计算机科学
COMPUTER SCIENCE

在知识图谱实体关系预测中对DistMult解码器的优化研究

韩以健，王宝会

引用本文

韩以健，王宝会. 在知识图谱实体关系预测中对DistMult解码器的优化研究[J]. 计算机科学，2024，51(11A)：231200118-5.

HAN Yijian, WANG Baohui. Study on DistMult Decoder in Knowledge Graph Entity Relationship Prediction [J]. Computer Science, 2024, 51(11A): 231200118-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231200118

在知识图谱实体关系预测中对 DistMult 解码器的优化研究

韩以健　王宝会
北京航空航天大学软件学院　北京 100191
(hyj1872@buaa.edu.cn)

摘　要　国家电网甘肃电力科学院希望通过大量科研文献构建电力行业知识图谱,并深度挖掘知识图谱中的的潜在关联。关系预测模型是解决这类问题的关键技术,也是知识图谱中的重要技术,是近年来科研工作者的研究热点。大量论文和实验已经证明使用编码器加解码器组合的框架在关系预测任务中有不错的表现。在这种框架下,由于图神经网络技术的进步,近年来有有不少工作通过以图神经网络为编码器并加以优化的方案来提升关系预测的效果,而忽略了解码器的作用。受到余弦相似度的启发,提出了基于 DistMult 的新型解码器 COS-DistMult,并在真实的数据集上进行对比实验。实验结果表明,关系预测模型的评价指标 Hits@10 的值提高了 2%左右,证明在以编码器加解码器为框架的关系预测任务中,优化解码器结构是一种行之有效的方法。

关键词:知识图谱;图神经网络;关系预测;DistMult 解码器
中图分类号　TP301

Study on DistMult Decoder in Knowledge Graph Entity Relationship Prediction

HAN Yijian and WANG Baohui
School of Software,Beihang University,Beijing 100191,China

Abstract　State Grid Gansu Electric Power Academy hopes to construct a knowledge graph of the power industry through a large amount of scientific research literature and deeply explore the potential correlations in the knowledge graph. The relationship prediction model is a key technology for solving such problems and an important technology in knowledge graphs,which has been a research hotspot for researchers in recent years. A large number of papers and experiments have demonstrated that the framework combining encoder and decoder performs well in relation prediction tasks. Under this framework,due to the advancement of graph neural network technology,there have been many works in recent years that have improved the performance of relationship prediction by using graph neural networks as encoders and optimizing them,while neglecting the role of decoders. Taking inspiration from cosine similarity,this paper proposes a novel decoder COS DistMult based on DistMult and conducts comparative experiments on real datasets,and the experimental results indicate that the evaluation indicator Hits@10of the relationship prediction model increases by 2%. It is proves that optimizing the decoder structure is an effective method in relation prediction tasks based on an encoder decoder framework.

Keywords　Knowledge graph,Graph neural network,Relation prediction,DistMult decoder

1　引言

知识图谱的概念最早由 Google 公司提出,他们建立并整合各种来源信息的知识库,称之为"知识图谱"。知识图谱是由实体和关系所组成的一个网络,其中实体对应网络中的节点,而关系可以理解为网络中不同类型的边。知识图谱已被证明对知识管理和数据分析具有革命性意义。

许多学者在构建知识图谱时,发现知识实体之间的关联信息不够充分,图谱存在稀疏性问题,没有充分揭示出数据集内的潜在信息。为了解决上述问题,深层次地挖掘图谱中的潜在关联信息,我们将侧重于研究知识图谱关系预测技术,它是解决这类问题的关键技术,也是知识图谱中的重要技术。

知识图谱的关系预测技术是从已知事实或知识中抽取并推断得出新事实和知识的过程,挖掘两个知识实体之间潜在的关联信息,这个过程或方法也可以称为链接预测。它可以扩充实体之间的链接,使图谱更完整,更有使用价值。关系预测技术也是当下知识图谱领域的热门研究方向。

近年来,国家电网甘肃电力公司快速发展,积极地完成了信息化的过程,积累了大量的企业内部文献数据等。这些数据以非结构化形式存储,不利于知识挖掘,难以得到高效利用。为了解决这些问题,建立高效的知识检索系统,立项开展了面向电力行业科技信息管理的知识图谱框架设计与构建工作。因此本文利用知识图谱关系预测技术来丰富知识表示,完善电力行业知识图谱,给研究者查阅文献资料提供便利。

早期的关系预测技术采用平移距离模型(TRANSE[1],TRANSR[2]等)或语义匹配模型(RESCAL[3],DistMult[4],HolE[5]等)来衡量知识事实三元组的语义距离,以三元组得分值为依据,判断其成立的概率。

Kipf 等[6]在 2017 年提出了图神经网络。此后,在知识表示学习领域,基于图神经网络方法已经逐渐成为了主流。

通信作者:王宝会(wangbh@buaa.edu.cn)

相较于传统的表示方法，图神经网络以其对图结构的建模能力和高效的图表示学习效果而备受青睐。

Michael 等[7]在 2018 年提出关系图卷积神经网络算法 R-GCN。它在图卷积的基础上进行了扩展，对不同类型的边使用不同的权重矩阵，让神经网络能够学习到异构图的拓扑结构。

由于知识图谱往往也属于异构图，因此知识图谱关系预测技术相关的许多工作采用编码器加解码器结构的网络模型，并取得了良好的性能表现。这进一步证明了编码器和解码器结构的有效性。其结构如图 1 所示。

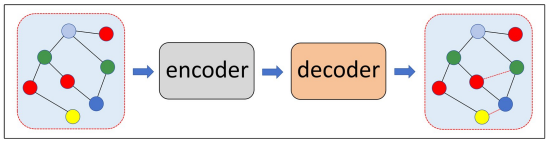

图 1 编码器和解码器框架

Fig. 1 Encoder and decoder framework

关于评价指标，由于传统的深度学习评价指标 AUC 等在关系预测任务中非常容易失真，往往以很快的速度就收敛于 99.9%，所以关系预测任务往往采用 hit@N 作为评价指标。其大致意思是对于每一个三元组样本，通过随机替换掉它的头顶点或是尾顶点的方式产生很多个负样本。对所有样本进行打分，并对这些打分结果进行升序排序。Hits@10，表示在所有样本结果中，计算正样本的得分排名前 10 的占比。Hits@1，Hits@5 同理。MRR 则表示所有正样本在各自打分结果中的排名的平均值。如果 MRR 和 Hits@10 得分高，就认为模型的效果是不错的。而对于打分结果比正样本还高的负样本，可以将其作为新的三元组关系添加到图谱中。这也是关系预测任务的目标。MRR 和 Hits@10 是知识图谱关系预测任务的常用指标，大量研究者都采用了这样的评价指标，本文也效仿之。

为了进一步提升关系预测的效果，不少研究通过各种方式设计或优化模型且取得了不错的效果。

Jin[8]提出了一种基于关系感知的时间嵌入方法（Relationaware Temproal Embedding，RTE），并提出 RTE 与 DistMult 和 SimplE 的融合机制，其在公开数据集上取得了较好的实验结果。

Xue 等[9]构建电力营销系统问答机器人时，针对三元组中的头实体、尾实体及约束关系建立了结构相同的 3 个神经网络，提出了三分支并行神经网络（TBPNN），降低了 MeanRank，但提升了 Hit@10 指标。

Shan 等[10]将实体类型和邻域信息编码为先验概率，将实例信息编码为似然概率，且按照贝叶斯规则将二者组合，并提出了一种基于贝叶斯规则的具有层次注意力的关系预测方法，其在 FB15k-237 上提升了 MRR 和 hits@10。

Chen 等[11]利用 Tucker 分解将三阶张量表示的知识图谱分解成一个核心张量与每个 mode 上因子矩阵的乘积的形式，提出了一种改进的 Tucker 分解知识图谱关系预测算法。

Wang 等[12]使用图注意力捕获每个实体邻域中的实体和关系特征，引入胶囊神经网络来解码三元组，通过胶囊神经网络节点嵌入特征的学习，生成连续向量与权重向量做点积运算，提出了一种融合图注意力网络和胶囊神经网络的知识图谱链接预测模型。

Pang 等[13]提出了一种基于注意力与卷积网络的链接预测方法（LPACN），采用改进的注意力机制将实体的注意力信息融入到关系嵌入中；并且将同元组内相邻实体个数信息融入卷积网络，进一步补足了实体卷积向量的信息含量。

上述研究通过优化编码器来实现更准确的知识表示，进而达到提升关系预测效果的目的。但鲜有人通过优化解码器的方式去提升知识图谱关系预测效果。因此，本文重点研究基于 DistMult 解码器的优化方法，受到向量余弦相似度的启发，提出了新型解码器 COS-DistMult（基于余弦相似度的 DistMult），其取得了较好的表现。

2 问题提出与模型研究

知识图谱关系预测任务的流程图如图 2 所示。在流程中，解码器负责衡量编码器给出的向量之间的相似度。相似度是损失函数优化的目标，也是评价指标判断三元组存在概率的依据。所以解码器在模型中起着承上启下，有至关重要的作用。

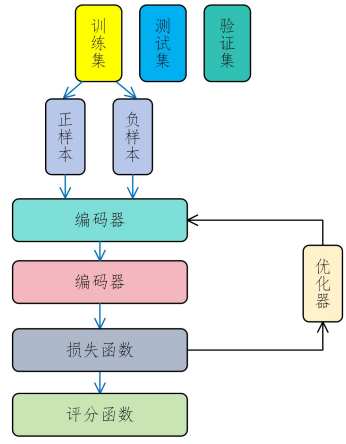

图 2 关系预测任务的流程图

Fig. 2 Flowchart of relationship prediction task

为了进一步提升关系预测质量，深入探讨了解码器 DistMult 的工作机制，并基于 DistMult，创新性地提出了新型解码器 COS-DistMult。

2.1 传统解码器 DistMult

DistMult（张量分解模型）是一种经典的语义匹配模型。它最早由 Yang 等[4]在 2015 年提出，是一种基于神经网络嵌入学习的方法，用于学习知识图谱中实体和关系的表示。其表达式为：

$$g(h,t)=h^{\mathrm{T}} \cdot M_r \cdot t \tag{1}$$

其中，h 和 t 分别代表头实体和尾实体向量；Mr 代表关系矩阵，表示一个图的拓扑结构。DistMult 是基于 RESCAL 扩展而来的，它将双线性模型 RESCAL 的关系矩阵限制为对角矩阵，其满足 $h*Mr*t=t*Mr*h$。这意味着交换头尾实体的顺序，等式仍然成立，也隐含着要求图谱中的每个三元组都是对称关系。换句话说就是，DistMult 丧失了对非对称关系的建模能力。但益处在于它也同时降低了对称关系的复杂性，大幅减少了参数量，进而提高了计算效率，也在一定程度上降低了过拟合的风险。在关系预测任务的实际应用场景中，尤其是在无向图中，DistMult 也被证明是高效的。DistMult 工作机制如图 3 所示。

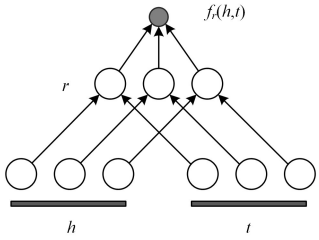

图 3　DistMult 工作机制

Fig. 3　DistMult working mechanism

2.2　新型解码器 COS-DistMult

知识图谱语义模型的核心目的是要衡量向量之间的距离。上文公式中的参数 Mr，对于同一个拓扑图来说是一个常数，而整个公式是将头、尾实体点乘后再乘固定常数。本文认为这样的方式不能有效体现两个实体的相似性或差异性。

受到余弦相似度算法的启发，两个向量的余弦值越大，说明两个向量的夹角越小，向量越相似，即解码器给出的得分值越高。因此，不妨假设直接将解码器给出的两个向量得分值与它们的余弦相似度相乘。所以得分函数公式演变为：

$$g(h,t)=h^{\mathrm{T}} \cdot M_r \cdot t \cdot \cos(h,t) \tag{2}$$

将新方法称为 COS-DistMult，其工作原理如图 4 所示。

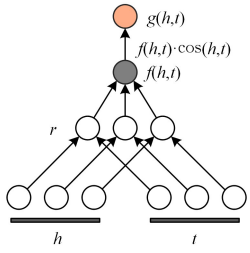

图 4　COS-DistMul 工作机制

Fig. 4　COS-DistMul working mechanism

评价指标 hit@10 和 MRR 的评分机制是衡量三元组语义距离得分的排名，而非得分值本身，因此只要证明样本的 COS-DistMult 得分值的标准差大于原生的 DistMult，即满足不等式：

$$Std(g(x))<Std(g(x)\cdot\cos x) \tag{3}$$

就可以从理论上证明 COS-DistMult 是有效的。但我们无法在数学上直接证明不等式成立，退而求其次，可以通过实验验证预测边的标准差是否变大了，来间接证明在特定数据集上可以满足上述不等式成立。但这里不妨提出一种猜想：只要能扩大关系预测模型输出的预测三元组样本之间的标准差，就可以间接证明算法的有效性。下一章节中通过实验来验证该假设。

3　实验与分析

3.1　实验背景和目的

为了解决电力行业文献知识图谱稀疏性的问题，结合业务需求和项目数据集的情况，设计采用编码器 R-GCN 和解码器 COS-DistMult 的框架完成知识图谱关系预测任务。

具体而言，通常论文的关键词有数量限制，一篇论文的关键词往往在 5 个以内，但在构建知识图谱时，我们希望单篇论文有更加丰富的关键词关联关系，这样在丰富知识表示的同时也解决了知识图谱稀疏性的问题。显然这种任务属于知识图谱领域的关系预测或链接预测工作。对于图表示学习任务，借助知识图谱和图神经网络的表示能力，来挖掘图谱中隐藏的关系，揭示潜在的链接。

3.2　数据集介绍

本实验分别采用自有数据集和公开数据集 WN18 进行对比实验。其中自有数据集采集自各种期刊文献网站上的电力行业相关科技论文数据，共 2 万行左右。每行数据有 7 列，分别为：标题、作者、分类号、机构、摘要、关键词和发表时间。

部分样本情况如图 5 所示。

	标题	作者	机构
0	中国火电行业多模型碳达峰情景预测	张金良	华北电力大学经济与管理学院
1	燃煤电厂水平衡模型与节水分析	刘广建	华北电力大学能源动力与机械工程学院
2	可再生能源配额制与碳排放权交易并行实施	鲍熊剑涛	上海电力大学经济与管理学院
3	多类型电源提升风光电集群高效消纳的源荷	周强	国网甘肃省电力有限公司
4	基于能量函数法的光伏和火电联合外送多机	亢朋朋	国网新疆电力有限公司
5	面向高比例可再生能源电力系统的容量补偿	刘硕	北京电力交易中心有限公司;电力系统及

图 5　数据样例图

Fig. 5　Data sample diagram

通过对数据的分析和建模，论文构建了 3 类实体和 2 类边，并抽象出了知识图谱的基本数据模型，如图 6 所示。其中，实体 1 是机构，其特征为机构名称、所属地区；实体 2 是论文，其特征为标题、分类号、所属关键词、发表时间；实体 3 是关键词。边 1 是实体机构与实体论文之间的链接；边 2 是实体论文与实体关键词之间的链接。

为了丰富知识图谱的内部链接，对论文所涉及的关键词进行了扩充。图 6 中由实体论文指向实体关键词的橙色线条，即关系预测任务要扩充的边。

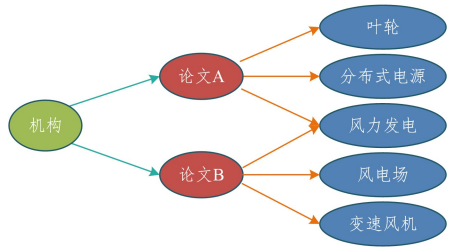

图 6　数据模型

Fig. 6　Data model

3.3　算法有效性实验

为了验证解码器 COS-DistMult 算法的有效性，随机抽取了若干条预测边，描绘了其概率分布，如图 7 所示。经统计，这些预测边的标准差为 0.136。

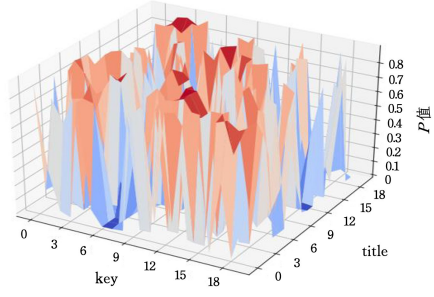

图 7　DistMult 输出的预测边的概率分布

Fig. 7　Probability distribution of predicted edges output by DistMult

图 7 中，key 轴表示实体关键词，title 轴表示实体论文，纵轴 probability 表示它们存在的概率。其概率值越高，颜色越接近暖色；概率值越低，越接近冷色。

同样地，我们也描绘了 COS-DistMult 的概率分布，如图 8 所示，其标准差为 0.178。显然可以得知 COS-DistMult 输出的预测边之间的标准差大于原生 DistMult，即式(3)中描述的不等式在该数据集上成立，说明前者可以在评价指标 hit@10 或 MRR 上得到更高的分值。

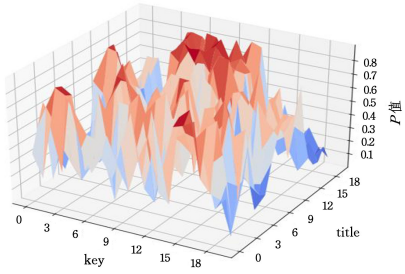

图 8 COS-DistMult 输出的预测边的概率分布

Fig. 8 Probability distribution of predicted edges output by COS-DistMul

分别使用原生 DistMult 和改进版 COS-DistMult 解码器在自有数据集上进行实验。在学习率、丢弃率、迭代次数等其他超参数不变，只改变解码器结构的情况下，运行结果如图 9 所示。

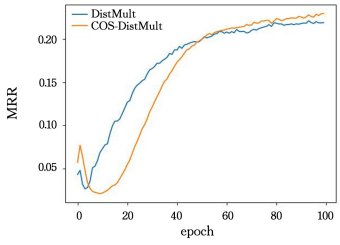

图 9 自有数据集 MRR 指标

Fig. 9 MRR metrics for proprietary datasets

图 9 中，纵轴表示评价指标 MRR 的值，横轴表示迭代次数。蓝色线为原生解码器 DistMult 结果，黄色线条为 COS-DistMult 结果。从图中可以看出，优化后的 COS-DistMult 的 MRR 并没有显著提升。

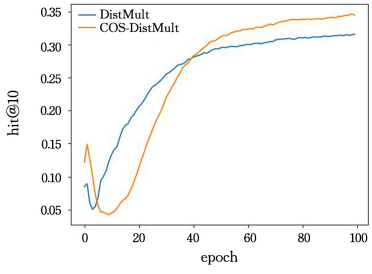

图 10 自有数据集 hits@10 指标

Fig. 10 hits@10 metrics for proprietary datasets

自有数据集评价指标 hit@10 如图 10 所示，横轴表示迭代次数，纵轴表示 hit@10 得分值。蓝色线为原生解码器 DistMult 结果，黄色线条为 COS-DistMult 结果。COS-DistMult 虽然收敛速度变慢了，但 hits@10 值提高了 2%左右。所以总体我们认为在自有数据集上，COS-DistMult 表现更好。

使用 COS-DistMult 模型，输出的知识图谱关系预测效果示例如表 1 所列。

表 1 模型预测效果示例

Table 1 Example of model prediction performance

实体论文	实体关键词
基于实时数据库的火电机组性能诊断与优化管理系统	监控信息系统
推广应用 600 MW 超临界机组的必要性和可行性研究	能损诊断
预测函数控制在火电厂单元机组协调控制系统中的应用	可编程控制器

为了验证模型是否会影响知识图谱的查询速度，收集模型给出的存在概率较高的预测边，并按一定比例将这些预测边三元组关系加入知识图谱中，然后测试查询知识图谱的平均响应速度(秒每次)，其结果如表 2 所列。

表 2 插入预测边后的知识图谱查询平均响应速度

Table 2 Average response speed of knowledge graph query after inserting predicted edges

预测边数量	DistMult	COS-DistMult
50	0.335	0.339
100	0.356	0.352
200	0.572	0.573

从表 2 中可以看出，原生的 DistMult 和改进版的 COS-DistMult 模型，对知识图谱插入等量的预测边后，其响应速度无显著差异。

3.4 公开数据集效果

使用 COS-DistMult 解码器，在公开数据集 WN18 上的 hit@10 值结果如图 11 所示。可以看出，迭代 100 次后 hit@10 值提高了 0.9%。

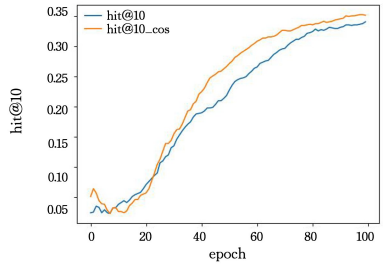

图 11 公开数据集 WN18 评价指标 hit@10

Fig. 11 Evaluation metrics for the publicly available dataset WN18 hits@10

使用 COS-DistMult 解码器，在公开数据集 WN18 上实验，MRR 值结果如图 12 所示。可以看出，迭代 100 次后 MRR 值无显著变化。

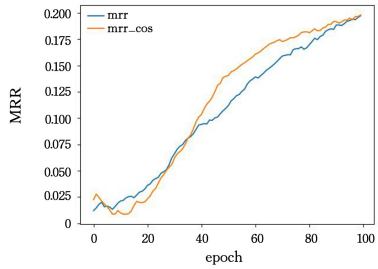

图 12 公开数据集 WN18 评价指标 MRR

Fig. 12 Evaluation metrics for the publicly available dataset WN18 MRR

结束语 本文在知识图谱实体关系预测任务中，对 DistMult 解码器的研究进行了深入的分析，并基于 DistMult 提出了优化策略；创新性开发出了新型解码器 COS-DistMult，且在自有数据集上取得了良好的实验效果，将 hit@10 提高了

2%，在公开数据集 WN18 上将 hit@1 值提升了 0.9%。

综上所述，本文主要创新点如下：

(1)在不含权的无向图的关系预测任务中，改进了解码器 DistMult 的工作机制。

(2)在真实的数据集和公开数据集 WN18 上进行了对比实验，验证了改进版的 COS-DistMult 具有更好的性能。

(3)针对关系预测的算法研究，提出了一种猜想：只要能扩大关系预测模型输出的预测三元组样本之间的标准差，就能从间接上证明算法有效性的一种假设。这在本文实验中取得了成功。

实体关系预测技术在包括知识图谱、社交网络、生物医学在内的诸多领域有广泛的应用和研究价值。只要场景满足以图论为基础的数学模型，实体关系预测技术都有发挥空间。近年来，图神经网络有了重大进展，图卷积优秀的图表示能力，给关系预测技术带来了新的技术支撑和解决方案。所以未来关系预测技术的重点研究方向之一可能是在以图神经网络为编码器的基础上，寻找与它更契合的解码器，让编码器和解码器之间的工作更协调，调试出实验场景下更高效的编码器与解码器组合。

参 考 文 献

[1] BORDES A，USUNIER N，GARCIADURAN A，et al. Translating Embeddings for Modeling Multi-relational Data[C]// Proceedings of the 26th InternationalConferenceon Neural Information Processing Systems. Red Hook，USA：Curran Associates Inc，2013：2787-2795.

[2] LIN Y K，LIU Z Y，SUN M S，et al. Learning Entity and Relation Embeddings for KnowledgeGraphCompletion[C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence. AAAI Press，2015：2181-2187.

[3] NICKEL M，TRESP V，KRIEGEL H P. A three-way model for collective learning onmulti- relational data[C]// Proceedings of the 28th International Conference on International Conference on Machine Learning. New York，USA：ACM，2011：809-816.

[4] YANG B S，YIH W T，HE X D，et al. Embedding Entities and relations for Learning and Inference in Knowledge Bases[C/OL]. https://arxiv.org/pdf/1412.6575.pdf.

[5] NICKEL M，ROSASCO L，POGGIO T. Holographic embeddings of knowledge graphs[C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. AAAI Press，2016：19551961.

[6] KIPF T N，WELLING M. Semi-supervised classification with graph convolutional networks[C]// ICLR. 2017.

[7] SCHLICHTKRULL M，KIPF T N，BLOEM P，et al. Modeling Relational Data with Graph Convolutional Networks[J]. Lecture Notes in Computer Science，2018，10843：593-607.

[8] JIN Z，YANG Z J. Timeknowledge representation learning based on relational time embedding[J]. Journal of Tianjin Urban Construction University，2022(4)：28.

[9] XUE X R，XU D L，LU Y. Research on Power DataAnalysis Algorithm Based on Knowledge Graph and ArtificialIntelligence[J]. Electronic Desig Engineering，2023，31(22)：13.

[10] SHAN X H，ZHAO X，CHEN W Y. Knowledge completion with hierarchical attention based on Bayesian rules[J]. Computer Science，2023，50(11)：234-240.

[11] CHEN H，LI G Y，QI R H. Improved Tucker decomposition knowledge graph completion algorithm[J]. Practice and Understanding of Mathematics，2020，50(16)：13.

[12] WANG K L，ZHOU Z L，CHEN D H. Researchon Link Prediction by Integrating GAT and CapsNet[J]. Communication Technology，2022(2)：55.

[13] PANG J，XU J，QIN H C. Knowledge hypergraph link prediction based on joint attention and convolutional networks[J]. Computer Science and Exploration，2023，17(11)：1.

HAN Yijian，born in 1995，postgraduate. His main research interests include graph neural networks and knowledge graph.

WANG Baohui，born in 1973，senior engineer，master supervisor. His main research interests include software architecture，big data，artificial intelligence，etc.