智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于相似性负采样的知识图谱嵌入

饶官军， 古天龙， 常亮， 宾辰忠， 秦赛歌， 宣闻

引用本文:
饶官军, 古天龙, 常亮, 等. 基于相似性负采样的知识图谱嵌入[J]. 智能系统学报, 2020, 15(2): 218–226.
RAO Guanjun, GU Tianlong, CHANG Liang, et al. Knowledge graph embedding based on similarity negative sampling[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(2): 218–226.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201811022

您可能感兴趣的其他文章

深度度量学习综述
A brief introduction to deep metric learning
智能系统学报. 2019, 14(6): 1064–1072 https://dx.doi.org/10.11992/tis.201906045

基于Hadoop的大规模网络安全实体识别方法
Large-scale network security entity recognition method based on Hadoop
智能系统学报. 2019, 14(5): 1017–1025 https://dx.doi.org/10.11992/tis.201809024

结合谱聚类的标记分布学习
Label distribution learning based on spectral clustering
智能系统学报. 2019, 14(5): 966–973 https://dx.doi.org/10.11992/tis.201809019

反馈式K近邻语义迁移学习的领域命名实体识别
Domain-named entity recognition based on feedback K-nearest semantic transfer learning
智能系统学报. 2019, 14(4): 820–830 https://dx.doi.org/10.11992/tis.201804013

旅游知识图谱特征学习的景点推荐
Tourism knowledge-graph feature learning for attraction recommendations
智能系统学报. 2019, 14(3): 430–437 https://dx.doi.org/10.11992/tis.201810032

一种结合词向量和图模型的特定领域实体消歧方法
A novel method using word vector and graphical models for entity disambiguation in specific topic domains
智能系统学报. 2016, 11(3): 366–375 https://dx.doi.org/10.11992/tis.201603048

微信公众平台

关注微信公众号，获取更多资讯信息

DOI: 10.11992/tis.201811022
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190520.1347.006.html

基于相似性负采样的知识图谱嵌入

饶官军, 古天龙, 常亮, 宾辰忠, 秦赛歌, 宣闻
(桂林电子科技大学 广西可信软件重点实验室, 广西 桂林 541004)

摘　要: 针对现有知识图谱嵌入模型通过从实体集中随机抽取一个实体来生成负例三元组, 导致负例三元组质量较低, 影响了实体与关系的特征学习能力。研究了影响负例三元组质量的相关因素, 提出了基于实体相似性负采样的方法来生成高质量的负例三元组。在相似性负采样方法中, 首先使用 K-Means 聚类算法将所有实体划分为多个组, 然后从正例三元组中头实体所在的簇中选择一个实体替换头实体, 并以类似的方法替换尾实体。通过将相似性负采样方法与 TransE 相结合得到 TransE-SNS。研究结果表明: TransE-SNS 在链路预测和三元组分类任务上取得了显著的进步。

关键词: 知识图谱; 表示学习; 随机抽样; 相似性负采样; K-Means 聚类; 随机梯度下降; 链接预测; 三元组分类
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2020)02−0218−09

中文引用格式: 饶官军, 古天龙, 常亮, 等. 基于相似性负采样的知识图谱嵌入 [J]. 智能系统学报, 2020, 15(2): 218–226.
英文引用格式: RAO Guanjun, GU Tianlong, CHANG Liang, et al. Knowledge graph embedding based on similarity negative sampling[J]. CAAI transactions on intelligent systems, 2020, 15(2): 218–226.

Knowledge graph embedding based on similarity negative sampling

RAO Guanjun, GU Tianlong, CHANG Liang, BIN Chenzhong, QIN Saige, XUAN Wen
(Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: For the existing knowledge graph embedding model, the random extraction of an entity from the entity set results in the generation of lower-quality negative triples, and this affects the feature learning ability of the entity and the relationship. In this paper, we study the related factors affecting the quality of negative triples, and propose an entity similarity negative sampling method to generate high-quality negative triples. In the similarity negative sampling method, all entities are first divided into a number of groups using the K-means clustering algorithm. Then, corresponding to each positive triple, an entity is selected to replace the head entity from the cluster, whereby the head entity is located in the positive triple, and the tail entity is replaced in a similar approach. TransE-SNS is obtained by combining the similarity negative sampling method with TransE. Experimental results show that TransE-SNS has made significant progress in link prediction and triplet classification tasks.

Keywords: knowledge graph; representation learning; random sampling; similarity sampling; K-means clustering; stochastic gradient descent; link prediction; triple classification

收稿日期: 2018−12−04.　**网络出版日期:** 2019−05−21.
基金项目: 国家自然科学基金资助项目 (U1501252, 61572146); 广西创新驱动重大专项项目 (AA17202024); 广西自然科学基金项目 (2016GXNSFDA380006); 广西高校中青年教师基础能力提升项目 (2018KYD203); 广西研究生教育创新计划项目 (YCSW2018139).
通信作者: 宾辰忠. E-mail: cz_bin@guet.edu.cn.

知识图谱 (knowledge graph) 的概念是谷歌在 2012 年正式提出的, 主要用于提升搜索引擎性能。随着大数据时代的到来, 知识图谱规模得到了快速的增长, 各种大规模知识图谱相继出现 (如 Freebase[1]、WordNet[2]、NULL[3] 等)。当前知识

图谱已在数据挖掘、人工智能等领域具有至关重要的作用, 促进了人工智能及其应用的发展, 如智能问答[4]、个性化旅游推荐等。

虽然现有知识图谱的规模已经相当大, 但其仍是不完整的, 如 Freebase 中 75% 的人不存在国籍信息, 71% 的人没有准确的出身地信息[5], 因此有必要对现有知识图谱进行自动补全。这是当前知识图谱研究中最主要的任务和挑战之一。近年来, 将知识图谱中实体与关系嵌入到向量空间进行知识图谱补全的方法显示出强大的可行性与鲁棒性。但是知识图谱嵌入的研究仍然面临着一个共同的问题, 即在现有知识图谱嵌入模型训练时, 是通过删除正例三元组 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ 中的 $\boldsymbol{h}$(或 $\boldsymbol{t}$), 然后从实体集中随机选择一个实体对删除 $\boldsymbol{h}$(或 $\boldsymbol{t}$) 不完整的三元组进行填充来生成负例三元组, 致使获得的大量负例三元组都是低质量的。低质量的负例三元组将导致知识图谱嵌入模型训练时无法对实体向量与关系向量进行有效的更新, 从而影响知识图谱的有效嵌入。

针对这一不足提出了一种通用的解决方法, 基于实体相似性负采样的负例三元组生成方法来提高知识图谱嵌入的质量。该方法能够在训练中生成一个高质量的负例三元组, 从而实现知识图谱嵌入模型的改进。我们将相似性负采样与 TransE 模型[6] 相结合得到 TransE-SNS 模型, 并且在 4 个通用数据集 (FB15K、FB13、WN11 和 WN18) 上进行了实验, 在链接预测与三元组分类任务中均获得了有效的提升。

1　相关研究

知识图谱嵌入 (knowledge graph embedding) 旨在将知识图谱中的实体与关系嵌入到连续的、稠密的、低维的和实值的向量空间, 将其表示为稠密低维实值向量。然后可以通过向量之间的欧氏距离、曼哈顿距离或马氏距离计算实现对知识图谱中对象间的相似度计算。

在各类知识图谱嵌入模型中, 基于翻译的表示学习[7] 模型实现了先进的性能。其中典型的翻译模型是由 Bordes 等[6] 于 2013 年提出的 TransE 模型。TransE 模型将三元组 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ 中的关系视为向量空间中头实体到尾实体的翻译操作。如果三元组 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ 成立, 则头实体向量 $\boldsymbol{h}$、关系向量 $\boldsymbol{r}$ 与尾实体向量 $\boldsymbol{t}$ 应满足 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$。由于 TransE 模型极为简单, 同时在处理大规模数据方面表现出优异的性能, 从而引起了基于翻译的表示学习研究热潮。在随后几年时间里, 基于翻译的模型衍生出一系列的模型。Wang 等[8] 提出让一个实体在不同关系下拥有不同的表示, 将实体投影到关系所在超平面, 然后在超平面上进行翻译操作。Lin 等[9] 认为实体与关系应当处于不同的语义空间, 提出了 TransR/CtransR 模型。TransR/CtransR 通过投影矩阵将实体从实体空间投影到关系空间, 然后在关系空间中建立翻译操作。Ji 等[10] 认为实体从实体空间投影到关系空间是实体与关系的交互过程, 提出了分别为头、尾实体提供不同的投影矩阵。TranSparse 模型[11] 考虑了实体与关系的不平衡性和异质性, 提出了一种根据关系的复杂程度来自适应的构造稀疏矩阵对实体进行投影, 这样防止了简单关系过拟合、复杂关系欠拟合的发生。Feng 等[12] 认为 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$ 的翻译规则过于严格, 于是建立了更加灵活的翻译规则 $\boldsymbol{h}+\boldsymbol{r}\approx\alpha\boldsymbol{t}$, 提高了模型的表达能力。Chang 等[13] 认为 FT 模型的翻译规则仍过于复杂, 进一步提出了 $(\boldsymbol{h}+\boldsymbol{\alpha}_h)+(\boldsymbol{r}+\boldsymbol{\alpha}_r)\approx(\boldsymbol{t}+\boldsymbol{\alpha}_t)$ 的翻译规则, 实现了翻译模型性能的提升。Tan 等[14] 考虑了不同关系空间中实体的不同状态和特征倾向, 将实体的本征态与拟态进行线性组合作为实体的嵌入特征, 并为每个关系都构造了一个动态关系空间, 提高了关系表示的能力, 减少了来自其他关系的噪声。Wang 等[15] 将生成对抗网络引入表示学习模型中, 利用生成器来获得高质量的负例三元组, 提高了知识表示学习的能力。

2　基于相似性负采样的知识图谱嵌入

2.1　实体的相似性

2.1.1　实体局部结构的相似性

在本节中, 将从两个角度对实体的相似性进行描述: 1) 知识图谱中实体局部结构的相似性; 2) 知识图谱通过 TransE 等翻译模型嵌入到向量空间中实体向量的相似性。

知识图谱中的每个实体间都存在着一定的联系, 包括直接联系与间接联系。直接联系是 2 个实体之间存在直接关系。间接联系是 2 个实体之间存在的关系路径。例如, 给定一个简单的知识图谱, 如图 1 所示。其中, 实体 $\boldsymbol{e}_1$ 与实体 $\boldsymbol{e}_3$ 之间存在着直接关系 (r_8) 和关系路径 (r_6, r_4); 实体 $\boldsymbol{e}_2$ 与实体 $\boldsymbol{e}_3$ 之间存在着直接关系 (r_8) 和关系路径 (r_6, r_4)、(r_6, r_3, r_7) 等。本文将一个实体与其他实体的直接联系形成的结构称为该实体的局部结构, 如图 2 所示。对于任意 2 个实体, 如果它们的局部

结构越相似，那么这 2 个实体也越相似。例如，$\boldsymbol{e}_1$ 与 $\boldsymbol{e}_2$ 的局部结构中分别含有 6 个关系，并且这些关系均相同，其中存在 4 个关系构成了相似的三元组 $(*, r_1, \boldsymbol{e}_4)$、$(*, r_2, \boldsymbol{e}_6)$、$(\boldsymbol{e}_{11}, r_5, *)$ 和 $(*, r_8, \boldsymbol{e}_3)$，其中 $*$ 可以用 $\boldsymbol{e}_1$ 或 $\boldsymbol{e}_2$ 代替，1 个关系连接 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_2$ 的三元组 $(\boldsymbol{e}_2, r_0, \boldsymbol{e}_1)$，1 个关系连接着不同的尾实体构成 2 个三元组 $(\boldsymbol{e}_1, r_6, \boldsymbol{e}_5)$ 和 $(\boldsymbol{e}_2, r_6, \boldsymbol{e}_7)$。通过对比 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_2$ 的局部结构可以判定 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_2$ 的相似性较高。与此相同，通过对比 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_3$ 的局部结构，我们可以判定 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_3$ 的相似性较低。

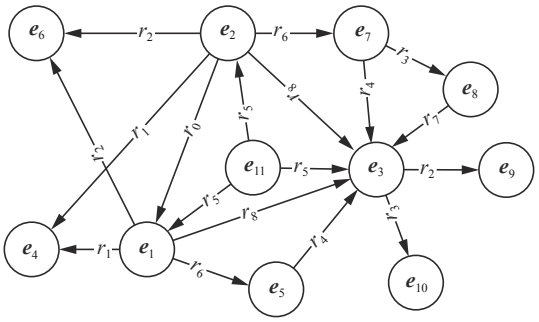
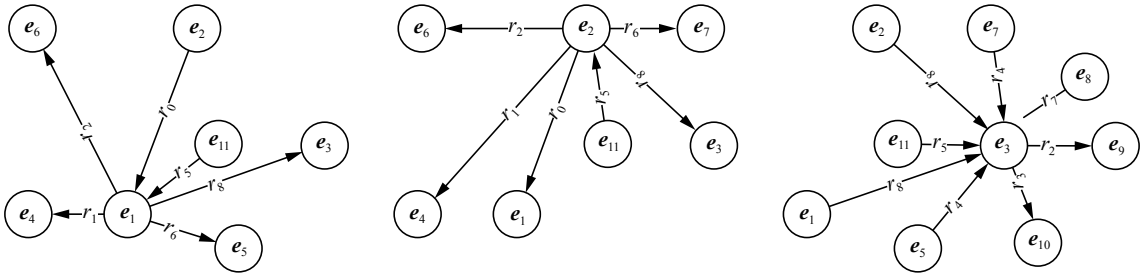


图 1　知识图谱

Fig. 1　Knowledge graph



(a) $\boldsymbol{e}_1$的局部结构　　(b) $\boldsymbol{e}_2$的局部结构　　(c) $\boldsymbol{e}_3$的局部结构

图 2　实体的局部结构

Fig. 2　Local structure of the entity

2.1.2　实体向量的相似性

当利用 TransE 模型将知识图谱嵌入到向量空间中时，对于知识图谱中的每一个三元组 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ 应当满足 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$。对于头实体 $\boldsymbol{h}$ 而言，它在向量空间中的向量 $\boldsymbol{h}\approx\boldsymbol{t}-\boldsymbol{r}$，即 $\boldsymbol{h}$ 可以由 $\boldsymbol{t}-\boldsymbol{r}$ 得到，将 $\boldsymbol{t}-\boldsymbol{r}$ 称之为 $\boldsymbol{h}$ 的空间约束。与此相同，$\boldsymbol{h}+\boldsymbol{r}$ 是 $\boldsymbol{t}$ 的空间约束，$\boldsymbol{t}-\boldsymbol{h}$ 是 $\boldsymbol{r}$ 的空间约束。因此，给定一个知识图谱，通过 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$ 将知识图谱中的实体与关系嵌入到向量空间时，可以将知识图谱中实体的特征与关系的特征在向量空间中的特征表示分别称之为实体向量与关系向量。例如图 2 中的实体 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_2$，当他们嵌入到向量空间中时，需要满足 4 个相同的空间约束，即 $*\approx\boldsymbol{e}_4-\boldsymbol{r}_1$、$*\approx\boldsymbol{e}_6-\boldsymbol{r}_2$、$*\approx\boldsymbol{e}_{11}-\boldsymbol{r}_5$ 和 $*\approx\boldsymbol{e}_3-\boldsymbol{r}_8$，其中 $*$ 可以用 $\boldsymbol{e}_1$ 或 $\boldsymbol{e}_2$ 替换，在这 4 个相同的空间约束下促使 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_2$ 趋近于相等。但与此同时，$\boldsymbol{e}_1$ 还需满足空间约束 $\boldsymbol{e}_1\approx\boldsymbol{e}_2-\boldsymbol{r}_0$ 和 $\boldsymbol{e}_1\approx\boldsymbol{e}_5-\boldsymbol{r}_6$，$\boldsymbol{e}_2$ 还需满足空间约束 $\boldsymbol{e}_2\approx\boldsymbol{e}_1-\boldsymbol{r}_0$ 和 $\boldsymbol{e}_2\approx\boldsymbol{e}_7-\boldsymbol{r}_6$，在这 2 个不同的空间约束下促使 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_2$ 又存在着一定的区别。因此，向量空间中的 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_2$，在相同的空间约束下驱使他们接近于相等，同时又在不同的约束下迫使他们产生一定的区别，这使得 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_2$ 在靠近的同时又存在一些距离。然而，对于图 2 中的实体 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_3$，$\boldsymbol{e}_1$ 在向量空间中受到的所有空间约束均与 $\boldsymbol{e}_3$ 受到空间约束完全不同，这将使得 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_3$ 在向量空间中相距较远。综上所述，在向量空间中，对于任意两个实体，如果他们受到的相同约束越多，那么这两个实体向量之间的距离越小，即实体越相似，反之亦然。

2.2　随机抽样的局限性

现有知识图谱嵌入模型都是采用随机抽样来生成一个负例三元组，即采用相同概率从实体集中抽取一个实体替换正例三元组中的头实体 (或尾实体)。然而，通过该方式生成负例三元组会存在一个问题：可能会在训练中生成大量的低质量负例三元组。导致上述问题的关键在于随机抽样忽略了实体之间的相似性。抽取的替换实体与被替换实体之间相似性可能是很低的，如图 1 中的实体 $\boldsymbol{e}_1$ 与 $\boldsymbol{e}_3$。一个低质量的负例三元组相对于正例三元组来说是极易区分的，这样的负例三元组对于学习知识图谱的有效嵌入是没有作用的。为了深入理解高质量的负例三元组与低质量负例三元组的区别，通过一个具体的例子进行阐述。

假设在知识图谱中有一个正例三元组 (广西，省会，南宁)，根据随机抽样原则，选择替换尾实体南宁来生成负例三元组。首先，通过移除南宁会得到一个不完整的三元组 (广西，省会，？)。然后，以相同的概率从实体集中抽取一个实体进行尾实体填充，假设抽取到一个 Person 类型实体马云，我们就会得到一个奇怪的负例三元组 (广西，

省会, 马云), 这样的负例三元组就是一个低质量的负例三元组。与此相反, 如果抽取到一个City 类型实体桂林, 将会得到一个高质量的负例三元组 (广西, 省会, 桂林)。由于南宁与桂林的相似度远高于南宁与马云的相似度。南宁与桂林拥有许多相同或者相似的属性或者关系类型, 比如, 它们均属于广西, 都拥有城市属性, 地理位置相近, 气候类型相同, 地形地貌相似, 历史文化也相似。与此相反, 南宁与马云, 一个是城市, 一个是人, 他们几乎没有相同的属性或者关系类型, 因此他们的相似度是极低的。

在知识图谱嵌入模型中, 如 TransE 模型使用基于边界的损失函数作为训练目标。训练过程中, 当在使用相似性低的实体进行替换来生成负例三元组时, 得到的是一个低质量的负例三元组, 这将导致生成的负例三元组与正例三元组的得分之差大于边界值, 使得损失值为 0。在损失值为 0 时, 模型将不会对实体向量与关系向量进行更新, 也就是说本次训练对于实体向量与关系向量的学习无益, 无法获得更多的样本特征。为了得到高质量的负例三元组, 促进实体向量与关系向量的有效更新, 实现知识图谱中实体与关系的有效嵌入, 应该使用与被替换实体具有一定程度相似性的替换实体。因此, 针对此问题我们提出了解决方案−相似性负采样。

2.3　相似性负采样

一个高质量的负例三元组可以帮助知识图谱嵌入, 而得到一个高质量的负例三元组的关键是获取一个与被替换实体相似的实体。将知识图谱嵌入到向量空间中时, 实体的局部结构相似性转化为 2 个实体向量的相似性。2 个实体向量之间的距离越小, 它们就越相似, 反之亦然。这促使我们萌发出对实体进行聚类后获得相似实体。于是使用简单有效的 K-Means 算法[16-18] 对实体进行聚类, 然后使得替换实体与被替换实体属于同一个簇。上述实体抽样过程, 称为相似性负采样 (similarity nagetive sampling, SNS)。

对于一个给定的知识图谱 $\boldsymbol{G}=(\boldsymbol{E}, \boldsymbol{R}, \boldsymbol{S})$, 其中 $\boldsymbol{E}=\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_N\}$表示知识图谱中包含 N 个实体的实体集合, $\boldsymbol{R}=\{r_1, r_2, \cdots, r_M\}$表示知识图谱中包含 M 个关系的关系集合, $\boldsymbol{S} \subseteq \boldsymbol{E} \times R \times \boldsymbol{E}$ 表示知识图谱中的三元组集合。本文目标是将 N 个实体划分到 K 个聚类中, 使得每个实体到所属聚类中心最近, 即, 每个实体到所属聚类中心的欧氏距离之和最小, 即满足式 (1):

$$\arg\min \sum_{i=1}^{K} \sum_{e \in C_i} \|\boldsymbol{e}-\boldsymbol{c}_i\|_{L_2} \tag{1}$$

式中: K 表示聚类中心的数量; $\boldsymbol{e}$ 是一个实体向量; $\boldsymbol{c}_i$ 表示第 i 个聚类中心向量; C_i 表示第 i 个聚类中实体 $\boldsymbol{e}$ 的集合; L_2 是第二范数欧氏距离。

知识图谱中实体的相似性负采样详细过程: 首先通过 TransE 模型训练 50 个 epoch 获取实体集 $\boldsymbol{E}$ 与关系集 $\boldsymbol{R}$ 的向量表示, 然后利用 K-Means 聚类将实体集划分为 K 个簇$\{\boldsymbol{E}_1, \boldsymbol{E}_2, \cdots, \boldsymbol{E}_K\}$, 每个簇中的实体之间具有较高的相似性。当给定一个被替换实体 $\boldsymbol{e} \in \boldsymbol{E}_k$ ($k \in \{1, 2, \cdots, K\}$) 时, 从簇 $\boldsymbol{E}_k$ 中选择替换实体 $\boldsymbol{e}'$。那么, 获得的替换实体 $\boldsymbol{e}'$与被替换实体 $\boldsymbol{e}$ 将具有较高的相似性。

2.4　TransE-SNS 模型

在本节中, 详述了将实体相似性负采样与 TransE 模型结合得到 TransE-SNS, 同时给出了 TransE-SNS 模型完整代码的训练过程, 即算法 1。

TransE-SNS 采用 $\boldsymbol{h}+\boldsymbol{r} \approx \boldsymbol{t}$ 的翻译原则将实体和关系嵌入同一个向量空间。因此为 TransE-SNS 定义了得分函数:

$$f_r(\boldsymbol{h}, \boldsymbol{t}) = \|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_{L_2} \tag{2}$$

式中: $\boldsymbol{h}$、$\boldsymbol{r}$、$\boldsymbol{t} \in \mathrm{R}^n$; L_2 是第二范数欧氏距离。

在 TransE-SNS 中, 采用了基于边界的损失函数作为训练目标, 基于边界的损失函数为

$$L = \sum_{(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}) \in S} \sum_{\mathrm{Neg}_{(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})} \in \boldsymbol{S}'} \nabla[f_r(\boldsymbol{h}, \boldsymbol{t}) + \gamma - f_r(\boldsymbol{h}', \boldsymbol{t}')]_+ \tag{3}$$

式中: $\boldsymbol{S}$ 是正例三元组集合; $\boldsymbol{S}'=\{(\boldsymbol{h}', \boldsymbol{r}, \boldsymbol{t}) \mid \boldsymbol{h}' \in \boldsymbol{E}_h; (\boldsymbol{h}', \boldsymbol{r}, \boldsymbol{t}) \notin \boldsymbol{S}\} \cup \{(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}') \mid \boldsymbol{t}' \in \boldsymbol{E}_t; (\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}') \notin \boldsymbol{S}\}$(其中, $\boldsymbol{E}_e$ 表示实体 $\boldsymbol{e}$ 所在的簇) 是负例三元组集合; $\mathrm{Neg}_{(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})}$ 是 $\boldsymbol{S}'$中 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ 对应的一个负例三元组; $[f_r(\boldsymbol{h}, \boldsymbol{t}) + \gamma - f_r(\boldsymbol{h}', \boldsymbol{t}')]_+ = \max(0, f_r(\boldsymbol{h}, \boldsymbol{t}) + \gamma - f_r(\boldsymbol{h}', \boldsymbol{t}'))$; γ 是边界值。本文利用随机梯度下降算法 (SGD)[19] 最小化基于边界的损失函数。

算法 1 给出了 TransE-SNS 模型的完整训练过程。在训练过程中, 前 50 个 epoch 采用的是随机抽样来生成负例三元组进行训练, 然后对实体向量进行第一次聚类。之后每完成训练 50 个 epoch, 就对训练得到的实体向量重新进行一次聚类。

算法 1　Learning TransE-SNS

输入　训练集 $S=\{(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})\}$和负例三元组集 $\boldsymbol{S}'=\{(\boldsymbol{h}', \boldsymbol{r}, \boldsymbol{t}) \mid \boldsymbol{h}' \in \boldsymbol{E}_h, (\boldsymbol{h}', \boldsymbol{r}, \boldsymbol{t}) \notin \boldsymbol{S}\} \cup \{(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}') \mid \boldsymbol{t}' \in \boldsymbol{E}_t, (\boldsymbol{h}', \boldsymbol{r}, \boldsymbol{t}) \notin \boldsymbol{S}\}$, 实体集 $\boldsymbol{E}$, 关系集 $\boldsymbol{R}$, 边界值 γ, 嵌入维度 n, 学习率 α, K-Means 聚类中心数 k, 实体聚类子集 $\boldsymbol{E}_i$ ($i=1, 2, \cdots, k$)。

输出　实体向量与关系向量

1) 初始化：

2) $\boldsymbol{r} \leftarrow$ uniform $(-6/\sqrt{n}, 6/\sqrt{n})$ 对于每一个 $\boldsymbol{r} \in \mathbf{R}$

3) $\boldsymbol{r} \leftarrow \boldsymbol{r}/\|\boldsymbol{r}\|$ 对于每一个 $\boldsymbol{r} \in \mathbf{R}$

4) $\boldsymbol{e} \leftarrow$ uniform $(-6/\sqrt{n}, 6/\sqrt{n})$ 对于每一个 $\boldsymbol{e} \in \mathbf{E}$

4) $\boldsymbol{e} \leftarrow \boldsymbol{e}/\|\boldsymbol{e}\|$ 对于每一个 $\boldsymbol{e} \in \boldsymbol{E}$

5) **loop**

6) $\boldsymbol{S}_{\text{batch}} \leftarrow$ sample $(\boldsymbol{S}, b)$ //从 $\boldsymbol{S}$ 中抽取一个大小为 b 的 mini-batch

7) $\boldsymbol{T}_{\text{batch}} \leftarrow \varnothing$ // 初始化正负例三元组对的集合

8) **for** $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}) \in S_{\text{batch}}$ **do**

9) $\text{Neg}_{(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})} \leftarrow$ sample$(S'_{(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})})$ //抽取一个负例三元组 $(\boldsymbol{h}', \boldsymbol{r}, \boldsymbol{t})$ 或 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}')$

10) $\boldsymbol{T}_{\text{batch}} \leftarrow \boldsymbol{T}_{\text{batch}} \cup \{(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}), \text{Neg}_{(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})}\}$

11) **end for**

12) 更新实体向量与关系向量

$$\sum_{(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}) \in \boldsymbol{S}} \sum_{\text{Neg}_{(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})} \in \boldsymbol{S}'} \nabla[f_r(\boldsymbol{h}, \boldsymbol{t}) + \gamma - f_r(\boldsymbol{h}', \boldsymbol{t}')]_+$$

13) **if** epoch % 50 == 0 **then**

14) 更新 $\boldsymbol{E}_i$// K-Means 聚类

15) **end if**

16) **end loop**

3　实验与分析

为了评估方法的性能，在 4 个公开数据集上进行实验，通过链接预测和三元组分类任务进行评价。

3.1　数据设置

我们使用的数据集是来自于 2 个被广泛使用的知识图谱 WordNet 和 Freebase。WordNet 是一个大型的英语词汇知识图谱。在 WordNet 中，将代表某一基本词汇概念的同义词集合作为实体，并在这些同义词集合之间建立各种语义关系。在本文中，使用 WordNet 中的 2 个子集：WN18[20] 和 WN11[21]。Freebase 是一个大型的人类知识的知识图谱，存储了真实世界中的一般事实。本文也使用了 Freebase 中的 2 个子集：FB15K[20] 和 FB13[21]。在表 1 中给出了这 4 种数据统计数据。

表 1　实验数据集

Table 1　Experimental datasets

Dataset	#Ent	#Rel	#Train	#Valid	#Test
WN11	38 696	11	112 581	2 609	10 544
WN18	40 943	18	141 442	5 000	5 000
FB13	75 043	13	316 232	5 908	23 733
FB15K	14 951	1 345	483 142	50 000	59 071

3.2　链接预测

链接预测旨在预测一个三元组 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ 中缺失的 $\boldsymbol{h}$(或 $\boldsymbol{t}$)。在这项任务中，将测试三元组 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ 缺失的 $\boldsymbol{h}$(或 $\boldsymbol{t}$) 称为正确实体，除正确实体以外的其他实体均被视为候选实体。首先，利用候选实体替换测试三元组 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ 中的 $\boldsymbol{h}$(或 $\boldsymbol{t}$) 以获得候选三元组。然后，计算候选三元组与测试三元组的得分。最后，根据实体对应的三元组得分从低到高对候选实体与正确实体进行升序排列。在 2 个数据集 WN18 和 FB15K 上进行链接预测任务，使用 2 项常用的评价标准作为实验中的评价指标[6]：正确实体排名前 10 的比例 (Hits@10) 和正确实体的平均排名 (Mean Rank)。显然，对于一个好的预测应该有一个高的 Hits@10 和低的 Mean Rank。

值得注意的是，在候选三元组集合中有一部分候选三元组可能存在于训练集、验证集和测试集中。虽然这些候选三元组不是当前测试的正确三元组，但它们应该被认为是正确的，并且它们的得分很可能比当前正确三元组的得分更低，从而影响正确实体的排名。我们将已经在训练集、验证集和测试集中出现过的候选三元组滤除。因此，在测试过程中设置了一个过滤器，并将 2 项评价指标中经过滤器滤除的称之为“Filt”，反之，将其称之为“Raw”。

在这项任务中，为了得到模型的最佳参数设置，尽可能多地尝试了参数的各种设置，参数主要从以下设置中选择：模型训练周期 epoch $\in$ {1 000, 2 000, 3 000}，学习率 $\alpha \in$ {0.01, 0.001, 0.000 1}，边界值 $\gamma \in$ {1, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6}，嵌入维度 $n \in$ {25, 50, 100, 200}，批处理大小 $B \in$ {100, 200, 500, 1 000}，聚类中心数 $K \in$ {16, 32, 64}，聚类迭代次数 $i \in$ {10, 20, 50}，三元组得分与聚类相似度均采用 L_2 第二范数进行计算。在 2 个数据集上，都获得了关于平均排名和排名前 10 的比例的最佳参数设置，如表 2 所示。

表 2　链接预测中的最佳参数设置

Table 2　Optimal parameter setting in link prediction

Dataset	Metric	epoch	α	γ	n	B	K	i	D.S
WN18	Mean Rank	2 000	0.001	5.5	50	100	14	20	L_2
	Hits@10	2 000	0.001	3	50	100	14	20	L_2
FB15K	Mean Rank	2 000	0.001	4	200	200	64	20	L_2
	Hits@10	2 000	0.001	2	200	200	64	20	L_2

在 WN18 和 FB15K 上的链路预测任务实验结果, 如表 3 所示。表中对比模型的实验结果来自于原文献, 加粗的结果为表中最优结果。从表中可以看出, 本文方法在大多数情况下都达到了最先进的效果。在 WN18 中, 本文方法在 Hits@10 (raw, bern) 中性能略低于 TranSparse-DT。在 FB15K 中, 本文方法在 Mean Rank (bern)和 Hits@10 (raw, bern) 未能获得所有模型中的最佳性能。我们认为 TransE-SNS 未能在所有的情况下达到最佳性能有以下 2 个原因: 1)FB15K 数据比较稀疏, 连接的多个相同关系的实体较少, 即每个实体本身对应的相似实体较少, 这导致聚类后每个簇中依旧包含一定数量的相似性较低的实体。2) 聚类中心 K 值选择比较困难, 并且 K 值选择被限制在几个固定值中。因此, K-Means 聚类不能很好地对实体进行聚类。

表 3　链接预测结果

Table 3　Link prediction results

Dataset	WN18				FB15K			
Metric	Mean Rank		Hits@10/%		Mean Rank		Hits@10/%	
	Raw	Filt	Raw	Filt	Raw	Filt	Raw	Filt
SE	1,011	985	68.5	80.5	273	162	28.8	39.8
SME(linear/bilinear)	542/526	533/509	65.1/54.7	74.1/61.3	274/284	154/158	30.7/31.3	40.8/41.3
LFM	469	456	71.4	81.6	283	164	26.0	33.1
TransE	263	251	75.4	89.2	243	125	34.9	47.1
TransH(unif/bern)	318/401	303/388	75.4/73.0	86.7/82.3	211/212	84/87	42.5/45.7	58.5/64.4
TransR(unif/bern)	232/238	219/225	78.3/79.8	91.7/92.0	226/198	78/77	43.8/48.2	65.5/68.7
CTransR(unif/bern)	243/231	230/218	78.9/79.4	92.3/92.3	233/199	82/75	44.0/48.4	66.3/70.2
TransD(unif/bern)	242/224	229/212	79.2/79.6	92.5/92.2	211/194	67/91	49.4/53.4	74.2/77.3
TranSparse(unif/bern)	233/223	221/211	79.6/80.1	93.4/93.2	216/190	66/82	50.3/53.7	78.4/79.9
TranSparse-DT(unif/bern)	248/234	232/221	80.0/**81.4**	93.6/94.3	208/**188**	58/79	**51.2**/**53.9**	78.4/80.2
GTrans-SW(unif/bern)	247/215	234/202	79.1/80.2	92.9/93.5	207/189	66/85	50.6/52.9	75.1/75.3
TransE+GAN-scratch	—	244	—	92.7	—	90	—	73.1
TransE+GAN-pretrain	—	240	—	91.3	—	81	—	74
TransE-SNS(unif/bern)	**220**/**207**	**208**/**195**	**80.2**/80.6	**94.0**/**94.6**	**198**/210	**56**/95	48.9/52.5	**80.1**/**83.0**

图 3 是在数据集 FB15K 中 1 345 个关系, 按照 4 种不同的关系类别分布情况, 其中 1-to-1 的简单关系占比为 24%, 1-to-N、N-to-1 和 N-to-N 的复杂关系分别占比 23%、29% 和 24%。

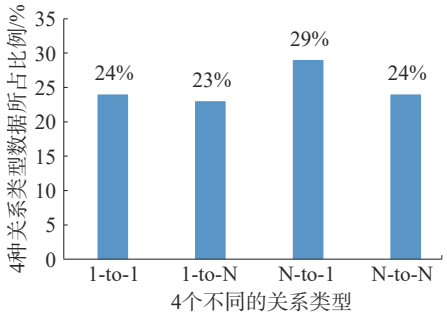


图 3　FB15K 中 1345 个关系的类型分布

Fig. 3　In the FB15K, the category distribution of 1345 relations

表 4 显示了 4 种不同关系类别下 Hits@10 的链接预测结果。值得注意的是, TransE-SNS 模型在大多数情况下都优于其他模型。特别是, 头部和尾部的预测在 N-to-N 关系中实现了最先进的性能。本文方法在 N-to-1 关系中略显不足。总体来说, 本文方法在处理复杂关系方面具有显著的优势。

3.3　三元组分类

三元组分类任务旨在判断一个给定的三元组 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ 是否正确。在本文中, 使用 3 个数据集 (即 WN11、FB13 和 FB15k) 来验证方法在不同数据集上的有效性。Socher 等[20] 提供了 2 个数据集 (即 WN11 和 FB13)。在 WN11 和 FB13 中, 已经包含正例三元组和负例三元组。其中, 每一个负例三元组都是通过破坏正例三元组来获得的。在 FB15K 中只存在正例三元组, 于是使用与 Socher 等[20] 相同的原理构造负例三元组。

表 4　FB15K 按照关系分类的链路预测结果

Table 4　Link prediction results on FB15K by relation category　%

Tasks	Predicting Head (Hits@10)				Predicting Tail (Hits@10)			
Relation Category	1-to-1	1-to-N	N-to-1	N-to-N	1-to-1	1-to-N	N-to-1	N-to-N
SE	35.6	62.6	17.2	37.5	34.9	14.6	68.3	41.3
SME(linear/bilinear)	35.1/30.9	53.7/69.6	19.0/19.9	40.3/38.6	32.7/28.2	14.9/13.1	61.6/76.0	43.3/41.8
TransE	43.7	65.7	18.2	47.2	43.7	19.7	66.7	50.0
TransH(unif/bern)	66.7/66.8	81.7/87.6	30.2/28.7	57.4/64.5	63.7/65.5	30.1/39.8	83.2/83.3	60.8/67.2
TransR(unif/bern)	76.9/78.8	77.9/89.2	38.1/34.1	66.9/69.2	76.2/79.2	38.4/37.4	76.2/90.4	69.1/72.1
CTransR(unif/bern)	78.6/81.5	77.8/89.0	36.4/34.7	68.0/71.2	77.4/80.8	37.8/38.6	78.0/90.1	70.3/73.8
TransD(unif/bern)	80.7/86.1	85.8/95.5	47.1/39.8	75.6/78.5	80.0/85.4	54.5/50.6	80.7/94.4	77.9/81.2
TranSparse(unif/bern)	83.2/87.1	85.2/**95.8**	51.8/44.4	80.3/81.2	82.6/87.5	60.0/57.0	85.5/94.5	82.5/83.7
TranSparse-DT(unif/bern)	83.0/**87.4**	85.7/**95.8**	**51.9**/47.7	80.5/81.6	82.8/86.7	59.9/56.3	85.5/**94.8**	82.9/84.0
GTrans-SW(unif/bern)	80.1/84.9	**93.0**/95.0	48.4/39.9	75.4/75.9	79.4/84.4	51.8/47.7	**91.2**/94.5	77.8/78.8
TransE-SNS(unif/bern)	**83.4**/84.1	88.8/**95.8**	45.6/**48.4**	**83.2**/**85.3**	**87.4**/**88.5**	**60.8**/**60.5**	83.3/94.5	**83.3**/**85.7**

在实验中，为每个关系 $\boldsymbol{r}$ 都设置了一个阈值 δ_r。在验证集上，通过最大化分类准确度来获取每一个关系所对应的 δ_r。对于给定三元组 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$，如果其得分函数的得分低于 δ_r，则将其归类为正例，否则为负例。使用与链接预测相同的方式来获得此任务的最佳参数设置，并得到了 3 个数据集上的最佳参数设置，如表 5 所示。

表 5　三元组分类中的最佳参数设置

Table 5　Optimal parameter setting in triple classification

Dataset	epoch	α	γ	n	B	K	i	D.S
WN18	2 000	0.001	5.5	50	100	16	20	L_2
FB13	2 000	0.001	2	100	200	32	20	L_2
FB15K	2 000	0.001	2.5	200	200	64	20	L_2

表 6 所示是 WN11、FB13 和 FB15K 三元组分类任务的实验结果。从表 6 中可知，TransE-SNS 在所有数据集上的分类性能都优于 TransE 和 TransH。在 FB13 上，TransE-SNS 更是取得了所有模型中的最佳性能。相对于 TranSparse-DT 和 GTans-SW，TransE-SNS 在 WN11 与 FB15K 上的性能略显不足。总体来说，尽管 TransE-SNS 并未在所有数据集上实现最佳性能，但 TransE-SNS 与大多数模型相比，仍具有较大优势。

表 6　三元组分类结果

Table 6　Triple classification results　%

Dataset	WN11	FB13	FB15K
SE	50.3	75.2	—
SME(bilinear)	70.0	63.7	—
SLM	69.9	85.3	—
LFM	73.8	84.3	—
NTN	70.4	87.1	68.2
TransE	75.9	81.5	79.8
TransH	78.8	83.3	79.9
TransR	85.9	82.5	82.1
CTransR	85.7	—	84.3
TranSparse-DT	**86.7**	85.3	88.9
GTrans-SW	86.3	81.7	**91.8**
TransE+GAN-scratch	85.1	83.1	—
TransE+GAN-pretrain	85.4	85.2	—
TransE-SNS	83.2	**87.1**	86.6

4　结束语

本文针对知识图谱嵌入模型中采用随机抽样无法很好地获取高质量的负例三元组，提出了一

种相似性负采样方法用于提高负例三元组的质量。与随机抽样相比, 相似性负采样在很大程度上提高了替换实体与被替换实体间的相似性, 从而提高了负例三元组的质量。在训练时, 相似性负采样生成的高质量负例三元组促进了模型对实体与关系特征的学习。通过将相似性负采样与 TransE 模型结合得到 TransE-SNS 模型。我们的方法能够通过高质量的负例三元组充分获取实体有效特征, 同时忽略了低质量的负例三元组。实验结果表明, TransE-SNS 模型在链路预测与三元组分类任务中均取得了较优的性能。特别是, 相较于基础模型 TransE, 引入相似性负采样后对模型性能具有较大提升。并且, TransE-SNS 模型与 TransE 一样简单且有效, 具有较强的可行性与鲁棒性。但是由于 K-Means 聚类算法本身在 K 值选择以及对数据具有一定要求, 造成相似性负采样对于较为稀疏的大规模知识图谱较难实现相似实体的聚类与采样, 从而影响模型的整体效果。在以后将进一步探索不同聚类算法和知识图嵌入模型的组合, 得到一个更加有效的知识图谱嵌入模型。

参考文献:

[1] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of 2008 ACM SIGMOD International Conference on Management of Data. Vancouver, Canada, 2008: 1247–1250.

[2] MILLER G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39–41.

[3] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning[C]//Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlanta, USA, 2010: 1306–1313.

[4] BORDES A, WESTON J, USUNIER N. Open question answering with weakly supervised embedding models[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Nancy, France, 2014: 165–180.

[5] DONG Xin, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 601–610.

[6] BORDES A, USUNIER N, GARCIA-DURÁN A, et al. Translating embeddings for modeling multi-relational data[C]//International Conference on Neural Information Processing Systems. South Lake Tahoe, USA, 2013: 2787–2795.

[7] 刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展 [J]. 计算机研究与发展, 2016, 53(2): 247–261.
LIU Zhiyuan, SUN Maosong, LIN Yankai, et al. Knowledge representation learning: a review[J]. Journal of computer research and development, 2016, 53(2): 247–261.

[8] WANG Zhen, ZHANG Jianwen, FENG Jianlin, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec City, Canada, 2014: 1112–1119.

[9] LIN Yankai, LIU Zhiyuan, SUN Maosong, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2181–2187.

[10] JI Guoliang, HE Shizhu, XU Liheng, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 687–696.

[11] JI Guoliang, LIU Kang, HE Shizhu, et al. Knowledge graph completion with adaptive sparse transfer matrix[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 985–991.

[12] FENG Jun, HUANG Minlie, WANG Mingdong, et al. Knowledge graph embedding by flexible translation[C]//Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning. Cape Town, South Africa, 2016: 557–560.

[13] CHANG Liang, ZHU Manli, GU Tianlong, et al. Knowledge graph embedding by dynamic translation[J]. IEEE access, 2017, 5: 20898–20907.

[14] TAN Zhen, ZHAO Xiang, FANG Yang, et al. GTrans: generic knowledge graph embedding via multi-state entities and dynamic relation spaces[J]. IEEE access, 2018: 8232–8244.

[15] WANG Peifeng, LI Shuangyin, PAN Rong. Incorporating GAN for negative sampling in knowledge representation learning[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 2005–2012.

[16] HARTIGAN J A, WONG M A. Algorithm AS 136: a K-Means clustering algorithm[J]. Journal of the royal statistical society, 1979, 28(1): 100–108.

[17] HAMERLY G, ELKAN C. Alternatives to the K-Means algorithm that find better clusterings[C]//Proceedings of the 11th International Conference on Information and Knowledge Management. McLean, USA, 2002: 600–607.

[18] CELEBI M E, KINGRAVI H A, VELA P A. A comparative study of efficient initialization methods for the k-means clustering algorithm[J]. Expert systems with applications, 2013, 40(1): 200–210.

[19] DUCHI J, HAZAN E, SINGER Y. Dearly adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of machine learning research, 2011, 12(7): 257–269.

[20] BORDES A, GLOROT X, WESTON J, et al. A semantic matching energy function for learning with multi-relational data: application to word-sense disambiguation[J]. Machine learning, 2014, 94(2): 233–259.

[21] SOCHER R, CHEN Danqi, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 926–934.

作者简介：

饶官军，硕士研究生，主要研究方向为知识图谱、表示学习。

古天龙，教授，博士生导师，主要研究方向为形式化方法、知识工程与符号推理、协议工程与移动计算、可信泛在网络、嵌入式系统。主持国家 863 计划项目、国家自然科学基金项目、国防预研重点项目、国防预研基金项目等 30 余项，出版学 术著作 3 部，发表学术论文 130 余篇。

常亮，教授，博士，中国计算机学会高级会员，主要研究方向为数据与知识工程、形式化方法、智能系统。主持并完成国家自然科学基金项目 1 项、广西省自然科学基金项目 1 项。发表学术论文 70 余篇。