是实体画像生成模块中的实体标签集合，最终将预测得到的新实体标签展示到网页上。

5.3 详细设计

知识图谱补全系统主要由实体画像生成模块和实体标签预测模块构成，系统的 Web 框架使用基于 Python 的 Django 框架，后端代码实现使用 Python，前端使用 Bootstrap 开源工具包。各模块的详细设计具体描述如下：

5.3.1 实体画像生成

输入：要查询画像结果的实体名。

输出：与目标实体名匹配的所有实体结果。根据匹配结果，显示该实体的画像标签结果和知识图谱中的原始描述信息。

流程：实体画像生成模块的处理流程如图5.3所示。输入要查询的实体名，在选定的数据集中进行模糊查询，在数据库中与实体相关的关键字进行匹配。将匹配得到的结果列表展示出来，并动态生成所选实体的画像结果和原始描述信息。

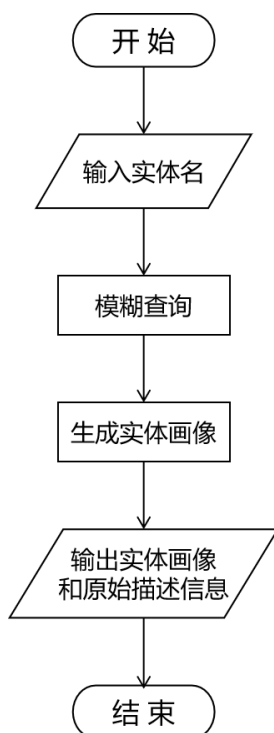


图 5.3: 实体画像生成流程图

主要接口函数：如表5.1所示，编程语言为 Python。

表 5.1: 实体画像生成模块主要接口函数

编号	函数
1	<code>def search_keywords(entity_name)</code> <code>entity_name</code> 为输入的要查询的实体名，进行模糊查询，返回目标实体结果。
2	<code>def init_entity_profiling(entity_name)</code> <code>entity_name</code> 为选择的目标实体名，在数据库中查找到对应的画像标签，返回目标实体的画像结果。

5.3.2 实体标签预测

输入：要查询预测结果的实体名。

输出：与目标实体名相关的标签预测结果及其对应得分。

流程：实体标签预测模块的处理流程如图5.4所示。输入要查询的实体名，在选定的数据集中查询到相关的标签预测结果，并且显示每个新标签的得分情况。

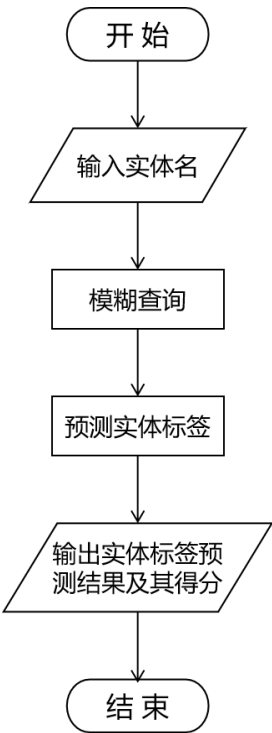


图 5.4: 实体标签预测流程图

主要接口函数：如表5.2所示，编程语言为 Python。

表 5.2: 实体标签预测模块主要接口函数

编号	函数
1	<code>def search_keywords(entity_name)</code> <code>entity_name</code> 为输入的要查询的实体名，进行模糊查询，返回目标实体结果。
2	<code>def init_label_prediction(entity_name)</code> <code>entity_name</code> 为选择的目标实体名，在数据库中查找到对应的新标签，返回实体标签预测结果及对应的得分情况。

5.4 系统演示

本研究按照上述的设计思路实现了基于网络角色的知识图谱补全系统，系统演示如下：

（1）主界面：显示 DBpedia 下子集 Band、University 和 Book 的统计信息，左侧导航栏中可以进行数据集的选择，如图5.5所示。

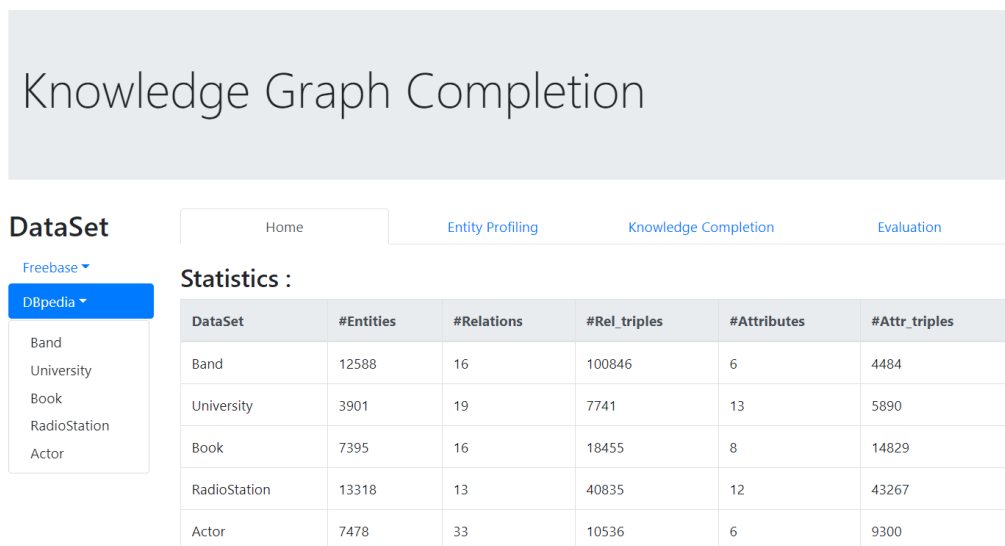


图 5.5: 系统主界面演示

（2）显示实体画像结果：点击菜单栏上的“Entity Profiling”，跳转到实体画像模块，如图5.6所示。用户可以在搜索栏中输入要查询的数据集中的实体名，能够动态地以星状结构显示目标实体的画像结果。

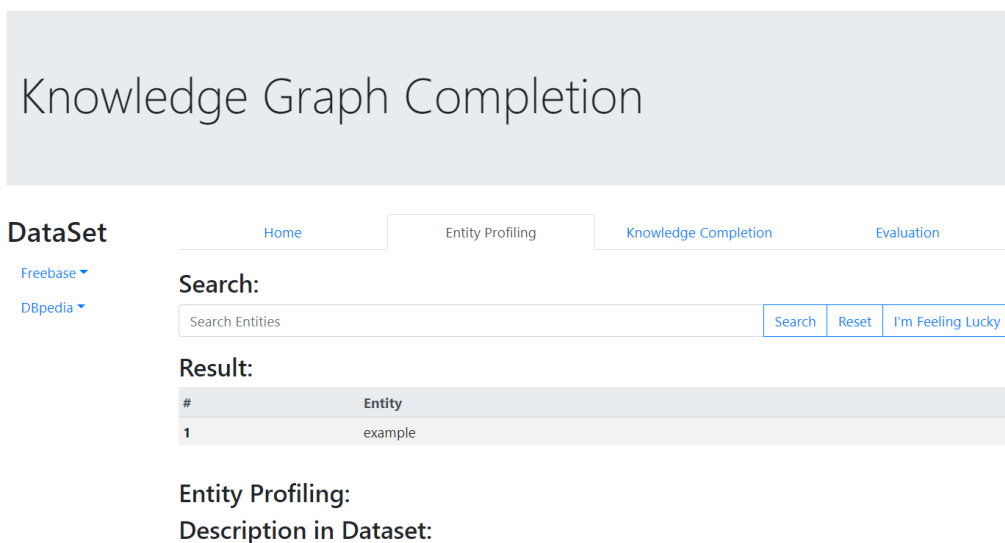


图 5.6: 实体画像模块演示

本系统也提供实体的模糊匹配查询，在图5.7中，输入要查询的实体名“Beastie”，选择“*Beastie_Boys*”实体，在下方显示对应实体的画像标签结果，如图5.8所示，包含了乐队活跃年份、乐队风格和所属国家等信息。点击实体的 URI 标识能够跳转到对应的 DBpedia 描述页面，显示其完整的描述信息，如图5.9所示。

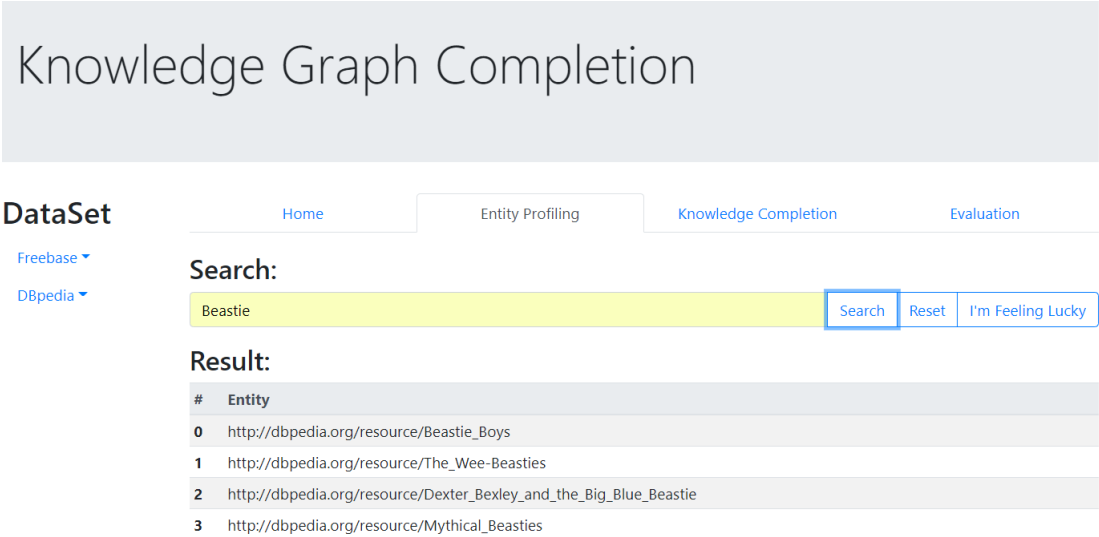


图 5.7: 模糊匹配查询实体

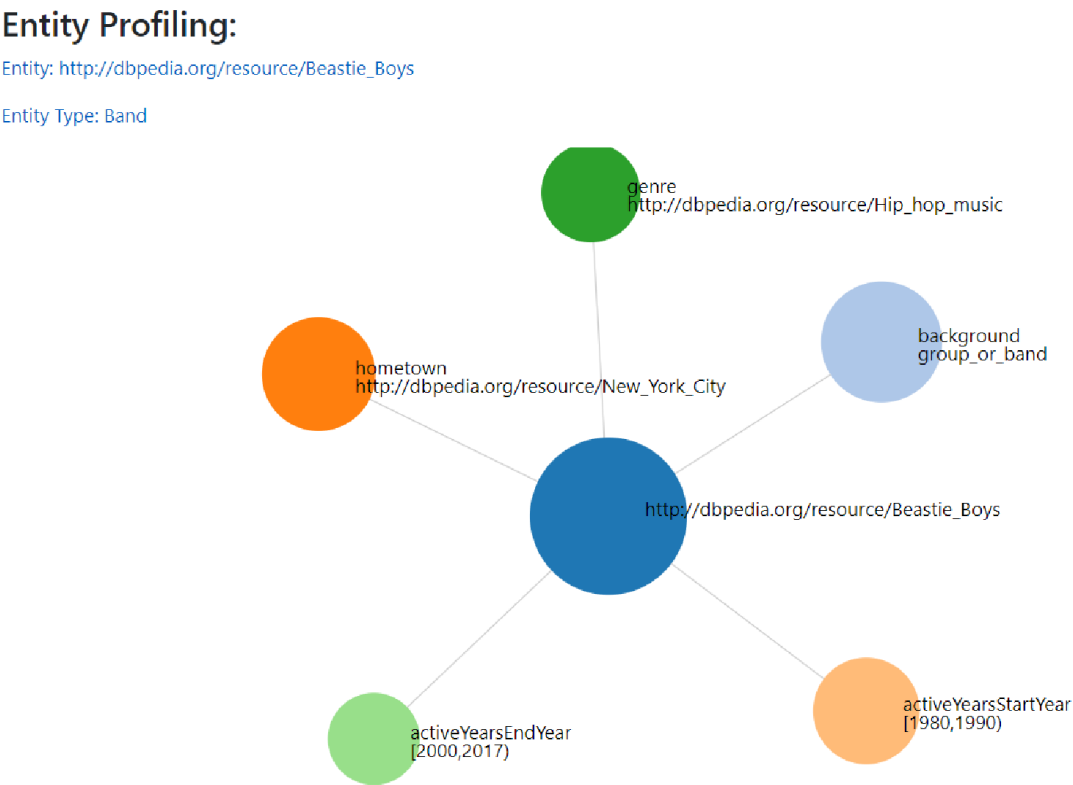
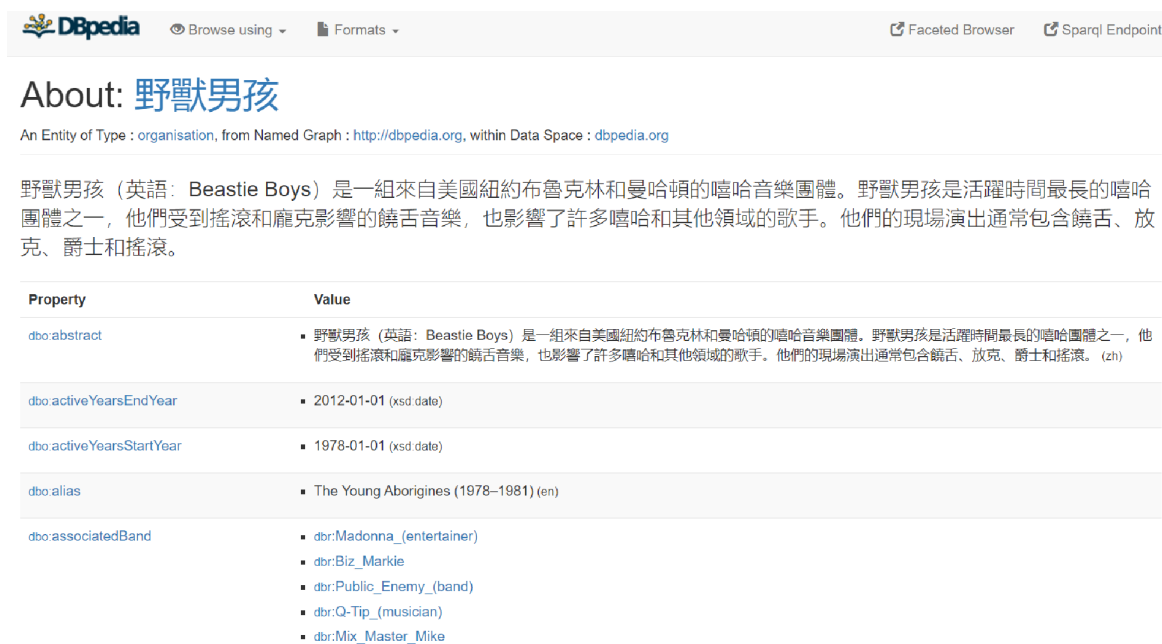


图 5.8: 实体画像结果



DBpedia

Browse using Formats Faceted Browser Sparql Endpoint

About: 野獸男孩

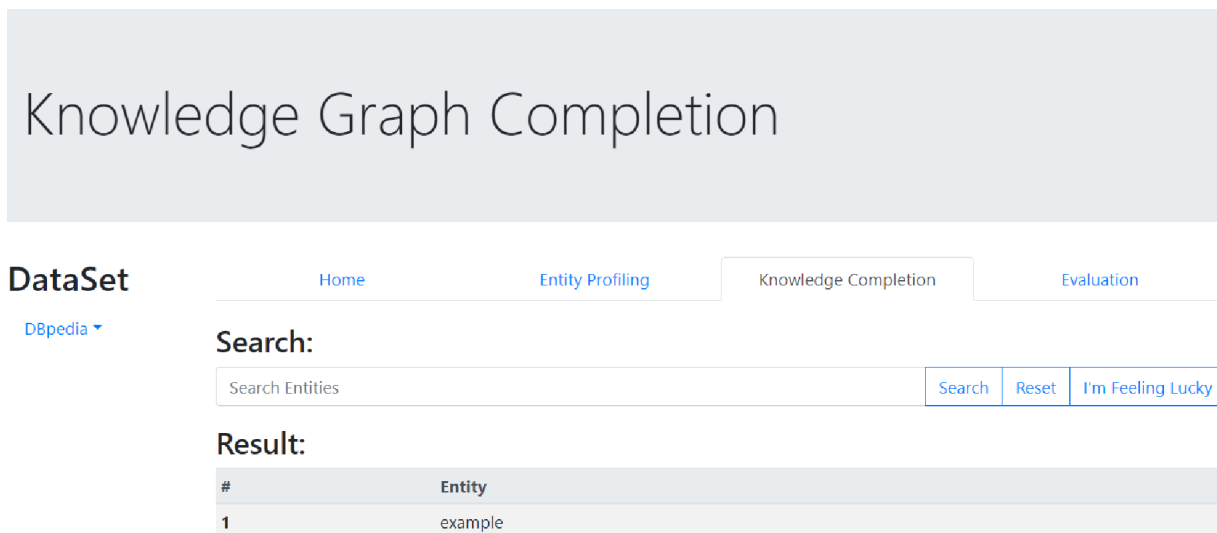
An Entity of Type : `organisation`, from Named Graph : `http://dbpedia.org`, within Data Space : `dbpedia.org`

野獸男孩（英語：Beastie Boys）是一組來自美國紐約布魯克林和曼哈頓的嘻哈音樂團體。野獸男孩是活躍時間最長的嘻哈團體之一，他們受到搖滾和龐克影響的饒舌音樂，也影響了許多嘻哈和其他領域的歌手。他們的現場演出通常包含饒舌、放克、爵士和搖滾。

Property	Value
<code>dbo:abstract</code>	<ul style="list-style-type: none">野獸男孩（英語：Beastie Boys）是一組來自美國紐約布魯克林和曼哈頓的嘻哈音樂團體。野獸男孩是活躍時間最長的嘻哈團體之一，他們受到搖滾和龐克影響的饒舌音樂，也影響了許多嘻哈和其他領域的歌手。他們的現場演出通常包含饒舌、放克、爵士和搖滾。 (zh)
<code>dbo:activeYearsEndYear</code>	<ul style="list-style-type: none">2012-01-01 (xsd:date)
<code>dbo:activeYearsStartYear</code>	<ul style="list-style-type: none">1978-01-01 (xsd:date)
<code>dbo:alias</code>	<ul style="list-style-type: none">The Young Aborigines (1978–1981) (en)
<code>dbo:associatedBand</code>	<ul style="list-style-type: none"><code>dbc:Madonna_(entertainer)</code><code>dbc:Biz_Markie</code><code>dbc:Public_Enemy_(band)</code><code>dbc:Q-Tip_(musician)</code><code>dbc:Mix_Master_Mike</code>

图 5.9: *Beastie_Boys* 在 DBpedia 中的详细描述信息

（3）显示实体标签预测结果：点击菜单栏上的“Knowledge Completion”，跳转到实体标签预测模块，如图5.10所示。用户可以在搜索栏中输入要查询的数据集中的实体名，将以实体标签的形式显示出标签预测的结果及其对应的得分。



Knowledge Graph Completion

DataSet

Home Entity Profiling Knowledge Completion Evaluation

DBpedia ▾

Search:

Search Entities Search Reset I'm Feeling Lucky

Result:

#	Entity
1	example

图 5.10: 实体标签预测模块演示

与实体画像生成模块类似，实体标签预测模块也提供实体的模糊匹配查询。输入的要查询的实体名“*Beastie_Boys*”，显示搜索结果，在页面下方显示出标签预测结果和得分情况，如图5.11所示。

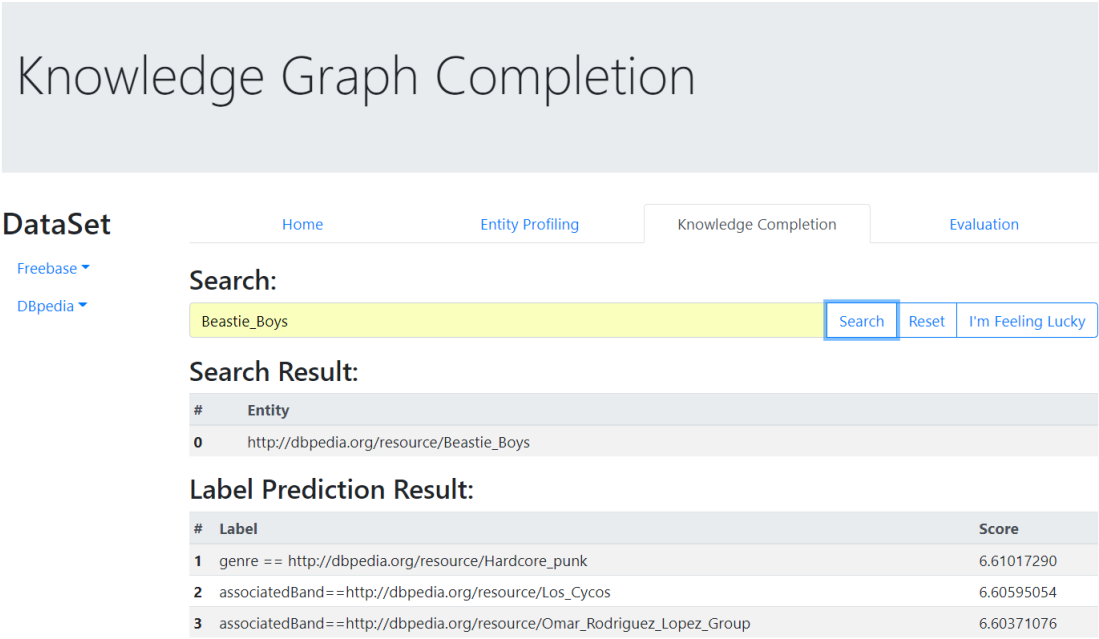


图 5.11: 标签预测结果

5.5 本章小结

本章设计并实现了基于网络角色的知识图谱补全系统，详细描述了该系统的需求分析、概要设计和系统演示。用户可以在该系统中了解实体的画像结果，并对实体标签预测结果进行自我评估，提高可理解性。

第六章 总结与展望

本章对基于网络角色的知识图谱缺失知识补全方法的研究工作做出总结和概括，并且结合在研究和实验过程中遇到的问题对未来的工作提出展望。

6.1 工作总结

知识图谱作为许多人工智能技术的基础，一直得到学术界和工业界的广泛关注和應用。但是，随着数据量的动态增长，知识图谱始终存在着数据稀疏的问题，知识图谱中缺失知识的补全问题亟待解决。本研究的知识补全方法主要涉及到两块相关的内容：知识补全方法和表示学习方法。现有的知识补全方法大多是基于表示学习的，所用的表示学习方法都是依赖于实体间的关系，这使得所学习的实体嵌入质量受关系稀疏性影响较大。故针对这些问题提出一种基于网络角色的知识图谱缺失知识补全方法，本研究的工作总结如下：

(1) 提出一种基于网络角色的知识图谱实体嵌入方法。根据实体间不同层面的相似等效性，从同质性、属性相似性、结构相似性三个角度来进行实体角色发现，生成不同的实体路径，解决了表示学习受知识图谱中关系稀疏影响较大的问题。除这三种实体相似性外，可以利用更多角度的相似性来进行角色发现，实现进一步的扩展。最终，利用 skip-gram 得到语义丰富的实体嵌入表示。

(2) 提出一种基于实体标签的知识补全方法。将实体标签作为预测模型的输入输出，用户通过少量的实体标签就可以快速理解预测结果，解决了可理解性问题。同时，将连续值属性离散化，属性预测的结果不再是一个确切的数值，而是一个值域范围，从另一种角度提高了预测结果的准确度，同时也能够满足用户对属性预测的需求。

(3) 设计并实现基于网络角色的知识图谱补全系统。根据选择的数据集，能够搜索相应的实体，动态显示知识图谱中的实体画像结果。并且，能够显示标签预测结果及其相应的排名情况。基于网络角色发现的实体嵌入表示可以用于实体画像标签的度量计算环节，展示高质量的实体标签。同时，实体嵌入表示能够用于初始化补全模型，实体标签作为输入输出，提高模型的性能和结果的可理解性。

综上所述，本研究提出了对于知识图谱中缺失知识补全的思路和方法，描述了基于网络角色的知识图谱缺失知识补全方法的模型框架和具体细节，并通过实验证明了本研究方法的可行性和有效性，以可视化系统的形式展示了实验成果。

6.2 未来展望

本研究提出的基于网络角色的知识图谱缺失知识补全方法工作已基本完成，并且同时实验证明了其可行性和有效性。但是，仍然存在以下可供继续深入研究的关键点：

（1）知识图谱的结构中仍有许多信息可供挖掘用于表示学习。例如，在网络角色发现中，结构洞理论、**motif** 特征等也可以作为实体的特征进行实体向量表示的学习，扩展了基于网络角色发现的实体嵌入方法，丰富实体嵌入表示的特征。

（2）学习自适应的路径建模方式。本研究在基于网络角色发现是实体嵌入方法中，考虑到数据集在属性特征和关系特征上的特点，进行一种有偏的路径混合方式。在后续的工作中，能够使模型能够针对数据集自动化的调节权重参数。

（3）利用 **attention** 机制提高补全模型的性能。在补全模型部分，本研究使用 GCN 作为图自编码模型的 **encoder** 部分，对于节点的邻域表示的聚集采用统一的归一化参数，忽略了标签对于实体的重要性。在后续的工作中，可以使用引入了 **attention** 机制的 GAT 模型，自动化地赋予实体邻居不同的权重，提供模型的精度。

（4）提高知识补全结果的可解释性。目前的深度学习模型在可解释性上的研究分析仍较少，使得该类模型应用到工程中仍有一定的安全隐患。而基于规则的方法可解释性强，但在大规模数据上效率较低，无法应用到生活中。针对该问题，可以对深度学习模型得到的预测结果进行溯源分析，向用户展示得到该结果的具体原因，以提高可解释性。

致谢

三年的硕士研究生时光转瞬即逝，在这三年期间，我得到了许多老师、同学和朋友们的关怀和帮助。在学术论文即将完成之际，我要向所有给予我支持、帮助和鼓励的人表示我最诚挚的谢意。

首先，我要衷心地感谢我的导师张祥老师。研究生入校以来，张老师始终在学习工作上给予我悉心的指导，在生活上给予我恰当的关心。在我参与的每一个项目中，张老师会仔细地为我分析问题，互相探讨如何解决困难。在这一次次的项目经历中，我的技术水平和写作水平也得到了很大提升。张老师的敬业精神，渊博的知识和严谨的治学态度都深深地感染并激励着我。非常感谢这三年老师对我的谆谆教诲，三年来学到的深入研究问题的方法将使我终生受益。

其次，我要感谢实验室的李慧颖老师。每次组会上，李老师都会针对科研方法和研究领域的相关问题展开讲解，帮助我们学习科研的方法，指出科研中需要关注的关键点。李老师在科研上的态度深深的影响着我，科研无止境，学习也不能停止。感谢三年来对我的帮助。

感谢师兄师姐们在科研中对我的悉心指导。在项目中，杨清清师姐和丁金如师姐始终耐心地指导我的工作，认真解答我的疑问，针对研究问题发表个人看法，帮助我开阔了科研思路。感谢师弟师妹们，帮助我完成项目实验的评估，大家一起通力合作，完成科研项目。很幸运这三年的研究生生活中有同门们的陪伴，正因为你们为实验室营造了良好的科研氛围，使我在与大家的相处过程中得到了成长，也收获了快乐。

感谢我的室友们，感谢这三年的互相关心与照顾。从青教搬到桃园，一路走来，我们始终相伴。在生活和学习工作中，大家共同努力，互相支持，每次在寝室总能收获满满的快乐。三年的时光太匆匆，可惜我们相识的太晚，相伴的时间太短，愿我们的情谊永不变。

感谢我的家人们，我的每一个决定都离不开他们的支持和付出，正是他们在身后支持着我，让我能够勇敢地去外面的世界闯荡，感受不一样的生活。许多感情无以言表，祝愿他们平安健康，开心快乐。

最后，感谢东南大学，我终生以母校为荣！

参考文献

- [1] Bordes, Antoine, Usunier, Nicolas, García-Durán, Alberto, et al. Translating Embeddings for Modeling Multi-relational Data[C]. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. 2013. 2787–2795.
- [2] Wang, Zhen, Zhang, Jianwen, Feng, Jianlin, et al. Knowledge Graph Embedding by Translating on Hyperplanes[C]. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI Press, 2014. 1112–1119.
- [3] Nickel, M., Tresp, V., and Kriegel, H. P. A Three-Way Model for Collective Learning on Multi-Relational Data[C]. In: International Conference on International Conference on Machine Learning. 2011.
- [4] Tay, Yi, Tuan, Luu Anh, Phan, Minh C., et al. Multi-Task Neural Network for Non-discrete Attribute Prediction in Knowledge Graphs[C]. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017. ACM, 2017. 1029–1038.
- [5] Zheng, Liang, Qu, Yuzhong, Qian, Xinqi, et al. A hierarchical co-clustering approach for entity exploration over Linked Data[J]. Knowl. Based Syst., 2018, 141:200–210.
- [6] Nentwig, Markus, Groß, Anika, and Rahm, Erhard. Holistic Entity Clustering for Linked Data[C]. In: IEEE International Conference on Data Mining Workshops, ICDM Workshops 2016. IEEE Computer Society, 2016. 194–201.
- [7] Auer, Sören, Bizer, Christian, Kobilarov, Georgi, et al. DBpedia: A Nucleus for a Web of Open Data[C]. In: The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007. Springer, 2007. 722–735.
- [8] Luczkovich, J. J., Borgatti, S. P., Johnson, J. C., et al. Defining and Measuring Trophic Role Similarity in Food Webs Using Regular Equivalence[J]. Journal of Theoretical Biology, 2003, 220(3):303–321.
- [9] Hafner-Burton, E. M., Kahler, M., and Montgomery, A. H. Network Analysis for International Relations[J]. International Organization, 2009, 63(3):559–592.

-
- [10] Holme, P. and Huss, M. Role-similarity based functional prediction in networked systems: Application to the yeast proteome[J]. *Journal of the Royal Society Interface*, 2005, 2(4).
- [11] Gilpin, Sean, Eliassi-Rad, Tina, and Davidson, Ian N. Guided learning for role discovery (GLRD): framework, algorithms, and applications[C]. In: *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013*. ACM, 2013. 113–121.
- [12] Airoldi, Edoardo M., Blei, David M., Fienberg, Stephen E., et al. Mixed Membership Stochastic Blockmodels[C]. In: *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*. Curran Associates, Inc., 2008. 33–40.
- [13] Brandes, Ulrik and Lerner, Jürgen. Structural Similarity: Spectral Methods for Relaxed Blockmodeling[J]. *J. Classif.*, 2010, 27(3):279–306.
- [14] Kleinberg, Jon M. Authoritative Sources in a Hyperlinked Environment[J]. *J. ACM*, 1999, 46(5):604–632.
- [15] Rossi, Ryan A. and Ahmed, Nesreen K. Role Discovery in Networks[J]. *IEEE Trans. Knowl. Data Eng.*, 2015, 27(4):1112–1131.
- [16] Henderson, Keith, Gallagher, Brian, Li, Lei, et al. It’s who you know: graph mining using recursive structural features[C]. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2011. 663–671.
- [17] Henderson, Keith, Gallagher, Brian, Eliassi-Rad, Tina, et al. RolX: structural role extraction & mining in large graphs[C]. In: *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*. ACM, 2012. 1231–1239.
- [18] Gilpin, Sean, Eliassi-Rad, Tina, and Davidson, Ian N. Guided learning for role discovery (GLRD): framework, algorithms, and applications[C]. In: *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013*. ACM, 2013. 113–121.
- [19] Landwehr, Niels, Kersting, Kristian, and Raedt, Luc De. nFOIL: Integrating Naïve Bayes and FOIL[C]. In: *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference*. AAAI Press / The MIT Press, 2005. 795–800.

- [20] Rossi, Ryan A. and Neville, Jennifer. Time-Evolving Relational Classification and Ensemble Methods[C]. In: *Advances in Knowledge Discovery and Data Mining - 16th Pacific-Asia Conference, PAKDD 2012*. Springer, 2012. 1–13.
- [21] Friedlander, Michael P. and Hatz, Kathrin. Computing non-negative tensor factorizations[J]. *Optim. Methods Softw.*, 2008, 23(4):631–647.
- [22] Yilmaz, Yusuf Kenan and Cemgil, A. Taylan. Probabilistic Latent Tensor Factorization[C]. In: *Latent Variable Analysis and Signal Separation - 9th International Conference, LVA/ICA 2010*. Springer, 2010. 346–353.
- [23] Singh, Ajit Paul and Gordon, Geoffrey J. Relational learning via collective matrix factorization[C]. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2008. 650–658.
- [24] Chan, Jeffrey, Lam, Samantha, and Hayes, Conor. Increasing the Scalability of the Fitting of Generalised Block Models for Social Networks[C]. In: *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. IJCAI/AAAI, 2011. 1218–1224.
- [25] Lin, Yankai, Liu, Zhiyuan, Sun, Maosong, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion[C]. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press, 2015. 2181–2187.
- [26] Nickel, Maximilian, Rosasco, Lorenzo, and Poggio, Tomaso A. Holographic Embeddings of Knowledge Graphs[C]. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 2016. 1955–1961.
- [27] Trouillon, Théo, Welbl, Johannes, Riedel, Sebastian, et al. Complex Embeddings for Simple Link Prediction[C]. In: *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016*. JMLR.org, 2016. 2071–2080.
- [28] Shi, Baoxu and Weninger, Tim. ProjE: Embedding Projection for Knowledge Graph Completion[C]. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI Press, 2017. 1236–1242.
- [29] Shen, Yelong, Huang, Po-Sen, Chang, Ming-Wei, et al. Modeling Large-Scale Structured Relationships with Shared Memory for Knowledge Base Completion[C]. In: *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017*. Association for Computational Linguistics, 2017. 57–68.

- [30] Liu, Hanxiao, Wu, Yuexin, and Yang, Yiming. Analogical Inference for Multi-relational Embeddings[C]. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017. PMLR, 2017. 2168–2178.
- [31] Guan, Saiping, Jin, Xiaolong, Wang, Yuanzhuo, et al. Shared Embedding Based Neural Networks for Knowledge Graph Completion[C]. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018. ACM, 2018. 247–256.
- [32] Shi, Baoxu and Weninger, Tim. Open-World Knowledge Graph Completion[C]. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. AAAI Press, 2018. 1957–1964.
- [33] Chen, Wenhui, Xiong, Wenhui, Yan, Xifeng, et al. Variational Knowledge Graph Reasoning[C]. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018. Association for Computational Linguistics, 2018. 1823–1832.
- [34] Xiong, Wenhui, Hoang, Thien, and Wang, William Yang. DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning[C]. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017. Association for Computational Linguistics, 2017. 564–573.
- [35] Das, Rajarshi, Dhuliawala, Shehzaad, Zaheer, Manzil, et al. Go for a Walk and Arrive at the Answer: Reasoning Over Paths in Knowledge Bases using Reinforcement Learning[C]. In: 6th International Conference on Learning Representations, ICLR 2018. Open-Review.net, 2018.
- [36] Lin, Xi Victoria, Socher, Richard, and Xiong, Caiming. Multi-Hop Knowledge Graph Reasoning with Reward Shaping[C]. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2018. 3243–3253.
- [37] Shen, Yelong, Chen, Jianshu, Huang, Po-Sen, et al. M-Walk: Learning to Walk over Graphs using Monte Carlo Tree Search[C]. In: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018. 2018. 6787–6798.
- [38] Fu, Cong, Chen, Tong, Qu, Meng, et al. Collaborative Policy Learning for Open Knowledge Graph Reasoning[C]. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural

- Language Processing, EMNLP-IJCNLP 2019. Association for Computational Linguistics, 2019. 2672–2681.
- [39] Lin, Yankai, Liu, Zhiyuan, Luan, Huan-Bo, et al. Modeling Relation Paths for Representation Learning of Knowledge Bases[C]. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015. The Association for Computational Linguistics, 2015. 705–714.
- [40] Lao, Ni and Cohen, William W. Relational retrieval using a combination of path-constrained random walks[J]. Mach. Learn., 2010, 81(1):53–67.
- [41] Gardner, Matt, Talukdar, Partha Pratim, Krishnamurthy, Jayant, et al. Incorporating Vector Space Similarity in Random Walk Inference over Knowledge Bases[C]. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, 2014. 397–406.
- [42] Neelakantan, Arvind, Roth, Benjamin, and McCallum, Andrew. Compositional Vector Space Models for Knowledge Base Completion[C]. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015. The Association for Computer Linguistics, 2015. 156–166.
- [43] Das, Rajarshi, Neelakantan, Arvind, Belanger, David, et al. Chains of Reasoning over Entities, Relations, and Text using Recurrent Neural Networks[C]. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017. Association for Computational Linguistics, 2017. 132–141.
- [44] Goethals, Bart and den Bussche, Jan Van. Relational Association Rules: Getting WARMER[C]. In: Pattern Detection and Discovery, ESF. Springer, 2002. 125–139.
- [45] Galárraga, Luis Antonio, Teflioudi, Christina, Hose, Katja, et al. AMIE: association rule mining under incomplete evidence in ontological knowledge bases[C]. In: 22nd International World Wide Web Conference, WWW '13. International World Wide Web Conferences Steering Committee / ACM, 2013. 413–422.
- [46] Guo, Shu, Wang, Quan, Wang, Lihong, et al. Knowledge Graph Embedding With Iterative Guidance From Soft Rules[C]. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. AAAI Press, 2018. 4816–4823.

- [47] Zhang, Wen, Paudel, Bibek, Wang, Liang, et al. Iteratively Learning Embeddings and Rules for Knowledge Graph Reasoning[C]. In: The World Wide Web Conference, WWW 2019. ACM, 2019. 2366–2377.
- [48] Rocktäschel, Tim and Riedel, Sebastian. End-to-end Differentiable Proving[C]. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. 2017. 3788–3800.
- [49] Yang, Fan, Yang, Zhilin, and Cohen, William W. Differentiable Learning of Logical Rules for Knowledge Base Reasoning[C]. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. 2017. 2319–2328.
- [50] Qu, Meng and Tang, Jian. Probabilistic Logic Neural Networks for Reasoning[C]. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019. 2019. 7710–7720.
- [51] Zhang, Yuyu, Chen, Xinshi, Yang, Yuan, et al. Efficient Probabilistic Logic Reasoning with Graph Neural Networks[C]. In: 8th International Conference on Learning Representations, ICLR 2020. OpenReview.net, 2020.
- [52] Trivedi, Rakshit, Dai, Hanjun, Wang, Yichen, et al. Know-Evolve: Deep Reasoning in Temporal Knowledge Graphs[J]. CoRR, 2017, abs/1705.05742.
- [53] Trivedi, R., Farajtabar, M., Wang, Y., et al. Know-Evolve: Deep Reasoning in Temporal Knowledge Graphs[J]. 2017.
- [54] García-Durán, Alberto, Dumancic, Sebastijan, and Niepert, Mathias. Learning Sequence Encoders for Temporal Knowledge Graph Completion[C]. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2018.
- [55] Goel, Rishab, Kazemi, Seyed Mehran, Brubaker, Marcus, et al. Diachronic Embedding for Temporal Knowledge Graph Completion[C]. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020. AAAI Press, 2020. 3988–3995.
- [56] Chekol, Melisachew Wudage, Pirrò, Giuseppe, Schoenfish, Joerg, et al. Marrying Uncertainty and Time in Knowledge Graphs[C]. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI Press, 2017. 88–94.

- [57] Omran, Pouya Ghiasnezhad, Wang, Kewen, and Wang, Zhe. An Embedding-Based Approach to Rule Learning in Knowledge Graphs[J]. *IEEE Trans. Knowl. Data Eng.*, 2021, 33(4):1348–1359.
- [58] Wijaya, Derry Tanti, Nakashole, Ndapandula, and Mitchell, Tom M. CTPs: Contextual Temporal Profiles for Time Scoping Facts using State Change Detection[C]. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. 1930–1936.
- [59] Trivedi, Rakshit, Dai, Hanjun, Wang, Yichen, et al. Know-Evolve: Deep Temporal Reasoning for Dynamic Knowledge Graphs[C]. In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*. PMLR, 2017. 3462–3471.
- [60] Dougherty, James, Kohavi, Ron, and Sahami, Mehran. Supervised and Unsupervised Discretization of Continuous Features[C]. In: *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, 1995. 194–202.
- [61] Xing-Sheng, L. I. and De-Yi, L. I. A New Method Based on Density Clustering for Discretization of Continuous Attributes[J]. *Acta Simulata Systematica Sinica*, 2003.
- [62] Perozzi, Bryan, Al-Rfou, Rami, and Skiena, Steven. DeepWalk: online learning of social representations[C]. In: *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2014. 701–710.
- [63] Grover, Aditya and Leskovec, Jure. node2vec: Scalable Feature Learning for Networks[C]. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016. 855–864.
- [64] Goldberg, Yoav and Levy, Omer. word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method[J]. *CoRR*, 2014, abs/1402.3722.
- [65] Li, Furong, Lee, Mong-Li, and Hsu, Wynne. Entity profiling with varying source reliabilities[C]. In: *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’14*. ACM, 2014. 1146–1155.
- [66] Rybak, Jan, Balog, Krisztian, and Nørnvåg, Kjetil. ExperTime: tracking expertise over time[C]. In: *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’14*. ACM, 2014. 1273–1274.
- [67] Katz, L. A new status index derived from sociometric analysis[J]. *Psychometrika*, 1953, 18(1):39–43.

- [68] Jeh, Glen and Widom, Jennifer. SimRank: a measure of structural-context similarity[C]. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002. 538–543.
- [69] Zhao, Peixiang, Han, Jiawei, and Sun, Yizhou. P-Rank: a comprehensive structural similarity measure over information networks[C]. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009. ACM, 2009. 553–562.
- [70] Kipf, Thomas N. and Welling, Max. Variational Graph Auto-Encoders[J]. CoRR, 2016, abs/1611.07308.
- [71] Socher, Richard, Chen, Danqi, Manning, Christopher D., et al. Reasoning With Neural Tensor Networks for Knowledge Base Completion[C]. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. 2013. 926–934.
- [72] Yang, Bishan, Yih, Wen-tau, He, Xiaodong, et al. Embedding Entities and Relations for Learning and Inference in Knowledge Bases[C]. In: 3rd International Conference on Learning Representations, ICLR 2015. 2015.
- [73] Bruna, Joan, Zaremba, Wojciech, Szlam, Arthur, et al. Spectral Networks and Locally Connected Networks on Graphs[C]. In: 2nd International Conference on Learning Representations, ICLR 2014. 2014.
- [74] Henaff, Mikael, Bruna, Joan, and LeCun, Yann. Deep Convolutional Networks on Graph-Structured Data[J]. CoRR, 2015, abs/1506.05163.
- [75] Kipf, Thomas N. and Welling, Max. Semi-Supervised Classification with Graph Convolutional Networks[C]. In: 5th International Conference on Learning Representations, ICLR 2017. OpenReview.net, 2017.
- [76] Wang, Hongwei, Zhao, Miao, Xie, Xing, et al. Knowledge Graph Convolutional Networks for Recommender Systems[C]. In: The World Wide Web Conference, WWW 2019. ACM, 2019. 3307–3313.
- [77] Miller, Alexander H., Fisch, Adam, Dodge, Jesse, et al. Key-Value Memory Networks for Directly Reading Documents[C]. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016. The Association for Computational Linguistics, 2016. 1400–1409.

作者攻读硕士学位期间的研究成果

王紫悦（1995-），女，汉族，江苏南京人。本科在南京农业大学人工智能学院就读，专业为计算机科学与技术。现为东南大学网络安全学院硕士研究生，研究方向为知识图谱。

• 攻读硕士学位期间发表的论文

- [1] Xiang Zhang, Qingqing Yang, Jinru Ding, **Ziyue Wang**. Entity Profiling in Knowledge Graphs[J]. IEEE Access 8: 27257-27266 (2020).
- [2] 张祥, **王紫悦**, 杨清清, 等. 基于知识图谱的实体标签可视化 [J]. 指挥信息系统与技术, 2020, 11(3): 1-9.

• 参与科研项目情况

- [1] 科技部重点研发计划: 国家中心城市数据管控与知识萃取技术和系统应用(2019.12-2022.11, 项目编号: 2019YFB2101800)

心於至善

