文章编号:1671-9352(2025)07-0094-10 DOI:10.6040/j.issn.1671-9352.4.2024.534

基于邻域粒度与三支决策的知识表示学习方法

钱文彬,彭嘉豪,蔡星星*
(江西农业大学软件学院, 江西 南昌 330045)

摘要:提出了一种基于邻域粒度与三支决策理论的知识表示学习方法,该方法采用 2 阶段的框架式增强算法,第 1 阶段通过知识表示学习方法拟合知识图谱中的节点与关系,映射其中蕴含的语义信息进入低维向量空间;第 2 阶段,通过划分低维向量表示的邻域粒度,捕捉和利用语义信息中的潜藏相似关系,并辅以三支决策对邻域粒度所挖掘的相似语义信息进行精准的划分,再将所挖掘出的潜藏信息对模型进行重训练,提升知识表示学习方法的准确性与鲁棒性。本文选定 5 种经典的知识表示学习模型,并在 4 个公开的大型知识图谱数据集上进行实验,通过实验结果验证了本方法的有效性。

关键词:知识图谱;知识表示学习;邻域粒度;三支决策
中图分类号:TP391 **文献标志码**:A
引用格式:钱文彬,彭嘉豪,蔡星星,等. 基于邻域粒度与三支决策的知识表示学习方法[J]. 山东大学学报(理学版),2025,60(7):94-103.

Knowledge graph representation learning based on neighborhood granularity and three-way decision

QIAN Wenbin, PENG Jiahao, CAI Xingxing*
(College of Software, Jiangxi Agricultural University, Nanchang 330045, Jiangxi, China)

Abstract: A knowledge representation learning method based on neighborhood granularity and three-way decision theory(NGTwD) is proposed. The method is implemented using a two-stage enhancement algorithm framework. In the first stage, knowledge representation learning is utilized to fit the nodes and relations in the knowledge graph, and map the embedded semantic information into a low-dimensional vector space. To better capture and exploit the latent similarities in the semantic information, the neighborhood granularity of the low-dimensional vector representations is divided in the second stage. This process is further complemented by applying three-way decision theory to precisely segment the similar semantic information. The extracted latent information is then used to retrain the model, thereby improving the accuracy and robustness of the knowledge representation learning method. Five classic knowledge representation learning models are selected, and experiments are conducted on four large publicly available knowledge graph datasets. The effectiveness of the proposed method is validated through the experimental results.
Key words: knowledge graph; knowledge graph representation learning; neighborhood granularity; three-way decision

0 引言

作为一种结构化的知识库,知识图谱[1]核心数据结构——三元组(实体—关系—实体)在描述复杂的人类知识方面提供了一个高度简化和统一的模式。在三元组的基础上,知识图谱构建了一个庞大的语义网络[2],涵盖了大量语义信息,且具有强组织性和可查询性。虽然三元组在一定程度上简化了知识表示的过程,却忽略了实体或关系之间潜藏的内在语义联系,如类似于“同义词”与“近义词”的关系等。

知识图谱中的实体和关系并非孤立存在、毫无关联的,它们之间的相互链接构成了一个复杂的语义网

收稿日期:2024-05-14; **网络出版时间**: 2025-02-28 12:39:53
基金项目:国家重点研发计划资助项目(2022YFD1600202); 国家自然科学基金资助项目(62366019); 江西省自然科学基金资助项目(20224BAB202015)
第一作者:钱文彬(1984—),男,教授,博士,研究方向为多标签学习、粒计算和知识发现. E-mail:qianwenbin1027@126.com
*通信作者**:蔡星星(1987—),女,讲师,硕士,研究方向为机器学习、知识图谱和多源数据融合. E-mail:caixx@jxau.edu.cn

络[3],这个网络中蕴含着丰富的隐含语义信息,通过挖掘和应用这些信息,能够增强知识图谱的应用能力。不同的关系可能也有语义上的重合,比如“创办者”和“创始人”的关系在语义上是可以互换的。传统的知识表示学习方法在这方面的处理简单,不能区分这些细微但重要的差异。

三支决策[4-6]是一种在不确定性环境下进行决策的方法,传统的二元决策模型中,决策者往往只能在接受和拒绝之间做出选择,而三支决策理论的核心在于它引入了第三种选择——推迟决策。这种选择允许决策者在面对不确定性时可以选择等待更多的信息或证据,做出更明智和合理的选择。

本文引入粒计算中的邻域粒度与三支决策作用于知识表示学习中,提出了一种 2 阶段的知识表示学习算法,使知识图谱模型能够充分挖掘关系间的相似语义联系。本文针对粒计算在知识图谱中的应用场景,对知识图谱内邻域粒度的划分进行了改进,通过挖掘关系之间的相似语义及其在知识图谱中的相互作用,基于三支决策理论提出了一种更为精准的正域划分策略,进而为知识表示学习提供了更多有价值的信息。

1　相关工作

知识表示学习是知识图谱的一个重要研究课题,表示学习是探索如何在计算机中高效表达知识的领域,它是构建智能知识图谱的基石,并在众多任务和应用中发挥着关键作用。知识表示学习的核心目标是将知识图谱中的实体和关系映射至一个低维连续的向量空间中[7],通过这种转化,计算机可以有效地执行代数运算揭示实体间复杂的语义联系,并基于这种数值化形式的知识图谱进行机器学习和逻辑推理。

最初的知识表示是基于距离的模型,例如结构嵌入(structured embedding, SE)模型[8]采用线性代数方法编码知识图谱中的结构关系,利用曼哈顿距离或欧氏距离测量实体和关系间的相关性。基于翻译的(translation embedding, TransE)模型[9]通过简单的几何变换模拟关系,该模型假设头实体向量加上关系向量应接近于尾实体向量。但 TransE 模型在处理多对一、一对多、多对多类型的关系时存在局限性。

为克服 TransE 模型在处理复杂关系方面的不足,学者们提出了其他模型。基于超平面的翻译(translating on hyperplanes, TransH)模型[10]为每个关系定义一个超平面,将实体映射到该平面上,同时在超平面上翻译运算,减少嵌入中的歧义。基于关系特定的投影矩阵翻译(translation-based model with relation-specific projection, TransR)模型[11]引入关系特定的映射矩阵,为每个关系构建了独立的特征空间,在该空间进行翻译,体现关系之间的多样性。利用知识表示学习带入复数和四元数空间以提高建模的精准度,旋转嵌入模型(rotational embedding model, RotatE)[12]在复数空间执行旋转操作表示关系,捕捉到关系的对称性、反对称性以及组合性,为知识图谱中关系建模提供了一个全新视角。相位感知旋转嵌入(phase-aware RotatE, pRotatE)[13]模型在 RotatE 模型的基础上引入复数相位的表示方法,捕获知识图谱中关系的数值特征。

知识表示学习的另一个重要分支是基于语义匹配的模型,其中判别性多模态知识图谱嵌入(discriminative multimodal knowledge graph embedding, DistMult)模型[14]和复数嵌入(complex embeddings for knowledge graphs, ComplEx)模型[15]是该方向的典型代表模型,通过相似度评分函数刻画实体和关系之间的匹配程度。这些模型的核心在于探索和揭示头实体与关系之间的深层语义相关性。为了表示语义内容,一些研究尝试将文本特征整合进嵌入过程中,利用实体名称或描述性文本,从而提升知识的多维度表达,增强嵌入表示的信息含量和质量。简易嵌入(simple embedding, SimplE)模型[16]能够全面准确地捕捉知识图谱中的复杂关系,而且模型结构中的参数冗余程度较低。矩阵分解(relational learning approach based on a tensor factorization, RESCAL)模型[17]和张量分解(tensor factorization based on Tucker decomposition, TuckER)模型[18]通过矩阵分解获取低维向量表示,揭示了实体间多维度的相互关系,提升了知识推理与关系预测的精度和效率。

邻域粒度是粒计算的一种重要技术[19-20],细化调整邻域粒度可以更精准地捕捉具有高置信度的相似样本集合。文献[21]将邻域粒度与注意力机制相结合,提出了邻域注意力神经细粒度实体类型,通过自适应聚合实体与关系的邻域粒度,增强知识表示学习的能力;文献[22]根据邻域粒度与生成对应的关系约束,提出了一种带有关系约束的邻域重新排序模型,并取得了良好的效果。

也有许多学者将三支决策理论融入知识表示学习方法,Peng 等[23]通过 K 最近邻算法(K nearest neighbors, KNN)获取关系间的相似度,利用三支决策将三元组进行划分,有效处理长尾关系和未知三元组的

不确定性。Duan 等[24]从三支决策与多粒度决策规则挖掘的角度实现了多粒度条件下的知识图谱概念认知。

2 基于邻域粒度与三支决策的知识表示学习

2.1 基本定义

知识图谱是一种离散化的语义知识库,基本组成单元为三元组结构,若干个三元组交错组成为一个知识图谱, 定义为[1]

$$G=\{(h,r,t)\}\subseteq E\times R\times E, \tag{1}$$

式中,h 为头实体,r 为关系,t 为尾实体,三元组结构为(h,r,t);E 表示实体集合,$h,t\in E$, R 表示关系集合,$r\in R$。头尾实体只有位置的差异,且可以相互对调位置。

知识表示学习旨在为知识图谱中的所有实体与关系寻找一个低维向量表示,令实体向量表示集合为$\boldsymbol{E}=\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n\}$,$\boldsymbol{E}$ 表示由 n 个实体组成的向量集合,每个实体向量具有 d 维特征,即 $\boldsymbol{e}_i=\{e_i^1,e_i^2,\cdots,e_i^d\}$;关系向量表示集合为 $\boldsymbol{R}=\{\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_n\}$,每一个关系向量 $\boldsymbol{r}$ 也是维度为 d 的向量单元。

知识表示学习的目标是寻找一个可以近似满足三元组语义信息的映射函数f',使得$f'(\boldsymbol{h},\boldsymbol{r})\approx\boldsymbol{t}$,其中$\boldsymbol{h}$、$\boldsymbol{t}$ 表示头实体与尾实体的低维向量表示。f'表示实体与关系之间存在的内在联系,$f'(\boldsymbol{h},\boldsymbol{r})$愈接近所映射的目标实体在嵌入空间中的向量 $\boldsymbol{t}$,说明映射关系愈准确,反之则说明映射函数存在语义建模上的缺失。

2.2 基于邻域粒度的相似关系发现

邻域粒度能高效且直观地挖掘知识图谱中的相似关系,通过控制粒度大小,有效地揭示知识图谱中具有潜藏相似语义的关系集合。根据粒计算理论中的邻域粒度概念[19]将知识图谱中的邻域粒度定义为一个给定对象在特定上下文中的直接关联对象集合

$$N(x_i)=\{y\in U\mid f_D(x_i,y)\leqslant D\}, \tag{2}$$

其中 x_i 为有限非空 $U=\{x_1,x_2,\cdots,x_n\}$ 对象全集(论域)中的任意对象,y 是满足某种距离关系 D 的条件下,与 x_i 相关联的对象, f_D 为对象(x,y)在 n 维特征空间 V 下的距离度量函数,定义为

$$f_D(x,y)=\begin{cases}\left(\sum\limits_{k=1}^{n}(\mid V_k(x)-V_k(y)\mid)^P\right)^{\frac{1}{P}}, & P=1,2,\\ \lim\limits_{P\to\infty}\left(\sum\limits_{k=1}^{n}(\mid V_k(x)-V_k(y)\mid)^P\right)^{\frac{1}{P}}, & P=\infty,\end{cases} \tag{3}$$

式中, $f_D(x,y)$称为 Minkowski 距离度量标准,$V_k(x)$表示对象 x 在特征 k 上的特征值。参数 P 表示距离度量函数采用不同的距离度量标准,当 $P=1$ 时,表示距离度量函数采用曼哈顿距离作为距离度量指标;当 $P=2$ 时,距离度量指标采用欧氏距离;当 $P=\infty$ 时,采用切比雪夫距离作为距离度量指标。

对于知识图谱而言,论域 U 相当于知识图谱中的节点或关系的合集,而特征空间 V 则为 U 在低维向量空间上所对应的张量单元的特征值。知识图谱中的邻域系统为

$$N_E=\{E,\boldsymbol{E},\delta\},\quad N_R=\{R,\boldsymbol{R},\delta\}, \tag{4}$$

式中,δ 是邻域阈值,邻域阈值是划分邻域空间范围的关键参数,直接决定了邻域系统在挖掘有效信息时的数量和质量[19]。邻域阈值定义为

$$\delta=\frac{1}{\gamma}\mu\left(\sum_{j=1}^{n}\frac{\sigma(\boldsymbol{v}_j)}{\gamma}\right), \tag{5}$$

式中,$\boldsymbol{v}_j$ 表示第 j 个实体的特征向量,$\sigma(\cdot)$表示 $\boldsymbol{v}_j$ 的标准差, $\mu(\cdot)$是均值函数,γ 是实验中调节邻域大小的超参数。

在预测缺失实体的任务中,邻域粒度求解过程的时间复杂度较高,关系数远少于实体数,为了节省有限的计算资源以获得高效且鲁棒的算法,本文的方法侧重于对相似关系的挖掘。

本文提出了一种新颖的 2 阶段框架式算法,$f_{\text{Embed}}(E,R)\to\boldsymbol{E},\boldsymbol{R}$ 代表通用的知识表示学习方法,此类知识表示学习方法的目标是将实体集合和关系集合中的自然语言信息映射到一个低维向量空间中,得到包含知识图谱中丰富语义的初步嵌入。

对于 $\forall\boldsymbol{r}_i\in\boldsymbol{R}$,采用式(2)计算邻域粒度,欧氏距离作为邻域的度量指标,定义为

$$N(\boldsymbol{r}_i)=\{\boldsymbol{r}_{\text{sim}}\in\boldsymbol{R}\mid f_D(\boldsymbol{r}_i,\boldsymbol{r}_{\text{sim}})\leqslant D\}, \tag{6}$$

其中 $f_D(\boldsymbol{r}_i,\boldsymbol{r}_{\text{sim}})$ 根据式(3)中的距离度量函数进行计算:

$$f_D(\boldsymbol{r}_i,\boldsymbol{r}_{\text{sim}})=\left(\sum_{k=1}^{n}(|V_k(\boldsymbol{r}_i)-V_k(\boldsymbol{r}_{\text{sim}})|)^2\right)^{1/2},\tag{7}$$

式中,$\boldsymbol{r}_{\text{sim}}$ 表示 $\boldsymbol{r}_i$ 在距离关系 D 下的相似关系,即邻域内的相邻相似关系。关系嵌入集合中邻域阈值为

$$\delta_{\boldsymbol{R}}=\frac{1}{\gamma}\mu\left(\sum_{i=1}^{n}\frac{\sigma(\boldsymbol{r}_i)}{\gamma}\right)。\tag{8}$$

根据邻域挖掘得到的相似关系集合,知识图谱嵌入模型可以获取具有相近语义的相似关系集合。对这些相邻的相似关系进行语义相似度计算,并根据语义相似度进行排序,获得在语义层面和特征维度上都较优的相似关系。语义相似度计算方法有:

(1)互信息用于度量两个随机变量之间的相关性,两相似关系间语义相似度的互信息为

$$S_{\text{MI}}(\boldsymbol{r}_i,\boldsymbol{r}_{\text{sim}})=\sum_{u\in\boldsymbol{r}_i}\sum_{v\in\boldsymbol{r}_{\text{sim}}}p(u,v)\log_2\frac{p(u,v)}{p(u)p(v)},\tag{9}$$

式中,u、v 表示关系及相似关系的嵌入单元,p 表示概率分布函数。

(2)余弦相似度度量 2 向量之间相似度的方法,即

$$S_{\text{cosine}}(\boldsymbol{r}_i,\boldsymbol{r}_{\text{sim}})=\frac{\boldsymbol{r}_i\cdot\boldsymbol{r}_{\text{sim}}}{|\boldsymbol{r}_i|\cdot|\boldsymbol{r}_{\text{sim}}|},\tag{10}$$

式中,$\boldsymbol{r}_i\cdot\boldsymbol{r}_{\text{sim}}$ 表示关系嵌入 $\boldsymbol{r}_i\cdot\boldsymbol{r}_{\text{sim}}$ 的点积,$|\boldsymbol{r}_i|$、$|\boldsymbol{r}_{\text{sim}}|$ 分别表示关系嵌入 $\boldsymbol{r}_i$、$\boldsymbol{r}_{\text{sim}}$ 的范数。

(3) 皮尔逊相关系数是衡量 2 个变量之间线性相关程度的统计指标,相似关系的嵌入单元的皮尔逊相关系数为

$$S_{\text{Pearson}}(\boldsymbol{r}_i,\boldsymbol{r}_{\text{sim}})=\frac{\sum_{k=1}^{n}(\boldsymbol{r}_{i_k}-\bar{\boldsymbol{r}}_i)(\boldsymbol{r}_{\text{sim}_k}-\bar{\boldsymbol{r}}_{\text{sim}})}{\sqrt{\sum_{k=1}^{n}(\boldsymbol{r}_{i_k}-\bar{\boldsymbol{r}}_i)^2}\sqrt{\sum_{k=1}^{n}(\boldsymbol{r}_{\text{sim}_k}-\bar{\boldsymbol{r}}_{\text{sim}})^2}}。\tag{11}$$

任意给定的关系 r_i 和相似关系 r_{sim},$S(r_i,r_{\text{sim}})$ 表示 r_i、r_{sim} 之间的相似度。

2.3　基于三支决策理论的知识融合

三元组所对应的关系定义为源区域,即正域(positive region, POS)的范畴;将邻域内挖掘出的邻接关系视为与源关系语义接近的相似关系,归纳为边界域(boundary region, BND),作为潜在隐形语义予以重点关注;对于那些与源关系无直接相关性的非邻接关系,则统一认定为与当前目标无关的关系,归纳为负域(negative region, NEG)。这 3 个区域分别定义为:

$$T_{\text{POS}}=\{(h,r_i,t)\mid(h,r_i,t)\in G\},\tag{12}$$

$$T_{\text{BND}}=\{(h,r_{\text{sim}},t)\mid r_{\text{sim}}\in N(r_i)-r_i,r_{\text{sim}}\in R\},\tag{13}$$

$$T_{\text{NEG}}=\{(h,r_{\text{neg}},t)\mid r_{\text{neg}}\in R-N(r_i)\}。\tag{14}$$

本文提出了一种基于邻域粒度与三支决策理论的知识表示学习方法(a knowledge representation learning method based on neighborhood granularity and three-way decision theory, NGTwD),将传统三支决策中的正域、边界域与负域进行了重构,NGTwD 方法依据三元组关系的观测结果与潜在的邻接实体关系特征,对三元组进行分类。在遵循式(13)的界定后,须设置一个相似度阈值。依据阈值,对比边界域中各个相似关系与源关系在语义相似度上的得分。在这一过程中,NGTwD 方法执行二次迭代,以精细化边界域的划分。若某一相似关系的得分不低于预设阈值,该关系属于新的正域;反之,则被判定为负域。为规避潜在的非确定性多项式时间(nondeterministic polynomial time, NP)难题,本文仅对边界域进行正、负域的划分,不再细分新的边界域。按照这一策略进行划分后,二次迭代的正域与负域的定义为

$$T'_{\text{POS}}=\{(h,r_{\text{sim}},t)\mid(h,r_{\text{sim}},t)\in T_{\text{BND}},\ S(r_i,r_{\text{sim}})\geqslant\alpha\},\tag{15}$$

$$T'_{\text{NEG}}=\{(h,r_{\text{sim}},t)\mid(h,r_{\text{sim}},t)\in T_{\text{BND}},\ S(r_i,r_{\text{sim}})<\alpha\},\tag{16}$$

式中,α 为语义相似度阈值。对关系集合完成两轮划分后,最终确定的正域为 2 次划分所得正域的并集,即

$$T^{\text{U}}_{\text{POS}}=T_{\text{POS}}+T'_{\text{POS}}。\tag{17}$$

2.4　训练过程

本文的训练过程分为 2 个阶段进行,损失函数优化策略均采用梯度下降法。NGTwD 方法使用邻域粒

度与三支决策发掘相似关系的集合后,相似关系集合作为新的监督信息进行重训练。重训练过程中,须要遍历整个相似关系集合,依据损失函数的收敛情况或预设的最大迭代次数确定训练的结束条件。本文的损失函数为

$$L_{\text{Loss}} = L_{\text{First}} + \omega \cdot L_{\text{Re}}, \tag{18}$$

式中,L_{First}表示首次对知识图谱进行嵌入时的损失函数,L_{Re}是对挖掘出的相似关系进行重训练时的损失函数,ω 为重训练的权重参数。本文采用梯度下降的方法对损失函数进行优化。

L_{First}的权重为 1,确保了算法的简化和标准化。尽管本方法主要处理知识图谱中的相似关系,但仍然将知识图谱中现有的、已被验证的三元组视作最高置信度的数据。这些高置信度的三元组为 NGTwD 提供了基准,确保训练过程的稳定性和可靠性。L_{First}定义为

$$L_{\text{First}} = \sum_{(h,r,t)\in G} \sum_{(h',r,t')\notin G} [\gamma_1 + f_{\text{Embed}}(h,r,t) - f_{\text{Embed}}(h',r,t')]_+, \tag{19}$$

式中,$(h',r,t') \notin G$为通过负采样技术获得的负样本三元组,$f_{\text{Embed}}(\cdot)$ 为所选用的知识表示学习模型的评分函数,$[x]_+ = \text{Max}(0,x)$ 代表最大值函数,取$(0,x)$ 中的最大值。γ_1 是训练过程中的超参数,它是损失函数中的边界参数。L_{Re} 为

$$L_{\text{Re}} = \sum_{(h,r,t)\in G} \sum_{r_{\text{sim}}\in R_{\text{POS}}(h',r_{\text{sim}},t')\notin G} [\gamma_2 + \lambda(f_{\text{Embed}}(h,r_{\text{sim}},t) - f_{\text{Embed}}(h',r_{\text{sim}},t'))]_+, \tag{20}$$

式中,参数 λ 是两关系间的语义相似度,γ_2 是重训练过程中的边界超参数。

NGTwD 框架如图 1 所示,第 1 阶段,将知识图谱初始化为随机张量单元之后,将初始化的张量单元导入选定的知识表示学习方法中训练,得到初步嵌入集合。在基于邻域粒度与三支决策理论的知识表示学习步骤中,步骤(3)—(6)是 NGTwD 方法的第 2 阶段,由知识图谱中每一个关系计算邻域粒度,根据邻域获取相似关系集合。计算相似关系集合中的相似关系与源关系间的语义相似度,使用三支决策理论划出置信度最高的正域,即最终相似关系集合。根据步骤(3)中获得的最终相似关系集合,将相似关系与源三元组中的实体相结合,形成新的相似关系三元组,代入 $f_{\text{Embed}}(h,r,t)$。最后根据知识表示学习方法预测得分情况,选取分数最高的预测结果作为最终的结果,并输出其对应的张量单元集合。

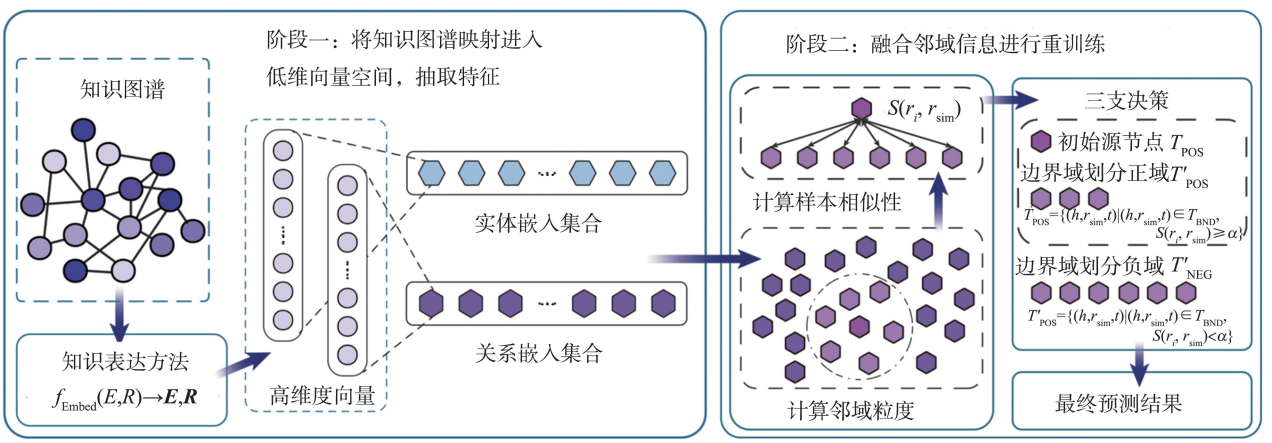

图 1 基于邻域粒度与三支决策理论的知识表示学习方法框架

Fig.1 The framework of knowledge graph representation learning based on neighborhood granularity and three-way decisions

方法 基于邻域粒度与三支决策理论的知识表示学习步骤。

输入 知识图谱 G、邻域阈值 δ、语义相似度阈值 α;

输出 $\boldsymbol{E}$ 和 $\boldsymbol{R}$

(1) 对 G 中所有节点与关系进行随机初始化,转换为 n 维的张量单元;

(2) 将初始化的张量单元导入知识表示学习方法 $f_{\text{Embed}}(h,r,t)$ 中,并获取其初步嵌入单元集合 $\boldsymbol{E}^0$、$\boldsymbol{R}^0$;

(3) 对于知识图谱内任意关系 $\forall r \in R$,结合邻域粒度及三支决策对其重复执行以下步骤:

① 获取 r 的低维向量空间上的张量单元 $\boldsymbol{r} \in \boldsymbol{R}^0$;

② 通过 r 的张量单元 $\boldsymbol{r}$ 与给定 δ 计算其邻域 $N(r)$;

③ 根据 $N(r)$ 划分 r 的 T_{POS1}、T_{BND1}、T_{NEG1};

④ 通过 T_{BND1} 获取 r 的邻接相似关系集合 $R_{\text{sim}} = \{r_{\text{sim}} \mid r_{\text{sim}} \in T_{\text{BND1}}\}$;

⑤ 对于 $\forall r_{\text{sim}} \in R_{\text{sim}}$,计算与 $r \in R$ 之间的语义相似度 $S(r_i, r_{\text{sim}})$;

⑥ 根据 $S(r_i, r_{\text{sim}})$ 与 α 对 T_{BND1} 进行第 1 次划分,分为一个新的正域 T_{POS2} 与新的负域 T_{NEG2};

⑦ 将 2 个正域拼接形成一个新的正域 T_{POS};

(4) 将步骤(3)所有关系的正域 T_{POS} 中相似关系带入 $f_{\text{Embed}}(h,r,t)$ 中重训练,直至遍历完相似关系集合;

(5) 对预测结果进行重排序,选择分数最高的预测结果;

(6) 输出结合了相邻关系内语义相似性的 $\boldsymbol{E}$ 和 $\boldsymbol{R}$。

3　实验设置及结果

为了验证本文方法有效性,采用 4 个公开数据集,分别是 FB15K(Freebase 15K)、FB15K-237(Freebase 15K-237)、WN18(WordNet 18)与 WN18RR(WordNet 18 Reversed Removed)数据集,每个数据集的信息与主要属性如表 1 所示。

表 1　实验数据集信息

Table 1　Information of experimental datasets

数据集	三元组数	关系数	实体数	训练集数	验证集数	测试集数
FB15K	592 213	1 345	14 951	483 142	50 000	59 071
FB15K-237	310 116	237	14 541	271 115	17 535	20 466
WN18	151 442	18	40 943	141 442	5 000	5 000
WN18RR	93 003	11	40 943	86 835	3 034	3 134

在实验设计中,本文选择了不同的对照实验模型,标明“NGTwD-”的模型是采用本文方法的模型,未标明“NGTwD-”的模型是采用基准方法的模型。本文选取了不同的评价指标对模型的整体性能进行评估,分别是平均排名(mean rank,MR)、平均倒数排名(mean reciprocal rank,MRR)以及前 N 个命中率(Hits@N,N=1,3,10)。其中 MR 是测试集中所有正确的三元组在预测结果中排名次序累加之后的平均值,该项评价指标越小,模型的整体性能越优;MRR 是正确三元组的排名次序的倒数累加后的平均值,该项指标越大,模型的整体性能越优;Hits@N 是在预测结果中排名小于 N 的正确三元组的平均占比,该项指标越大,模型的整体性能越优。本文对算法的表现进行了排序,并用粗体表示模型性能最优,下划线表示模型性能次优。

如表 2、3 所示,当嵌入维度为 100 时, NGTwD-pRotatE 模型选用 FB15K 数据集的 MRR 为 57.41%,Hits@10 为 75.96%。NGTwD-pRotatE 模型选用 WN18 数据集的 MRR 为 94.88%,Hits@10 为 95.77%。NGTwD-ComplEx 模型选用 FB15K-237 数据集的 MRR 为 28.09%,Hits@10 为 44.45%。NGTwD-ComplEx 模型选用 WN18RR 数据集的 MRR 为 47.77%,Hits@10 为 56.43%。说明采用本文方法的模型的 MRR 和 Hits@10 取得最优值。

表 2　嵌入维度为 100,选用 FB15K 与 FB15K-237 数据集,基准方法模型和本文方法模型的评价指标

Table 2　Evaluation metrics of models using different methods with dimension 100 on the FB15K and FB15K-237 datasets

模型	FB15K 数据集					FB15K-237 数据集				
	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率
TransE	114.8	38.19%	23.51%	47.58%	63.31%	300.0	24.76%	16.80%	27.30%	40.48%
DistMult	81.61	39.07%	27.16%	44.78%	61.87%	250.9	24.55%	17.08%	26.56%	39.38%
ComplEx	56.59	46.02%	33.41%	53.01%	69.24%	213.4	26.77%	18.74%	29.15%	42.69%
RotatE	89.84	53.87%	41.76%	61.99%	74.28%	400.9	25.74%	17.66%	28.65%	41.65%
pRotatE	69.53	57.19%	47.08%	63.18%	75.68%	264.9	24.57%	16.33%	26.99%	41.03%
NGTwD-TransE	78.21	42.10%	27.08%	52.02%	67.39%	232.2	26.23%	17.49%	29.30%	43.42%
NGTwD-DistMult	65.93	40.82%	28.85%	46.68%	63.58%	209.4	25.50%	17.85%	27.68%	40.62%
NGTwD-ComplEx	**42.17**	48.82%	36.22%	56.02%	71.82%	**163.4**	**28.09%**	**19.80%**	**30.58%**	**44.45%**
NGTwD-RotatE	73.10	54.11%	41.99%	62.20%	74.56%	351.9	26.07%	17.76%	29.15%	42.38%
NGTwD-pRotatE	54.99	**57.41%**	**47.25%**	**63.45%**	**75.96%**	202.6	24.70%	16.37%	27.11%	41.35%

表 3 嵌入维度为 100,选用 WN18 与 WN18RR 数据集,基准方法模型和本文方法模型的评价指标

Table 3 Evaluation metrics of models using different methods with dimension 100 on the WN18&WN18RR datasets

模型	WN18 数据集					WN18RR 数据集				
	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率
TransE	414.0	48.23%	8.29%	88.59%	94.52%	6 498	19.89%	0.989%	36.82%	47.00%
DistMult	399.6	62.78%	48.17%	74.71%	86.22%	5 240	43.10%	39.42%	44.86%	49.94%
ComplEx	323.7	78.30%	70.44%	84.60%	91.01%	4 709	47.00%	42.76%	49.30%	54.83%
RotatE	422.5	94.40%	93.78%	94.81%	95.42%	6 416	44.63%	41.88%	45.69%	50.05%
pRotatE	247.4	94.78%	94.25%	95.10%	95.62%	3 909	46.18%	42.79%	47.29%	52.99%
NGTwD-TransE	320.8	48.59%	8.350%	89.33%	94.67%	5875	20.42%	0.989%	37.94%	47.94%
NGTwD-DistMult	385.8	62.82%	48.17%	74.75%	86.24%	4 861	43.18%	39.45%	45.01%	50.08%
NGTwD-ComplEx	144.2	78.42%	70.56%	84.70%	91.11%	3 111	**47.77%**	**43.04%**	**50.50%**	**56.43%**
NGTwD-RotatE	305.9	94.48%	93.84%	94.91%	95.59%	5 221	44.84%	41.99%	45.87%	50.34%
NGTwD-pRotatE	**125.3**	**94.88%**	**94.35%**	**95.16%**	**95.77%**	**2 010**	46.69%	42.93%	48.04%	54.07%

如表 4、5 所示,当嵌入维度为 500 时,NGTwD-pRotatE 模型选用 FB15K 数据集的 MRR 为 71.62%,Hits@1 为 63.41%,Hits@3 为 77.15%。NGTwD-TransE 模型选用 FB15K-237 数据集的 MR 为 149.2,MRR 为 31.04%,Hits@1 为 21.12%,Hits@3 为 34.96%,Hits@10 为 50.52%。NGTwD-pRotatE 模型选用 WN18 数据集的 MR 为 117.4,MRR 为 94.70%,Hits@1 为 94.04%。NGTwD-RotatE 选用 WN18RR 数据集的 MRR 为 47.75%,Hits@1 为 42.82%,Hits@3 为 49.57%,Hits@10 为 57.58%。说明采用本文方法的模型在 MR、MRR、Hits@1、Hits@3 和 Hits@10 取得最优值。

表 4 嵌入维度为 500,选用 FB15K 与 FB15K-237 数据集,基准方法模型和本文方法模型的评价指标

Table 4 Evaluation metrics of models using different methods with dimension 500 on the FB15K and FB15K-237 datasets

模型	FB15K 数据集					FB15K-237 数据集				
	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率
TransE	34.82	64.15%	51.05%	74.19%	84.55%	188.2	28.89%	19.56%	32.26%	47.32%
DistMult	53.31	45.10%	32.66%	51.88%	68.21%	242.2	25.29%	17.83%	27.31%	40.15%
ComplEx	41.09	50.12%	37.31%	57.54%	73.46%	205.1	27.12%	19.08%	29.51%	43.04%
RotatE	34.85	68.40%	57.07%	77.06%	85.76%	196.6	29.23%	20.15%	32.37%	47.23%
pRotatE	38.76	71.32%	63.11%	76.81%	85.69%	213.5	27.39%	18.75%	30.11%	44.52%
NGTwD-TransE	**25.95**	66.36%	53.53%	76.35%	**85.98%**	**149.2**	**31.04%**	**21.12%**	**34.96%**	**50.52%**
NGTwD-DistMult	45.36	46.56%	34.13%	53.40%	69.58%	206.1	26.21%	18.49%	28.44%	41.50%
NGTwD-ComplEx	31.31	52.69%	39.94%	60.30%	75.62%	154.4	28.45%	20.12%	30.95%	45.10%
NGTwD-RotatE	32.63	68.47%	57.14%	77.11%	85.84%	189.0	29.45%	20.17%	32.83%	47.69%
NGTwD-pRotatE	30.75	**71.62%**	**63.41%**	**77.15%**	85.96%	171.2	27.61%	18.84%	30.42%	44.95%

表 5 嵌入维度为 500,选用 WN18 与 WN18RR 数据集,基准方法模型和本文方法模型的评价指标

Table 5 Evaluation metrics of models using different methods with dimension 100 on the WN18 and N18RR datasets

模型	WN18 数据集					WN18RR 数据集				
	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率
TransE	182.0	60.80%	29.41%	92.62%	95.54%	3 347	20.34%	1.60%	35.42%	49.81%
DistMult	212.0	48.09%	31.73%	57.77%	82.89%	6 002	43.69%	39.90%	45.29%	51.10%
ComplEx	269.4	66.62%	52.07%	78.95%	89.58%	4 183	45.00%	38.93%	48.39%	55.82%
RotatE	178.4	94.70%	93.80%	95.21%	96.15%	3 240	46.99%	42.25%	48.60%	56.33%
pRotatE	162.4	94.62%	93.96%	95.01%	95.74%	2 805	46.23%	41.96%	47.69%	54.58%
NGTwD-TransE	148.7	61.11%	29.59%	93.08%	95.72%	3 277	20.74%	1.66%	36.28%	50.22%
NGTwD-DistMult	204.2	48.12%	31.76%	57.79%	82.94%	5 610	43.86%	40.01%	45.45%	51.44%
NGTwD-ComplEx	151.4	67.09%	52.49%	79.53%	89.85%	2 898	45.51%	39.07%	49.14%	56.88%
NGTwD-RotatE	161.6	94.77%	93.85%	**95.34%**	**96.26%**	2 957	**47.75%**	**42.82%**	**49.57%**	**57.58%**
NGTwD-pRotatE	**117.4**	**94.70%**	**94.04%**	95.06%	95.85%	**1 973**	46.71%	42.10%	48.25%	56.00%

如表 6、7 所示,当嵌入维度为 1 000 时,NGTwD-TransE 模型选用 FB15K 数据集的 Hits@ 10 为 87.35%,比采用基准方法的 TransE 模型的 Hits@ 10 高出 1.21%。NGTwD-pRotatE 模型选用 FB15K-237数据集的 Hits@ 10 为 53.69%,比采用基准方法的 pRotatE 模型的 Hits@ 10 高出 1.57%。NGTwD-RotatE 模型选用 WN18RR 数据集的 Hits@ 10 为 57.99%,比采用基准方法的 RotatE 模型高出 1.56%。

表 6　嵌入维度为 1 000,选用 FB15K 与 FB15K-237 数据集,采用基准方法模型和本文方法模型的评价指标

Table 6　Evaluation metrics of models using different methods with dimension 1 000 on the FB15K and FB15K-237 datasets

模型	FB15K 数据集					FB15K-237 数据集				
	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率
TransE	32.50	66.84%	54.68%	76.05%	86.14%	180.8	29.24%	19.79%	32.69%	48.08%
DistMult	51.53	45.24%	32.76%	51.95%	68.60%	239.9	25.06%	17.48%	27.21%	39.98%
ComplEx	40.68	50.86%	38.02%	58.57%	74.09%	203.9	27.04%	19.00%	29.36%	42.94%
RotatE	32.61	69.48%	58.11%	78.25%	86.82%	187.5	29.48%	20.33%	32.60%	47.75%
pRotatE	35.01	71.39%	62.53%	77.60%	86.39%	174.2	32.42%	22.70%	36.17%	52.12%
NGTwD-TransE	**24.53**	68.94%	57.13%	78.01%	**87.35%**	143.4	31.42%	21.35%	35.50%	51.27%
NGTwD-DistMult	43.83	46.74%	34.27%	53.53%	69.96%	204.9	25.91%	18.12%	28.19%	41.27%
NGTwD-ComplEx	31.41	53.35%	40.60%	61.14%	76.30%	149.6	28.27%	19.91%	30.61%	44.80%
NGTwD-RotatE	30.90	69.58%	58.22%	**78.35%**	86.90%	181.0	29.62%	20.36%	32.83%	48.06%
NGTwD-pRotatE	28.09	**71.94%**	**63.19%**	78.11%	86.70%	**129.6**	**33.49%**	**23.37%**	**37.59%**	**53.69%**

表 7　嵌入维度为 1 000,选用 WN18 与 WN18RR 数据集,基准方法模型和本文方法模型的评价指标

Table 7　Evaluation metrics of models using different methods with dimension 1 000 on the WN18 and WN18RR datasets

模型	WN18 数据集					WN18RR 数据集				
	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率	MR	MRR	前 1 个命中率	前 3 个命中率	前 10 个命中率
TransE	168.6	67.02%	41.38%	93.05%	95.61%	3 035	20.47%	1.99%	35.16%	49.92%
DistMult	237.9	46.75%	30.58%	55.87%	81.56%	3 772	39.23%	31.76%	43.36%	52.87%
ComplEx	224.2	66.13%	51.63%	78.42%	89.10%	4 122	44.91%	38.75%	48.34%	56.05%
RotatE	147.6	94.49%	93.39%	95.22%	96.19%	2 920	46.20%	41.10%	48.15%	56.43%
pRotatE	161.6	94.73%	94.11%	95.04%	95.88%	2 787	45.99%	41.50%	47.37%	55.12%
NGTwD-TransE	137.5	67.39%	41.56%	93.63%	95.78%	2 981	21.07%	2.41%	35.93%	50.48%
NGTwD-DistMult	223.2	46.83%	30.67%	56.00%	81.59%	3 550	39.29%	31.76%	43.51%	52.95%
NGTwD-ComplEx	141.6	66.86%	52.24%	79.49%	89.57%	3 055	45.46%	38.99%	49.15%	57.04%
NGTwD-RotatE	132.8	94.58%	93.45%	**95.34%**	**96.34%**	2 692	**46.89%**	41.31%	**49.17%**	**57.99%**
NGTwD-pRotatE	**96.6**	**94.83%**	**94.19%**	95.17%	96.06%	**1 912**	46.36%	**41.59%**	47.88%	56.33%

本文方法的模型测试随着嵌入维度的增大,模型性能越优,说明在知识表示学习的链接预测任务中,本文设计的增强方法的精准度有明显的提升。将挖掘局部相似关系的语义融入模型后,提高了本文模型的整体链接预测能力,验证了本文方法的有效性。

对 NGTwD 方法第 2 阶段的邻域阈值参数 γ 分析,实验嵌入维度为 1 000,采用 4 个数据集、NGTwD-TransE 模型。实验结果如图 2 所示,不同数据集的 γ 曲线整体趋势保持一致。FB15K、FB15K-237 数据集的关系数分别为 1 345 和 237, $\gamma \in [0.3,0.4]$,曲线转折后趋于稳定。WN18、WN18RR 数据集的关系数分别为 18 和 11, $\gamma \in [0.2,0.3]$,曲线转折后趋于稳定。数据集的关系数越少,γ 曲线趋于稳定的阈值越小。当 γ 较小时,邻域粒度的半径也会减小,导致模型无法获得更多的相似关系,所以模型整体精度会偏小。随着 γ 的增大,整体的模型精度趋于稳定,从而验证本文方法具有一定的鲁棒性。

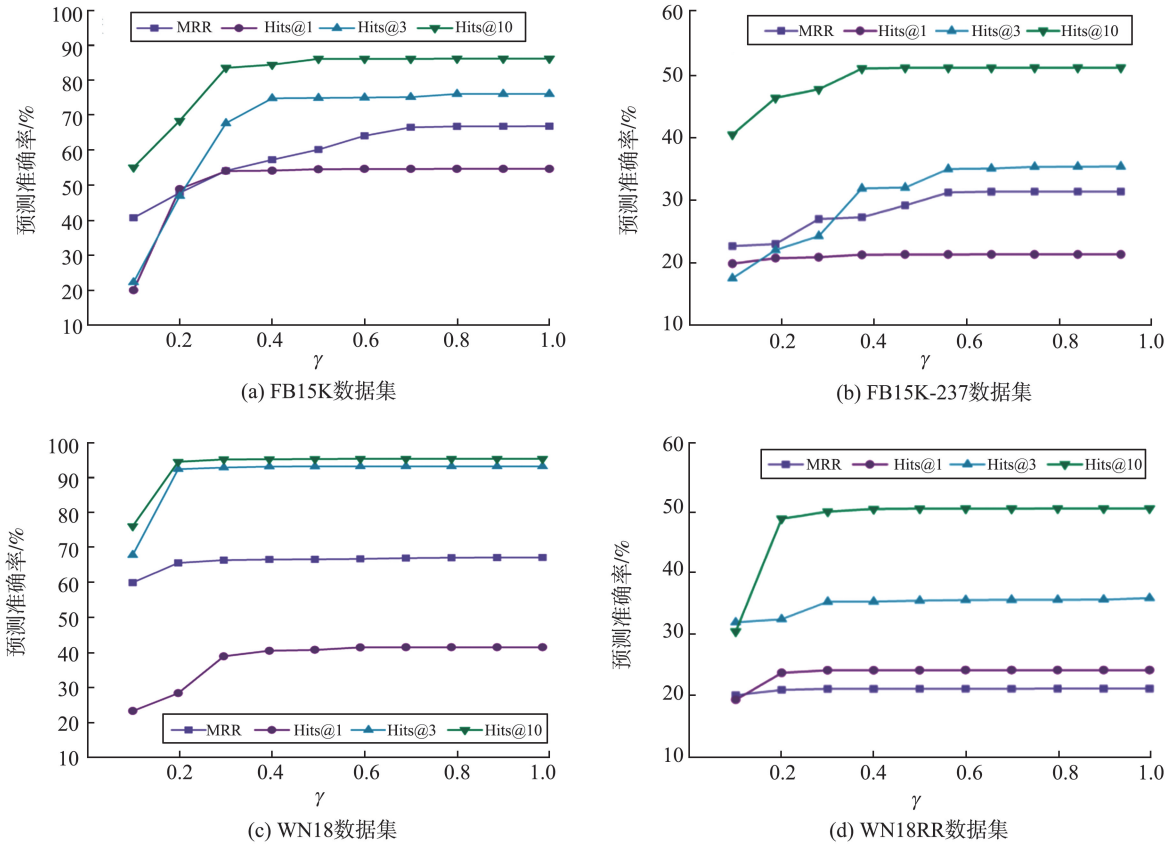

图 2 邻域阈值参数 γ 的分析

Fig.2 Parameter sensitivity analysis of neighborhood threshold parameters γ

4 结论

本文通过将知识表示学习和邻域粒度与三支决策理论相结合,提出了一种新的知识表示学习增强方法,以挖掘和利用知识图谱中的语义信息。通过邻域粒度精细划分相似的近邻关系,利用三支决策理论对邻域粒度挖掘的潜层信息进行再次划分,辅以关系间的语义相似度,将知识图谱中挖掘出的隐含语义关系作为监督信息以重新拟合模型,借助三支决策理论优化了知识的表示和链接预测过程,加强模型对语义相似度的敏感性和建模准确性。实验验证了本文所提方法在知识表示学习中的有效性。本文从粒计算的角度拓展了知识表示学习的理论和方法,取得了较好的效果。

参考文献:

[1] JIShaoxiong, PAN Shirui, CAMBRIA Erik, et al. A survey on knowledge graphs:representation, acquisition, and applications [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(2):494-514.

[2] 张天成,田雪,孙相会,等. 知识图谱嵌入技术研究综述[J]. 软件学报,2023,34(1):277-311.
ZHANGTiancheng, TIAN Xue, SUN Xianghui, et al. Overview on knowledge graph embedding technology research[J]. Journal of Software, 2023, 34(1):277-311.

[3] 王萌,王昊奋,李博涵,等. 新一代知识图谱关键技术综述[J]. 计算机研究与发展,2022,59(9):1947-1965.
WANG Meng, WANG Haofen, LI Bohan, et al. Survey on key technologies of new generation knowledge graph[J]. Journal of Computer Research and Development, 2022, 59(9):1947-1965

[4] DING Juanjuan, ZHANG Chao, LI Deyu, et al. Three-way decisions in generalized intuitionistic fuzzy environments: survey and challenges[J]. Artificial Intelligence Review, 2024, 57(2):38.

[5] YIN Longjun, ZHANG Qinghua, ZHAO Fan, et al. Superiority of three-way decisions from the perspective of probability[J]. Artificial Intelligence Review, 2023, 56(2):1263-1295.

[6] YAO Yiyu. Three-way decision: an interpretation of rules in rough set theory[C]//Rough Sets and Knowledge Technology:

4th International Conference,Gold Coast: Springer, 2009:642-649.
[7] CAO Jiahang, FANG Jinyuan, MENG Zaiqiao, et al. Knowledge graph embedding: a survey from the perspective of representation spaces[J]. ACM Computing Surveys, 2024, 56(6):1-42.
[8] YANG Jing, YANG Laurence T, WANGHao, et al. Representation learning for knowledge fusion and reasoning in cyber-physical-social systems: survey and perspectives[J]. Information Fusion, 2023, 90:59-73.
[9] ANTELMI Alessia, CORDASCO Gennaro, POLATO Mirko, et al. A survey on hypergraph representation learning[J]. ACM Computing Surveys, 2023, 56(1):1-38.
[10] WANG Zhen, ZHANG Jianwen, FENG Jianlin, et al. Knowledge graph embedding by translating on hyperplanes[C]// Proceedings of the AAAI conference on artificial intelligence. Quebec City: AAAI, 2014:1112-1119.
[11] ZHONG Lingfeng, WU Jia, LI Qian, et al. A comprehensive survey on automatic knowledge graph construction[J]. ACM Computing Surveys, 2023, 56(4): 94:1-94.
[12] PENG Ciyuan, XIA Feng, NASERIPARSA Mehdi, et al. Knowledge graphs: opportunities and challenges[J]. Artificial Intelligence Review, 2023, 56(11):13071-13102.
[13] HONG Qinghang, BAI Yushi, TAO Guanyu, et al. Improving knowledge graph embedding with numerical edge features: a pRotatE approach[C]//Proceedings of the 2020 IEEE International Conference on Data Mining, Los Alamitos: IEEE, 2020: 210-219.
[14] AHMED Shams Forruque, ALAM MD Sakib Bin, HASSAN Maruf, et al. Deep learning modelling techniques: current progress, applications, advantages, and challenges[J]. Artificial Intelligence Review, 2023, 56(11):13521-13617.
[15] PHAN Huyen Trang, NGUYEN Ngoc Thanh, HWANG Dosam. Fake news detection: a survey of graph neural network methods[J]. Applied Soft Computing, 2023, 139:110235.
[16] LIANG Ke, LIU Yue, ZHOU Sihang, et al. Knowledge graph contrastive learning based on relation-symmetrical structure [J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(1):226-238.
[17] WANG Xiao, CHENGuangyao, QIAN Guangwu, et al. Large-scale multi-modal pre-trained models: a comprehensive survey [J]. Machine Intelligence Research, 2023, 20(4):447-482.
[18] WANG Jingxiong, ZHANG Qi, SHI Fobo, et al. Knowledge graph embedding model with attention-based high-low level features interaction convolutional network[J]. Information Processing & Management, 2023, 60(4):103350.
[19] HU Qinghua, YU Daren, XIE Zongxia. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2):866-876.
[20] SEWWANDI M A N D, LI Yuefeng, ZHANG Jinglan. A class-specific feature selection and classification approach using neighborhood rough set and *K*-nearest neighbor theories[J]. Applied Soft Computing, 2023, 143:110366.
[21] ZHUO Jianhuan, ZHU Qiannan, YUE Yinliang, et al. A neighborhood-attention fine-grained entity typing for knowledge graph completion[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. New York:Association for Computing Machinery, 2022:1525-1533.
[22] LI Yu, HU Bojie, LIU Jian, et al. A neighborhood re-ranking model with relation constraint for knowledge graph completion [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31:411-425.
[23] PENG Zhihan, YU Hong. Knowledge graph representation learning for link prediction with three-way decisions[C]// International Joint Conference on Rough Sets. Bratislava: Springer, 2021:266-278.
[24] DUAN Jiangli, WANG Guoyin, XIN Hu, et al. Mining multigranularity decision rules of concept cognition for knowledge graphs based on three-way decision[J]. Information Processing & Management, 2023, 60(4):103365.

(编辑:陈丽萍)