

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 9 月 10 日 (10.09.2020)



(10) 国际公布号
WO 2020/177142 A1

(51) 国际专利分类号:
G06F 16/36 (2019.01)

(21) 国际申请号: PCT/CN2019/077728

(22) 国际申请日: 2019 年 3 月 12 日 (12.03.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910154252.2 2019 年 3 月 1 日 (01.03.2019) CN

(71) 申请人: 东北大学 (NORTHEASTERN UNIVERSITY) [CN/CN]; 中国辽宁省沈阳市和平区文化路三巷 11 号, Liaoning 110819 (CN)。

(72) 发明人: 王大玲 (WANG, Daling); 中国辽宁省沈阳市和平区文化路三巷 11 号, Liaoning 110819 (CN)。 刘泓辰 (LIU, Hongchen); 中国辽宁省沈阳市和平区文化路三巷 11 号, Liaoning 110819 (CN)。 冯时 (FENG, Shi); 中国辽宁省沈阳市和平区文化路三巷 11 号, Liaoning 110819 (CN)。 张

一飞 (ZHANG, Yifei); 中国辽宁省沈阳市和平区文化路三巷 11 号, Liaoning 110819 (CN)。

(74) 代理人: 沈阳东大知识产权代理有限公司 (SHENYANG DONGDA INTELLECTUAL PROPERTY AGENCY CO., LTD); 中国辽宁省沈阳市和平区文化路三巷 11 号, Liaoning 110819 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(54) Title: TRANSFER RELATIONSHIP-BASED LOCAL ADAPTIVE KNOWLEDGE GRAPH OPTIMIZATION METHOD

(54) 发明名称: 一种基于传递关系的本地自适应知识图谱优化方法

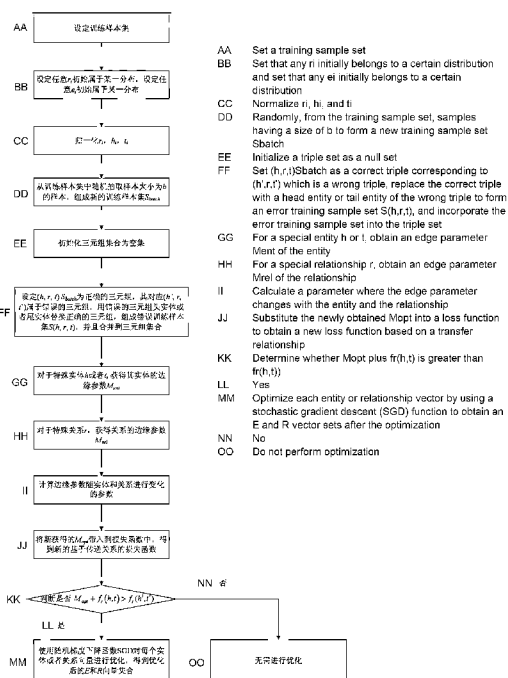


图 1

(57) Abstract: The present invention provides a transfer relationship-based local adaptive knowledge graph optimization method, comprising: setting a training sample set; setting that any r_i and e_i initially belong to a certain distribution; performing normalization; forming a new training sample set; initializing a triple set as a null set; setting a correct triple corresponding to a wrong triple, replacing the correct triple with a head entity or tail entity of the wrong triple to form an error training sample set, and incorporating the error training sample set into the triple set; obtaining an edge parameter of an entity; obtaining an edge parameter of a relationship; calculating a parameter where the edge parameter changes with the entity and the relationship; obtaining a new loss function based on a transfer relationship; and performing determination and optimizing each entity or relationship vector by using a stochastic gradient descent (SGD) function. According to the present invention, the incompleteness of data can be made up, different potential semantics between the relationship and the entity can be better expressed, and the new knowledge map constructed after the optimization has higher accuracy.

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

- 关于申请人有权申请并被授予专利(细则4.17(ii))

本国际公布:

- 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本发明提出一种基于传递关系的本地自适应知识图谱优化方法, 包括: 设定训练样本集; 设定任意 r_i 和 e_i 初始属于某一分布; 归一化; 组成新的训练样本集; 初始化三元组集合为空集; 设定正确的三元组, 其对应错误的三元组, 用错误的三元组头实体或者尾实体替换正确的三元组, 组成错误训练样本集, 并且合并到三元组集合中; 获得其实体的边缘参数; 获得关系的边缘参数; 计算边缘参数随实体和关系进行变化的参数; 得到新的基于传递关系的损失函数; 进行判断并使用随机梯度下降函数SGD对每个实体或者关系向量进行优化; 本发明可以弥补数据的不完全性, 并能够更好地表达出关系与实体间潜在的不同语义, 优化后所构建的新的知识图谱具有更高的准确率。

一种基于传递关系的本地自适应知识图谱优化方法

技术领域

本发明属于知识管理和信息检索领域，具体涉及一种基于传递关系的本地自适应知识图谱优化方法。

背景技术

知识图谱 (Knowledge Graph) 是一种大规模的结构化数据集合，它用于描述现实世界中各种存在的实体与概念，以及实体和实体、实体和概念以及概念和概念的联系，扩充知识结构，进而能够通过计算机来识别并进行分析处理。

知识图谱由三元组(h, r, t)组成 (h 、 r 、 t 分别表示头实体、关系、尾实体)。在目前的基于嵌入翻译的知识图谱中，主要方法是将实体与关系嵌入到同一个低维度空间中，或者是将实体和关系分别嵌入到不同的低维度空间中。这些方法都能极好地构建出知识图谱模型，并能获得良好的效果。其中有很多效果比较突出的方法，如 TransE、TransH、TransR、TransA 等。其中 TransA 模型根据知识图谱的实体和关系能够动态地决定损失函数中的参数，并从理论上证明了损失函数中的参数在知识图谱建模的本地化中的重要作用。但是就知识具有的传递性这一特点而言，现有的这些基于嵌入翻译的知识图谱模型大多没有对其予以考虑，即在构建三元组时，没有将三元组中的关系、实体等具有与其他三元组中的实体、关系的关联进行考虑或者深入分析。

知识是具有传递性的，我们可以举例进行说明。给出一系列三元组：(猫狮金鱼，属于，虎头金鱼)，(虎头金鱼，属于，金鱼)，(金鱼，属于，鱼)，(鱼，属于，动物)。从这四个三元组中，我们可以提取到很多符合常理且符合生物学的新三元组，分别是(猫狮金鱼，属于，金鱼)、(猫狮金鱼，属于，鱼)、(猫狮金鱼，属于，动物)、(虎头金鱼，属于，鱼)、(虎头金鱼，属于，动物)、(金鱼，属于，动物)。这些三元组都是客观事实存在，但是并未显示出来。知识图谱的基本思想是：如果 $h+r$ 与 t 相关，那么 $h+r \approx t$ ，即 $h+r$ 与 t 越相关， $h+r$ 与 t 的距离越近，反之 $h+r$ 与 t 的距离越远。所以在上述例子中“猫狮金鱼”+“属于”不仅仅应该与“虎头金鱼”的位置更相近，同时也应该与“金鱼”、“鱼”、“动物”等实体的距离更近。不能仅仅因为在三元组中没有这些内容，就将这些定义为不正确的三元组，从而将这些从推理方面符合客观事实且科学的内容相互远离。

虽然知识具有传递性，但如果不加限制地传递，却又是片面和不科学的。例如，我们也可以用三元组进行表示(人，像，猿)、(猿，像，猴)、(猴，像，猪)。这里的“像”可以视为知识图谱三元组数据中经常出现的“关系”。在这里，我们就不能也不敢获得类似于第一个例

子的三元组，例如（人，像，猪）这一类型的三元组。从常理上就可以看出，这类三元组就是错误的三元组。但是像（人，像，猴）这一类型的三元组确实可以在一定程度上能够让人接受，但又不是完全接受。这类三元组的定义模糊不清，既可以让人认为是对的，同时也可以让人认为是错的，没有明确的定义。但是，如果继续这样传递下去，一定可以认为是错的。

从上述的两个例子中，我们可以看出，一方面，知识确实具有传递性，因此由实体和关系构成的三元组也应该表达这种传递性；另一方面，三元组在传递的过程中，会发生失真现象或者某些关系不具有知识传递的能力，使得新获得的三元组变得面目全非。所以，在知识进行传递中，需要有一个度量标准，满足该标准才具有传递性。

发明内容

针对以上技术问题，本发明提出一种基于传递关系的本地自适应知识图谱优化方法，即 TransAT，该方法将传递关系引入到前述的 TransA 模型中，形成新的知识图谱模型，即基于传递关系的本地自适应知识图谱模型 TransAT。首先，针对知识图谱中的随机梯度下降算法以及本地自适应知识图谱模型进行扩充，所得到的随机梯度下降算法将替换后的实体根据与原实体的距离不同，采用不同的优化策略，得到不同的优化效果，从而获得新的知识图谱模型。然后，根据实体和关系的位置，结合知识的传递性，产生边缘参数，应用到损失函数中，进行模型优化，获得基于传递关系的本地自适应知识图谱。

一种基于传递关系的本地自适应知识图谱优化方法，具体步骤如下：

步骤 1：设定训练样本集为 $S=\{(h_i, r_i, t_i)\}$ ， h_i 为一个头实体向量， t_i 为一个尾实体向量， r_i 为一个连接头实体和尾实体的关系向量， $i=1, 2, \dots, N$ ， $h_i \in E$ ， $t_i \in E$ ， $r_i \in R$ ，其中， E 为实体集， R 为关系集，设定嵌入空间维度为 n ，知识能够传递的最远距离为 d ，约束因子为 μ ；

步骤 2：设定任意 r_i 初始属于某一分布，并且 $r_i \in R$ ；设定任意 e_i 初始属于某一分布，与 r_i 所属分布为同一分布或者为不同分布，并且 $e \in E$ ， e 为 h 或者 t ；

步骤 3：归一化 r_i ， h_i ， t_i ，归一化公式分别为：

$$r_i = \frac{r_i}{\|r_i\|} \quad (1)$$

$$h_i = \frac{h_i}{\|h_i\|} \quad (2)$$

$$t_i = \frac{t_i}{\|t_i\|} \quad (3)$$

其中， $\|r_i\|$ 、 $\|h_i\|$ 、 $\|t_i\|$ 分别为 r_i ， h_i ， t_i 的一阶范数；

步骤 4：从训练样本集中随机抽取样本大小为 b 的样本，组成新的训练样本集 S_{batch} ；

$S_{batch}=sample(S, b)$; 其中, $sample(S, b)$ 表示在训练样本集 S 中随机抽取 b 个样本;

步骤 5: 初始化三元组集合为空集 $T_{batch}=\Phi$;

步骤 6: 基于归一化后的 r_i, h_i, t_i , 设定 $(h, r, t) \in S_{batch}$ 为正确的三元组, 其对应 (h', r, t') 属于错误的三元组, 用错误的三元组头实体或者尾实体替换正确的三元组, 组成错误训练样本集 $S(h', r, t')$, 并且合并到三元组集合 T_{batch} 中, 得到合并后三元组集合 $T_{batch}=T_{batch} \cup \{(h', r, t'), (h, r, t)\}$;

步骤 7: 对于实体 h 或者 t , 获得其实体的边缘参数 M_{ent} , 公式如下:

$$M_{ent} = \frac{\sum_{r \in R_h} \left(\min_{t, t'} \sigma(\|h - t'\| - \|h - t\|) (1 - \sum_{i=1}^d f(i)) \right)}{n} + \frac{\sum_{r \in R_h} \left(\sum_{i=1}^d f(i) \min_{l, l'} \sigma(\|h - l'\| - \|h - l\|) \right)}{n} \quad (4)$$

$$\text{当 } \sigma(x) = \begin{cases} x & \text{当 } x \geq 0 \\ -x & \text{当 } x < 0 \end{cases}$$

其中, $t \in P_r, t' \in N_r \cup T_r, l \in T_r, l' \in N_r, \sigma(x)$ 为求绝对值的函数, d 表示知识能够传递的最远距离, $f(i)$ 表示传递后每部分实体具有的能量且 $f(i) \geq 0$, 并且整个公式中的传递总能量应该为 1, 所以 $\sum_{i=0}^d f(i) = 0, R_h$ 代表实体 h 在三元组集合中涉及的关系集合, r 表示 h 在三元组中涉及的某一个关系, n 表示 R_h 集合中关系的个数, $\min_{t, t'} \sigma(\|h - t'\| - \|h - t\|)$ 表示获得正确实体和 h 的距离与错误实体和 h 的距离的差值, 而 $\min_{l, l'} \sigma(\|h - l'\| - \|h - l\|)$ 表示获得传递错误实体和 h 的距离与部分错误实体与 h 的距离的差值;

步骤 8: 对于关系 r , 获得关系的边缘参数 M_{rel} , 公式如下:

$$M_{rel} = \min_{r_i \in R_{h,r}} (\|r_i\| - \|r\|) \quad (5)$$

其中, $R_{h,r}$ 是一个关系集合, $R_{h,r} = \{r_1, r_2, \dots, r_n\}$, 表示对于实体 h 除了关系 r 外, 涉及到其他关系的集合, $\|r_i\| - \|r\|$ 表示给定的特定关系到 $R_{h,r}$ 集合中的关系的距离, $\min_{r_i \in R_{h,r}} (\|r_i\| - \|r\|)$ 表示给定的关系和 $R_{h,r}$ 中的关系距离的最小差值, 且当 $\|r_i\| \geq \|r\|$ 情况不存在、或者 $R_{h,r}$ 为空时, $M_{rel} = 0$;

步骤 9: 计算边缘参数随实体和关系进行变化的参数, 公式如下:

$$M_{opt} = \mu M_{ent} + (1 - \mu) M_{rel} \quad (6)$$

其中, M_{ent} 表示实体的边缘参数, M_{rel} 表示关系的边缘参数, 参数 μ 控制 M_{ent} 和 M_{rel} 对 M_{opt} 所占的比例, 且 $0 \leq \mu \leq 1$;

步骤 10: 将新获得的 M_{opt} 带入到损失函数中, 得到新的基于传递关系的损失函数, 即为:

$$L = \sum_{(h, l, t) \in S} \sum_{(h', l, t') \in S'} (M_{opt} + f_r(h, t) - f_r(h', t')) \quad (7)$$

其中, f_r 是基于关系 r 的距离函数, 具体地, $f_r(h, t)$ 为 $h+r$ 与 t 之间的距离函数, $f_r(h', t')$ 为 $h'+r$ 与 t' 之间的距离函数;

步骤 11: 判断若 $M_{opt} + f_r(h, t) > f_r(h', t')$, 输入三元组集合 T_{batch} , 使用随机梯度下降函数 SGD 对每个实体或者关系向量进行优化, 得到优化后的 E 和 R 向量集合, 即完成知识图谱优化, 若判断 $M_{opt} + f_r(h, t) \leq f_r(h', t')$, 无需进行优化;

所述随机梯度下降函数 SGD 进行优化, 公式如下:

$$m' = m \pm (\lambda - t_r(m_0, d)) \nabla f \quad (8)$$

其中, m 表示实体或者关系, m' 表示优化后的实体或者关系, λ 表示随机梯度下降的学习速率, ∇f 表示对三元组 $h+r$ 与 t 的相似度距离求导的结果, m_0 与 m 分别表示知识传递的启发实体和传播收到的实体, $t_r(m_0, d)$ 表示对于关系 r 的基于知识传递的范围函数, 且 $0 < t_r(m_0, d) < 1$, d 表示知识传播的距离, 且 $d > 0$ 。

有益技术效果:

目前, 在基于嵌入翻译的知识图谱中, 主要方法是将实体与关系嵌入到同一个低维度空间中, 或者是将实体和关系分别嵌入到不同的低维度空间中。这些方法都能极好地构建出知识图谱模型, 并能获得良好的效果。但是, 现有的知识图谱模型大多没有考虑到知识的传递性, 即在构建三元组时, 没有将三元组中的关系、实体等与其他三元组中的实体、关系的关联进行考虑或者深入地分析。本发明通过关系传递性的引入以及这种传递性在关系推演过程中的不断衰弱, 可以在很大程度上弥补数据的不完全性, 并能够更好地表达出关系与实体间潜在的不同语义, 因此来得到更好地效果。

采用本发明的基于知识传递关系的本地自适应知识图谱优化方法, 优化后所构建的新的知识图谱, 即优化后的 E 和 R 向量集合链接预测和三元组分类方面具有更高的准确率。

附图说明

图 1 为本发明实施例的种基于传递关系的本地自适应知识图谱优化方法流程图。

具体实施方式

下面结合附图和具体实施实例对发明做进一步说明, 一种基于传递关系的本地自适应知识图谱优化方法, 如图 1 所示, 具体步骤如下:

步骤 1: 设定: 训练样本集为 $S = \{(h_i, r_i, t_i)\}$, h_i 为一个头实体向量, t_i 为一个尾实体向量,

r_i 为一个连接头实体和尾实体的关系向量, $i=1, 2, \dots, N$, $h_i \in E$, $t_i \in E$, $r_i \in R$, 其中, E 为实体集, R 为关系集, 设定嵌入空间维度为 n , 知识能够传递的最远距离为 d , 约束因子为 μ ;

步骤 2: 设定任意 r_i 初始属于某一分布, 并且 $r_i \in R$; 设定任意 e_i 初始属于某一分布, r_i 所属分布为同一分布或者为不同分布, 并且 $e \in E$, e 为 h 或者 t ;

步骤 3: 归一化 r_i , h_i , t_i , 归一化公式分别为:

$$r_i = \frac{r_i}{\|r_i\|} \quad (1)$$

$$h_i = \frac{h_i}{\|h_i\|} \quad (2)$$

$$t_i = \frac{t_i}{\|t_i\|} \quad (3)$$

步骤 4: 从训练样本集中随机抽取样本大小为 b 的样本, 组成新的训练样本集 S_{batch} : $S_{batch} = \text{sample}(S, b)$; 其中, $\text{sample}(S, b)$ 表示在训练样本集 S 中随机抽取 b 个样本;

步骤 5: 初始化三元组集合为空集 $T_{batch} = \Phi$;

步骤 6: 基于归一化后的 r_i , h_i , t_i , 设定 $(h, r, t) \in S_{batch}$ 为正确的三元组, 其对应 (h', r, t') 属于错误的三元组, 用错误的三元组头实体或者尾实体替换正确的三元组, 组成错误训练样本集 $S(h', r, t)$, 并且合并到三元组集合 T_{batch} 中: $T_{batch} = T_{batch} \cup \{(h', r, t'), (h, r, t)\}$;

步骤 7: 对于实体 h 或者 t , 获得其实体的边缘参数 M_{ent} , 公式如下:

$$M_{ent} = \frac{\sum_{r \in R_h} \left(\min_{t, t'} \sigma(\|h - t'\| - \|h - t\|) (1 - \sum_{i=1}^d f(i)) \right)}{n} + \frac{\sum_{r \in R_h} \left(\sum_{i=1}^d f(i) \min_{t, t'} \sigma(\|h - t'\| - \|h - t\|) \right)}{n} \quad (4)$$

$$\text{当 } \sigma(x) = \begin{cases} x & \text{当 } x \geq 0 \\ -x & \text{当 } x < 0 \end{cases}$$

其中, $t \in P_r$, $t' \in N_r \cup T_r$, $l \in T_r$, $t' \in N_r$, $\sigma(x)$ 为求绝对值的函数, d 表示知识能够传递的最远距离, $f(i)$ 表示传递后每部分实体具有的能量且 $f(i) \geq 0$, 并且整个公式中的传递总能量应该为 1, 所以 $\sum_{i=0}^d f(i) = 0$, R_h 代表实体 h 在三元组集合中涉及的关系集合, r 表示 h 在三元组中涉及的某一个关系, n 表示 R_h 集合中关系的个数, $\min_{t, t'} \sigma(\|h - t'\| - \|h - t\|)$ 表示获得正确实体和 h 的距离与错误实体和 h 的距离的差值, 而 $\min_{t, t'} \sigma(\|h - t'\| - \|h - t\|)$ 表示获得传递错误实体和 h 的距离与部分错误实体与 h 的距离的差值;

步骤 8: 对于关系 r , 获得关系的边缘参数 M_{rel} , 公式如下:

$$M_{rel} = \min_{r_i \in R_{h,r}} (\|r_i\| - \|r\|) \quad (5)$$

其中, $R_{h,r}$ 是一个关系集合, $R_{h,r} = \{r_1, r_2, \dots, r_n\}$, 表示对于实体 h 除了关系 r 外, 涉及到其他关系的集合, $\|r_i\| - \|r\|$ 表示给定的特定关系到 $R_{h,r}$ 集合中的关系的距离, $\min_{r_i \in R_{h,r}} (\|r_i\| - \|r\|)$ 表示给定的关系和 $R_{h,r}$ 中的关系距离的最小差值, 且当 $\|r_i\| \geq \|r\|$ 情况不存在、或者 $R_{h,r}$ 为空时, $M_{rel} = 0$;

步骤 9: 计算边缘参数随实体和关系进行变化的参数, 公式如下:

$$M_{opt} = \mu M_{ent} + (1 - \mu) M_{rel} \quad (6)$$

其中, M_{ent} 表示实体的边缘参数, M_{rel} 表示关系的边缘参数, 参数 μ 控制 M_{ent} 和 M_{rel} 对 M_{opt} 所占的比例, 且 $0 \leq \mu \leq 1$;

步骤 10: 将新获得的 M_{opt} 带入到损失函数中, 得到新的基于传递关系的损失函数, 即为:

$$L = \sum_{(h,l,t) \in S} \sum_{(h',l',t') \in S'} (M_{opt} + f_r(h,t) - f_r(h',t')) \quad (7)$$

其中, f_r 是基于关系 r 的距离函数, 具体地, $f_r(h, t)$ 为 $h+r$ 与 t 之间的距离函数, $f_r(h', t')$ 为 $h'+r$ 与 t' 之间的距离函数;

步骤 11: 判断若 $M_{opt} + f_r(h, t) > f_r(h', t')$, 输入三元组集合 T_{batch} , 使用随机梯度下降函数 SGD 对每个实体或者关系向量进行优化, 得到优化后的 E 和 R 向量集合, 即完成知识图谱优化, 若判断 $M_{opt} + f_r(h, t) \leq f_r(h', t')$, 无需进行优化;

所述随机梯度下降函数 SGD 进行优化, 公式如下:

$$m' = m \pm (\lambda - t_r(m_0, d)) \nabla f \quad (8)$$

其中, m 表示实体或者关系, m' 表示优化后的实体或者关系, λ 表示随机梯度下降的学习速率, ∇f 表示对三元组 $h+r$ 与 t 的相似度距离求导的结果, m_0 与 m 分别表示知识传递的启发实体和传播收到的实体, $t_r(m_0, d)$ 表示对于关系 r 的基于知识传递的范围函数, 且 $0 < t_r(m_0, d) < 1$, d 表示知识传播的距离, 且 $d > 0$ 。

下面给出 TransAT 算法的实施例, 针对该算法编程即可实现基于知识传递关系的本地自适应知识图谱的优化, 构建新的知识图谱, 如表 1 所示, 用伪代码实现:

表 1: 基于知识传递的本地自适应知识图谱模型

算法 TransAT: 基于知识传递的本地自适应知识图谱模型

输入: $S=\{(h, r, t)\}$, 实体集 E 和关系集 R , 嵌入空间维度 n 传播距离 d , 约束因子 μ

输出: 优化后的实体和关系向量集合

算法描述:

1. 对于任意 r 初始分布于 $\left(-\frac{6}{\sqrt{k'}}, \frac{6}{\sqrt{k}}\right)$, 并且 $r \in R$;
2. 归一化 r , $r = \frac{r}{\|r\|}$;
3. 对于任意 e 初始分布于 $\left(-\frac{6}{\sqrt{k'}}, \frac{6}{\sqrt{k}}\right)$, 并且 $e \in E$, e 是 h 或者 t ;
4. loop
5. 归一化 h 和 t , $h = \frac{h}{\|h\|}$, $t = \frac{t}{\|t\|}$;
6. 样本大小 b , $S_{batch} = \text{sample}(S, b)$;
7. 初始化三元组集合 $T_{batch} = \Phi$;
8. for $(h, r, t) \in S_{batch}$ do
9. 从 (h', r, t') 中提取样本 $(S(h', r, t'))$;
10. $T_{batch} = T_{batch} \cup \{(h', r, t'), (h, r, t)\}$;
11. end for
12. 对于实体 h 或者 t , 获得 M_{ent} ;
13. 对于关系 r , 获得 M_{rel} ;
14. $M_{opt} = \mu M_{ent} + (1 - \mu) M_{rel}$;
15. 更新损失函数 $\sum_{(h, t, r) \in S} \sum_{(h', t', r') \in S'} (M_{opt} + f_r(h, t) - f_r(h', t'))$
If $(M_{opt} + f_r(h, t) > f_r(h', t'))$ do
16. SGD(h, t, r, h', t');
17. endif
18. end loop
19. return 优化后 E 和 R 向量集合;
20. finish

该算法中输入是三元组集合、实体集合、关系集合、实体和关系需要嵌入的低维度空间的维度、对于实体知识传播的距离以及边缘参数中对实体部分边缘参数和关系部分边缘参数的约束参数。这里的“三元组集合”可以从 WordNet、Freebase 等现有资源中获得。算法输出的

是基于传递关系的本地自适应知识图谱模型。第 1、2 行表示对关系的初始操作，对每个关系在低维度空间中随机赋予约束内一个位置，并且能够更好地处理，将关系进行归一化。第 3 行表示对所有实体嵌入到低维度空间向量中，并将位置随机分配在一个区域内。第 4-18 行表示对基于传递关系的自适应知识图谱模型进行循环优化，其中第 5 行表示对涉及到的实体，分别是头实体和尾实体进行归一化，第 6 行表示从三元组集合中选取一部分作为训练集。第 7-11 行表明将正确的三元组和错误的三元组进行集合，其中第 9、10 行表明替换正确三元组中的头实体或者尾实体，形成错误的三元组，第 12-14 行是通过公式(4)、公式(5)和公式(6)计算 M_{ent} 、 M_{rel} 、 M_{opt} 。第 15-17 行是对正确三元组和错误三元组涉及的参数进行优化，第 15 行是基于传递关系的本地自适应损失函数，判断是否对涉及的参数进行优化操作。第 16 行是对正确三元组中的 h 、 t 、 r 和错误三元组中的 h' 、 t' 、 r 进行随机梯度下降操作，所调用的随机梯度下降函数 SGD 描述如表 2，用伪代码实现：

表 2：基于传递关系的随机梯度下降

函数 SGD: 基于传递关系的随机梯度下降

输入: 训练三元组集合 (h, r, t) , 错误三元组集合 (h, r, t') , 嵌入维度 n , 学习速率 λ , 传输距离 d ;

输出: 优化后的正确三元组集合 (h, r, t) 和错误三元组集合 (h, r, t') 中实体和关系向量;

算法描述:

1. $d' = t_r(t, d)$;
 2. for $i=1$ to n
 3. $x = \nabla f(t[i] - h[i] - r[i])$;
 4. $h[i] = h[i] + \lambda x$;
 5. $r[i] = r[i] + \lambda x$;
 6. $t[i] = t[i] - \lambda x$;
 7. $x = \nabla f(t'[i] - h[i] - r[i])$;
 8. $h[i] = h[i] - (\lambda - d')x$;
 9. $r[i] = r[i] - (\lambda - d')x$;
 10. $t'[i] = t'[i] + (\lambda - d')x$;
 11. endfor
 12. return 优化后的 (h, r, t) 和 (h, r, t')
 13. finish
-

函数 SGD 的输入是：知识图谱中正确的三元组、错误的三元组、嵌入的低维度空间的维度、随机梯度下降算法的学习速率和知识的传播距离，然后获得对实体和关系处理后、在低维度嵌入空间中的位置。第 1 行显示错误三元组中换掉的实体与正确三元组中的原实体传播后的能量值。第 2-11 行显示对知识图谱模型中的实体与关系的循环优化，其中第 3-6 行显示正确三元组优化的结果，使用公式(8)，第 7-10 行显示错误三元组在基于支持传播后的优化结果，使用公式(8)。其中，第 3 行和第 7 行分别表示对正确三元组中的差值和错误三元组中的差值进行求导。

以下是本发明算法的原理详细描述：

梯度下降算法是机器学习中作为模型优化的基本方法之一，该算法通过训练集中随机选取某个训练样本，在模型进行迭代求解过程中去计算梯度，并通过计算后的梯度对模型中的参数进行更新。随机梯度下降算法的主要思想是：在给定的训练样本集 M 中，随机选取其中 N 个样本作为新的训练样本集，去替代原训练样本集 M ，完成模型进行训练。在知识图谱中，随机梯度下降算法是对损失函数 f 进行参数优化。它是按照梯度反方向更新参数 w 直到函数收敛。算法为：

$$w' = w - \lambda \nabla f \quad (9)$$

其中， w 表示需要优化的参数，而 w' 表示优化后得到新的参数的值， λ 表示学习速率， ∇f 表示对损失函数求导的结果。

优化的随机梯度下降算法实现。这里的优化随机梯度下降算法实际上是将传递关系引入进来。具体地，设 E 是实体集合， R 是关系集合，把任意一个三元组定义为 (h_i, r_i, t_i) ($i=1, 2, \dots, n$)。其中， h_i 是一个头实体向量， t_i 是一个尾实体向量， r_i 是一个连接头实体和尾实体的关系向量。且 $h_i \in \mathbf{R}^k$ 和 $t_i \in \mathbf{R}^k$ 且 $r_i \in \mathbf{R}^d$ ，其中 k 可以等于 d ，也可以不等于 d 。 $\mathbf{R}^k$ 与 $\mathbf{R}^d$ 可以相同也可以不相同。 h_i 、 t_i 和 r_i 可以处于同一个嵌入空间 $\mathbf{R}^k$ 中，同样 h_i 、 t_i 可以处于 $\mathbf{R}^k$ 空间中，而 r_i 处于 $\mathbf{R}^d$ 空间中。 Δ 表示正确的三元组集合，即这个三元组在训练集中出现， Δ' 表示错误的三元组。

有多个三元组，分别是 (h_1, r_1, t_1) 、 (t_1, r_1, h_2) 、 (h_2, r_1, t_2) 。可以看出， h_1 与 h_2 可能具有关系的关联性，而 t_1 与 t_2 同样可能具有关系的关联性。根据闭包原理中关系传递性原理， h_1 与 t_2 同样可能具有关联性。由于关系本身含义的原因，知识的传递性可能会在传递时失真，无法确保其有效性。所以当传递距离越远时，越是应该减少其关联性，直到为 0。因此，基于传递关系的随机梯度下降算法的主要思想是：由于不能完全确定传递的关系是否在传递时失

真，所以在进行对知识图谱模型进行优化时，当传递具有的能量越剧烈，即关系传播的距离越近，错误三元中相应变换的位置远离的梯度就越低；当关系传播的距离越远，传递的能量越低，错误三元中相应变换的位置远离的梯度就越高，直到变回到正常的梯度。

当出现具有传递关系的三元组，进行模型优化时应将错误的三元组梯度对应的梯度相应减少，传递关系越强烈，优化的梯度就应该越低。由此，基于传递关系的随机梯度下降算法为：

$$m' = m \pm (\lambda - t_r(m_0, d)) \nabla f \quad (8)$$

其中， m 表示实体或者关系， m' 表示优化后的实体或者关系， λ 表示随机梯度下降的学习速率， ∇f 表示三元组 $h+r$ 与 t 的相似度距离的导数， m_0 与 m 分别表示知识传递的启发实体和传播收到的实体， $t_r(m_0, d)$ 表示对于关系 r 的基于知识传递的范围函数，且 $0 < t_r(m_0, d) < 1$ 。 d 表示知识传播的距离，为了将该模型与随机梯度下降算法作为区别，设置 $d > 0$ 。并且 $t_r(m_0, d)$ 可以为线性函数、指数型函数或者对数型函数，根据具体情况做具体的选择。从公式(2)可以看出，当 $\lambda < t_r(m_0, d)$ 时，则学习速率因子就会为正，即使三元组是错误三元组，同样会使得内容相互靠近。当 $\lambda > t_r(m_0, d)$ 时，那么学习速率为负，此时的三元组是错误三元组，同样会使得内容相互远离。但是与原来相比，远离的距离会变小。 $t_r(m_0, d)$ 是一种连续函数，随 m_0 和 d 变化而变化。当 $t_r(m_0, d)$ 是线性函数时，说明随着传播距离的增大，梯度呈线性变化，前一个传递与后一个传递对于 m 来说，损失的度相同；当 $t_r(m_0, d)$ 是指数型函数时，说明随着传播距离的增大，梯度指数型变化，前一个传递与后一个传递对于 m 来说，传播能量在传递过程中能量损失呈指数型变化，即知识传递越远，新知识与 $m+r$ 最初的含义越不相近，如（人，像，狗）这种出现错误三元组的可能性急速增加；当 $t_r(m_0, d)$ 是对数型函数时，说明随着传播距离的增大，梯度呈对数型函数变化，最终会慢慢趋向一个平稳的梯度，前一个传递与后一个传递对于 m 来说，传播能量在传递过程中呈对数型变化，即知识越传递，新知识与 $m+r$ 本身的内容虽越不相近，但是仍然有一定的关系，错误三元组出现的可能性增加，与上一个传播的知识相比，错误三元组出现的概率会缓慢增加。

基于传递关系的本地自适应知识图谱模型构建。主要目的是通过引入传递关系和自适应机制构建优化后的知识图谱。主要分为以下 2 部分：

1. 基于传递关系的知识图谱模型构建。上述优化的随机梯度下降算法可以应用于其他基于嵌入翻译的知识图谱，即可以将优化函数应用到如 TransE、TransH、TransD、TransR 等知识图谱模型中去。对于 TransE 来说，优化函数就是对 $h+r$ 与 t 计算距离后的随机梯度更新；对于 TransH 来说，优化函数就是 h 与 t 映射到 r 的超平面计算距离后的梯度更新；对于 TransD、

TransR 来说, 就是 h 与 t 映射到 r 关系空间中计算距离后的梯度更新。当传播范围 $d=0$ 时, 结合后的 TransE、TransH、TransD、TransR 模型的优化函数就是随机梯度下降算法。

进而, 将基于传递关系的随机梯度下降算法融入到现有的基于嵌入翻译的知识图谱的优化算法中, 形成基于知识传递的知识图谱模型 (TransT)。该模型在知识图谱进行优化时, 充分考虑了三元组中传递性, 也就是知识具有的传递性。当知识具有传递性时, 相互远离的距离相应变小。

2. 结合传递关系和自适应机制的知识图谱构建。知识图谱是将实体和关系嵌入到同一低维度或者不同低维度的向量空间中, 通过最小化损失函数获得的值来进行知识图谱模型优化。

$$L = \sum_{(h,r,t) \in S} \sum_{(h',r,t') \in S'} (\gamma + f_r(h,t) + f_r(h',t')) \quad (7)$$

损失函数是由正确三元组 (h, r, t) 、边缘参数 γ 与错误三元组 (h', r, t') 之间差值构成的。TransE、TransH、TransD 等知识图谱模型都是通过模型训练得出边缘参数的值。因此, 每个损失函数中边缘参数的值都会随着知识图谱模型的不同发生相应的改变。

本地自适应知识图谱模型证明了边缘参数在知识图谱中的重要性以及实体与关系对边缘参数的影响, 但是它并没有考虑知识图谱中的关系间具有的传递性。本步骤将传递关系考虑到 TransA 中, 形成新的知识图谱模型——基于传递关系的本地自适应知识图谱模型 TransAT。

边缘参数随实体和关系进行改变, 这里将其设为 M_{opt} , 其值为:

$$M_{opt} = \mu M_{ent} + (1 - \mu) M_{rel} \quad (6)$$

其中, M_{ent} 表示实体的边缘参数, M_{rel} 表示关系的边缘参数, 参数 μ 控制 M_{ent} 和 M_{rel} 对 M_{opt} 所占的比例, 且 $0 \leq \mu \leq 1$ 。

对于某一个实体 h 和一个关系 r , $P_r = \{t | (h, r, t) \in \Delta\}$ 被称作正确的实体, $T_r = \{t | (h, r, t) \notin \Delta, (h, r, t') \in \Delta, (t', r, t) \in \Delta, (h, r', t) \in \Delta, \exists r' \in R\}$ 被称为传播错误实体, 其中对 r 而言, h 到 t 之间可能存在多个三元组, $N_r = \{t | (h, r, t) \notin \Delta, \exists ((h, r, t') \in \Delta \cup (t', r, t) \notin \Delta, (h, r', t) \in \Delta, \exists r' \in R)\}$ 被称为部分错误实体。当知识图谱不考虑知识的传播性时, $T_r \subseteq N_r$, 此时的知识图谱模型即为本地自适应知识图谱模型 TransA。边缘参数 M_{ent} 由正确的实体和错误实体的前两类共同组成。 M_{ent} 是为了将 P_r 向实体 h 靠近, 而 N_r 和 T_r 远离 h , 其中 T_r 远离 h 一部分, 而另一部分则被认为是正确实体, 保持距离不动, 而 N_r 则彻底远离 h 。 M_{ent} 的值是由正确实体、错误实体、传递实体的间距加权求和获得。那么在基于传递的知识图谱模型中 M_{ent} 的计算如下:

$$\begin{aligned}
M_{ent} = & \frac{\sum_{r \in R_h} \left(\min_{t, t'} \sigma(\|h - t'\| - \|h - t\|) (1 - \sum_{i=1}^d f(i)) \right)}{n} \\
& + \frac{\sum_{r \in R_h} \left(\sum_{i=1}^d f(i) \min_{l, l'} \sigma(\|h - l'\| - \|h - l\|) \right)}{n} \\
& \text{当 } \sigma(x) = \begin{cases} x & \text{当 } x \geq 0 \\ -x & \text{当 } x < 0 \end{cases}
\end{aligned} \tag{4}$$

其中, $t \in P_r$, $t' \in N_r \cup T_r$, $l \in T_r$, $l' \in N_r$, $\sigma(x)$ 为求绝对值的函数, d 表示知识能够传递的最远距离, $f(i)$ 表示传递后每部分实体具有的能量且 $f(i) \geq 0$, 并且整个公式中的传递总能量应该为 1, 所以 $\sum_{i=0}^d f(i) = 0$ 。 R_h 代表实体 h 在三元组集合中涉及的关系集合, r 表示 h 在三元组中涉及的某一个关系, n 表示 R_h 集合中关系的个数。 $\min_{t, t'} \sigma(\|h - t'\| - \|h - t\|)$ 表示获得正确实体和 h 的距离与错误实体和 h 的距离的差值, 而 $\min_{l, l'} \sigma(\|h - l'\| - \|h - l\|)$ 表示获得传递错误实体和 h 的距离与部分错误实体与 h 的距离的差值。当 T_r 不为空, 且在知识进行传递的距离不到 d , 知识就无法继续进行传递下去, 此时就停止对知识传递的求和操作, 不继续进行知识传递。当 $\sum_{i=0}^d f(i)$ 的值较大时, M_{ent} 的值主要取决于传递知识获得的边缘参数部分; 而 $\sum_{i=0}^d f(i)$ 的值较小时, M_{ent} 的值主要取决于正确实体和错误实体获得的边缘参数部分; 在特殊情况下, $\sum_{i=0}^d f(i)$ 为 0 时, M_{ent} 的值完全取决于正确实体和错误实体获得的边缘参数部分, 此时的知识图谱模型为本地自适应知识图谱模型。当 $N_r \cup T_r$ 为空集, 即所有的实体都是与 h 有 r 的关系、或者与 h 完全没有关系时, M_{ent} 为 0。当 T_r 为空集时, 此时的知识图谱模型就成为本地自适应知识图谱模型。

由于采用对实体具有关系的 r 进行统计运算, 且实体包含头实体 h 和尾实体 t 两个部分, 在进行训练时随机替换 h 和 t 其中的一个部分, 剩下的一个实体通过公式(5)计算出其自身的 M_{ent} , 可以是头实体 h 的 M_{ent} , 也可以是尾实体 t 的 M_{ent} , 然后进行基于知识传递的本地自适应知识图模型构建。所以 M_{ent} 可以在 1-to-1、1-to- N 、 N -to-1、 N -to- N 等实体与关系不同比例下均表现出良好的特性。通过对所有具有关联性的实体和关系进行循环求和运算, 然后求出求和运算后的平均值, 就得到了 M_{ent} 。

对于关系的边缘参数部分 M_{ent} , 由于其实体的数量与关系的数量相比, 关系的数量差距太大, 所以在进行知识传递时, 关系能够在距离非常短的情况下传递到大多数其他关系。传播后的三元组会与传输的意义几乎完全不同, 传播时损失的含义会变得非常多, 所以考虑基

于知识的传播在关系上意义不大。基于此，本算法采用本地自适应知识图谱模型中基于关系的边缘参数作为关系部分的边缘参数，其计算如下：

$$M_{rel} = \min_{r_i \in R_{h,r}} (\|r_i\| - \|r\|) \quad (5)$$

其中， $R_{h,r}$ 是一个关系集合， $R_{h,r} = \{r_1, r_2, \dots, r_n\}$ ，表示对于实体 h 除了关系 r 外，涉及到其他关系的集合， $\|r_i\| - \|r\|$ 表示给定的特定关系到 $R_{h,r}$ 集合中的关系的距离， $\min_{r_i \in R_{h,r}} (\|r_i\| - \|r\|)$ 表示给定的关系和 $R_{h,r}$ 中的关系距离的最小差值。且当 $\|r_i\| \geq \|r\|$ 情况不存在、或者 $R_{h,r}$ 为空时， $M_{rel} = 0$ 。

M_{opt} 是由 M_{rel} 与 M_{ent} 共同组成，因此将 M_{rel} 与 M_{ent} 的值带入到 M_{opt} ，即将公式(5)、公式(4)带入公式(6)中，获得基于传递关系的边缘参数 M_{opt} ， M_{opt} 的计算如下：

$$M_{opt} = \mu \left(\frac{\sum_{r \in R_h} \left(\min_{t, t'} \sigma(\|h - t'\| - \|h - t\|) \right) (1 - \sum_{i=1}^d f(i))}{n} + \frac{\sum_{r \in R_h} \left(\sum_{i=1}^d f(i) \min_{t, t'} \sigma(\|h - t'\| - \|h - t\|) \right)}{n} \right) + (1 - \mu) \min_{r_i \in R_{h,r}} (\|r_i\| - \|r\|) \quad (10)$$

将新获得的 M_{opt} 带入到公式(3)中，得到新的基于传递关系的损失函数，即为：

$$L = \sum_{(h, l, t) \in S} \sum_{(h', l', t') \in S'} (M_{opt} + f_r(h, t) - f_r(h', t')) \quad (8)$$

得到的 L 即基于传递关系的本地自适应知识图谱的损失函数。将该损失函数带入知识图谱中，获得相应的基于知识传递的本地自适应知识图谱模型。当实体和关系改变时，该模型能够自动调整边缘参数，不仅能够完成自适应调整，同时能够减少构建模型的复杂度。该模型是对边缘参数进行优化自动调整，所有的知识图谱模型都需要边缘参数对图谱进行优化，因此可以将该模型应用到 TransE、TransH、TransR、TransD 等知识图谱模型中。

附表说明：

表 3 为本发明实施例的数据集描述。其中，WN 和 FB 分别为 Wordnet 和 Freebase 的简写，它们都是公开的数据集。其中，WN11 和 WN18 分别表示 Wordnet 中包含 11 和 13 个关系类型，FB13 和 FB15K 分别表示 Freebase 中包含 13 和 15K 个关系类型。

表 4 为本发明实施例的软硬件环境描述。

表 5 为本发明实施例的链接预测实验结果。

链接预测是指对于一个三元组 (h, r, t) 去掉其中的头实体 h 或者尾实体 t , 预测缺失的实体 h 或者 t 。从中可以看出, 给定 (h, r) 来预测 t , 给定 (r, t) 来预测 h 。链接预测并非从知识图谱中返回一个最好的结果, 而是返回一系列备选的实体结果。本发明从公开数据集中选取 WN18 和 FB15K 这两个数据集进行实验。

在进行数据测试时, 对于每个测试中的三元组 (h, r, t) , 通过使用知识图谱中全部实验实体来分别替换 (h, r, t) 中的 h 和 t , 接着通过评分函数 f_r 计算 (h, r) 与 t 和 (r, t) 与 h 的距离。并通过获得的相似度距离对所有实体进行降序排序。对于知识图谱模型的评价标准采用平均排序和正确实体排在所有测试实体在前十名的概率 (Hits@10)。通常, 对于一个好的链接预测结果应该能够获得一个较低的平均排序的数值或者一个较高的 Hits@10 的数值。然而在进行测试排序时, 实验认为的一个错误三元组也有可能存在于知识图谱中的测试三元组集合中。对于这些错误三元组, 实验中也应该被认为是正确的三元组。然而在测试时, 这些三元组获得非常高的排名, 以至于会大幅度降低知识图谱模型的评价标准。因此为了避免上述情况, 可以通过过滤出现在知识图谱中的错误三元组, 得到真正匹配知识图谱模型的评价指标。本发明将第一种评估设置为“Raw”, 第二这个评价设置为“Filter”。

本发明采用多种方法进行对比实验, 其中包括基于嵌入翻译的知识图谱模型, 例如 TransE、TransH、TransR, 以及一些基于能量的知识图谱模型, 例如 SME、Unstructured 等方法。本实验选择边缘参数 γ 在 $\{1, 0.1, 0.01\}$ 之中, 随机梯度下降的学习速率 λ 在 $\{1, 0.1, 0.01, 0.001\}$ 之中, 低维度空间的维度 n 在 $\{20, 50\}$ 之间, 知识的传播距离 d 在 $\{1, 2, 3\}$ 之间, 相似度函数即距离空间在欧几里得距离 L_1 和曼哈顿距离 L_2 之间, 传播的能量损失函数在线性函数、指数函数和对数函数之间。在 WN18 数据集中, $n=20$, $\lambda=0.01$, $d=2$, $\mu=0.6$, 传播的能量损失函数为对数函数, 取 L_1 为相似性度量距离; 在 FB15K 数据集中, $n=50$ 、 $\lambda=1$ 、 $d=1$ 、 $\mu=0.5$, 传播的能量损失函数为指数函数, 取 L_2 为相似性度量距离。

从表 5 的实验结果可见, 在两个数据集上, TransAT 在平均排序方面均优于其他方法。而在 Hits@10 时, 由于考虑到了知识的传递性, 存在一些没有传递的知识却占据较好位置的实体, 使得在进行排序时, 不能获得前 10 的排名, 使得在 Hits@10 测试集上, 总体效果略差于 TransA。

表 6 为本发明实施例的三元组分类实验结果。

三元组分类的目的是预测一个给定三元组 (h, r, t) 是正确三元组还是错误三元组。这是一个二元分类问题。由于数据集中没有公开表明哪些是正确三元组, 哪些是错误三元组, 因此采用链接预测使用的方法来构造错误三元组。具体的分类方法是为每一个关系 r 设置其特有的阈值 δ_r , 对于一个三元组 (h, r, t) , 如果 $f_r(h, t)$ 的相似性度量小于该阈值, 则该三元组被

认为是正确三元组，否则为错误三元组。该阈值 δ_r 通过验证集最大化结果来优化。

在 WN11 数据集中， $n=20$ ， $\lambda=0.1$ ， $d=2$ ， $\mu=0.6$ 传播的能量损失函数为对数函数，取 L_1 为相似性度量距离；在 FB13 数据集中， $n=50$ 、 $\lambda=0.1$ 、 $d=1$ 、 $\mu=0.5$ ，传播的能量损失函数为指数函数、取 L_2 为相似性度量距离。

从表 6 的实验结果可见，在 WN11 中，TransAT 效果优于 TransA，因为在此数据集中，知识传递有可靠的基础。而在 FB13 中则没有，因而在 FB13 中效果略差于 TransA，并比其他方法相比，如 NTN 仍有一定的差距。

综合表 5 和表 6 的实验结果可见，在具有知识传递可靠基础的数据集中，本发明提出的基于传递关系的本地自适应知识图谱优化方法 TransAT 总体上占据优势。

表 3 本发明实验采用的数据集描述

数据集	关系	实体	训练集	验证集	测试集
WN11	11	38,696	112,581	2,609	10,544
WN18	18	40943	141,442	5,000	5,000
FB13	13	75,043	316,232	5,908	23,733
FB15K	1,345	14,951	483,142	50,000	59,071

表 4 本发明实验的软硬件环境

硬件	CPU	Inter Core i7-2600 CPU @ 3.40GHz
	内存	8.00GB
	硬盘	500GB
软件	操作系统	Windows 7-64 位
	开发语言	C++
	开发工具	CodeBlock

表 5 本发明链接预测的实验结果

数据集	WN18				FB15K			
评价标准	Mean Rank		Hits@10(%)		Mean Rank		Hits@10(%)	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
Unstructured	315	304	35.3	38.2	1074	979	4.5	6.3
RESCAL	1180	1163	37.2	52.8	828	683	28.4	44.1
SE	1011	985	68.5	80.5	273	162	28.8	39.8
SME(linear)	545	533	65.1	74.1	274	154	30.7	40.8
SME(bilinear)	526	509	54.7	61.3	284	158	31.3	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	318	303	75.4	86.7	212	87	45.7	64.4
TransD	224	212	79.6	92.2	194	91	53.4	77.3
TransR	232	219	78.3	91.7	168	77	48.2	68.7
CTransR	231	218	79.4	92.3	199	75	48.4	70.2
TransA	165	153	90.6	93.2	164	58	50.7	76.5
TransAT	160	149	88.2	91.1	162	55	47.9	74.6

表 6 本发明三元组分类准确度实验结果(%)

数据集	WN11	FB13
SE	53.0	75.2
SME (linear)	70.0	63.7
SLM	69.9	85.3
LFM	73.8	84.3
NTN	70.4	87.1
TransE	75.9	81.5
TransH	78.8	83.3
TransR	85.9	82.5
TransA	93.2	82.8
TransAT	93.8	82.2

权 利 要 求 书

1. 一种基于传递关系的本地自适应知识图谱优化方法，其特征在于，所述基于传递关系的本地自适应知识图谱优化方法，具体步骤如下：

步骤 1：设定训练样本集为 $S=\{(h_i, r_i, t_i)\}$ ， h_i 为一个头实体向量， t_i 为一个尾实体向量， r_i 为一个连接头实体和尾实体的关系向量， $i=1, 2, \dots, N$ ， $h_i \in E$ ， $t_i \in E$ ， $r_i \in R$ ，其中， E 为实体集， R 为关系集，设定嵌入空间维度为 n ，知识能够传递的最远距离为 d ，约束因子为 μ ；

步骤 2：设定任意 r_i 初始属于某一分布，并且 $r_i \in R$ ；设定任意 e_i 初始属于某一分布，与 r_i 所属分布为同一分布或者为不同分布，并且 $e \in E$ ， e 为 h 或者 t ；

步骤 3：归一化 r_i ， h_i ， t_i ，归一化公式分别为：

$$r_i = \frac{r_i}{\|r_i\|} \quad (1)$$

$$h_i = \frac{h_i}{\|h_i\|} \quad (2)$$

$$t_i = \frac{t_i}{\|t_i\|} \quad (3)$$

其中， $\|r_i\|$ 、 $\|h_i\|$ 、 $\|t_i\|$ 分别为 r_i ， h_i ， t_i 的一阶范数；

步骤 4：从训练样本集中随机抽取样本大小为 b 的样本，组成新的训练样本集 S_{batch} ：
 $S_{batch} = \text{sample}(S, b)$ ；其中， $\text{sample}(S, b)$ 表示在训练样本集 S 中随机抽取 b 个样本；

步骤 5：初始化三元组集合为空集 $T_{batch} = \Phi$ ；

步骤 6：基于归一化后的 r_i ， h_i ， t_i ，设定 $(h, r, t) \in S_{batch}$ 为正确的三元组，其对应 (h', r, t') 属于错误的三元组，用错误的三元组头实体或者尾实体替换正确的三元组，组成错误训练样本集 $S(h', r, t')$ ，并且合并到三元组集合 T_{batch} 中，得到合并后三元组集合 $T_{batch} = T_{batch} \cup \{(h', r, t'), (h, r, t)\}$ ；

步骤 7：对于实体 h 或者 t ，获得其实体的边缘参数 M_{ent} ，公式如下：

$$M_{ent} = \frac{\sum_{r \in R_h} \left(\min_{t, t'} \sigma(\|h - t'\| - \|h - t\|) (1 - \sum_{i=1}^d f(i)) \right)}{n} + \frac{\sum_{r \in R_h} \left(\sum_{i=1}^d f(i) \min_{l, l'} \sigma(\|h - l'\| - \|h - l\|) \right)}{n} \quad (4)$$

$$\text{当 } \sigma(x) = \begin{cases} x & \text{当 } x \geq 0 \\ -x & \text{当 } x < 0 \end{cases}$$

其中， $t \in P_r$ ， $t' \in N_r \cup T_r$ ， $l \in T_r$ ， $l' \in N_r$ ， $\sigma(x)$ 为求绝对值的函数， d 表示知识能够传递的最远距离， $f(i)$ 表示传递后每部分实体具有的能量且 $f(i) \geq 0$ ，并且整个公式中的传递总能量应该

为 1, 所以 $\sum_{i=0}^d f(i) = 0$, R_h 代表实体 h 在三元组集合中涉及的关系集合, r 表示 h 在三元组中涉及的某一个关系, n 表示 R_h 集合中关系的个数, $\min_{t,t'} \sigma(\|h-t'\| - \|h-t\|)$ 表示获得正确实体和 h 的距离与错误实体和 h 的距离的差值, 而 $\min_{l,l'} \sigma(\|h-l'\| - \|h-l\|)$ 表示获得传递错误实体和 h 的距离与部分错误实体与 h 的距离的差值;

步骤 8: 对于关系 r , 获得关系的边缘参数 M_{rel} , 公式如下:

$$M_{rel} = \min_{r_i \in R_{h,r}} (\|r_i\| - \|r\|) \quad (5)$$

其中, $R_{h,r}$ 是一个关系集合, $R_{h,r} = \{r_1, r_2, \dots, r_n\}$, 表示对于实体 h 除了关系 r 外, 涉及到其他关系的集合, $\|r_i\| - \|r\|$ 表示给定的特定关系到 $R_{h,r}$ 集合中的关系的距离, $\min_{r_i \in R_{h,r}} (\|r_i\| - \|r\|)$ 表示给定的关系和 $R_{h,r}$ 中的关系距离的最小差值, 且当 $\|r_i\| \geq \|r\|$ 情况不存在、或者 $R_{h,r}$ 为空时, $M_{rel} = 0$;

步骤 9: 计算边缘参数随实体和关系进行变化的参数, 公式如下:

$$M_{opt} = \mu M_{ent} + (1 - \mu) M_{rel} \quad (6)$$

其中, M_{ent} 表示实体的边缘参数, M_{rel} 表示关系的边缘参数, 参数 μ 控制 M_{ent} 和 M_{rel} 对 M_{opt} 所占的比例, 且 $0 \leq \mu \leq 1$;

步骤 10: 将新获得的 M_{opt} 带入到损失函数中, 得到新的基于传递关系的损失函数, 即为:

$$L = \sum_{(h,l,t) \in S} \sum_{(h',l',t') \in S'} (M_{opt} + f_r(h,t) - f_r(h',t')) \quad (7)$$

其中, f_r 是基于关系 r 的距离函数, 具体地, $f_r(h, t)$ 为 $h+r$ 与 t 之间的距离函数, $f_r(h', t')$ 为 $h'+r$ 与 t' 之间的距离函数;

步骤 11: 判断若 $M_{opt} + f_r(h, t) > f_r(h', t')$, 输入三元组集合 T_{batch} , 使用随机梯度下降函数 SGD 对每个实体或者关系向量进行优化, 得到优化后的 E 和 R 向量集合, 即完成知识图谱优化, 若判断 $M_{opt} + f_r(h, t) \leq f_r(h', t')$, 无需进行优化。

2. 根据权利要求 1 所述基于传递关系的本地自适应知识图谱优化方法, 其特征在于, 所述随机梯度下降函数 SGD 进行优化, 公式如下:

$$m' = m \pm (\lambda - t_r(m_0, d)) \nabla f \quad (8)$$

其中, m 表示实体或者关系, m' 表示优化后的实体或者关系, λ 表示随机梯度下降的学习速率, ∇f 表示对三元组 $h+r$ 与 t 的相似度距离求导的结果, m_0 与 m 分别表示知识传递的启发实体和传播收到的实体, $t_r(m_0, d)$ 表示对于关系 r 的基于知识传递的范围函数, 且 $0 < t_r(m_0, d) < 1$,

d 表示知识传播的距离, 且 $d > 0$ 。

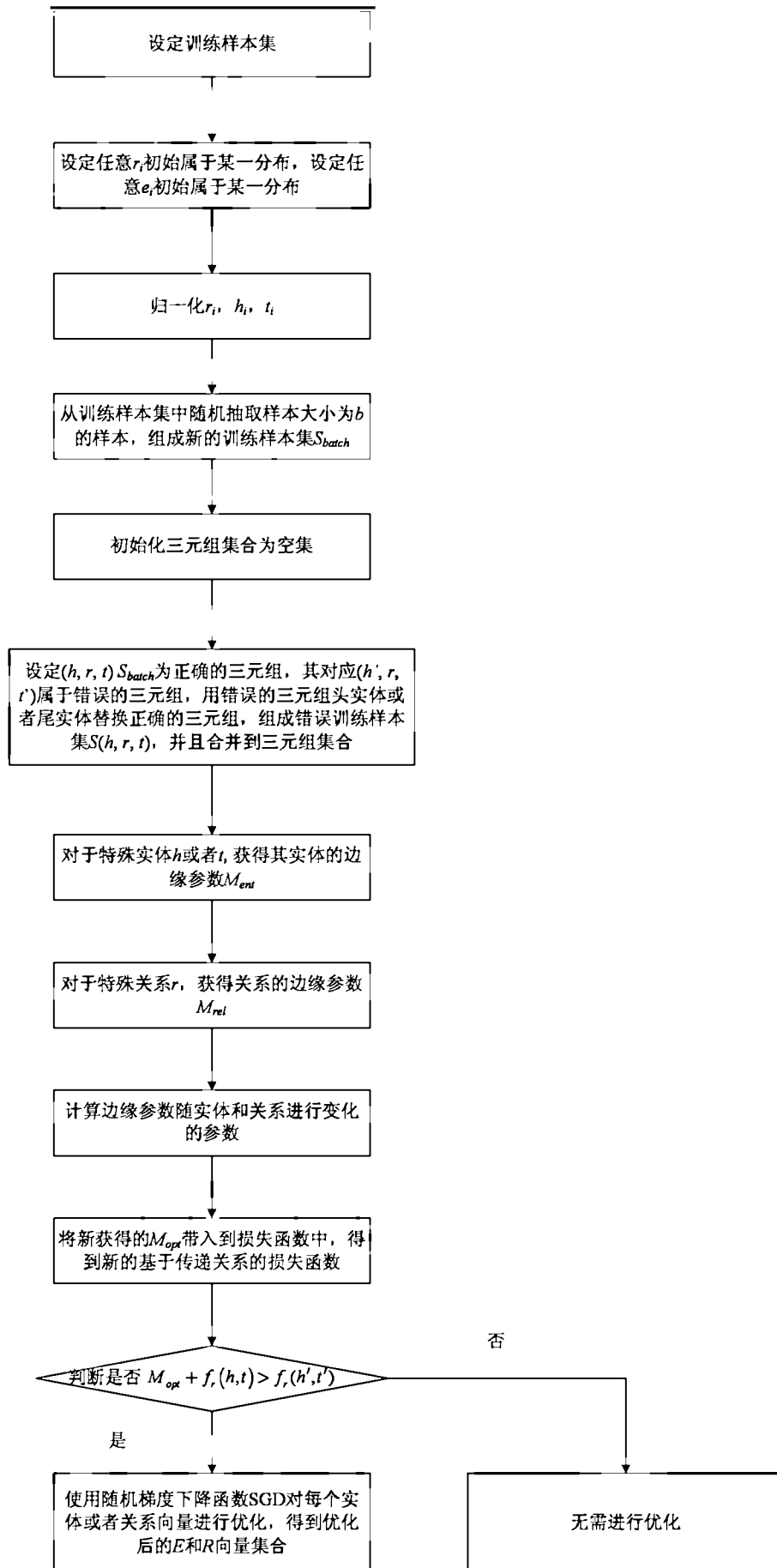


图 1

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/077728

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/36(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPTXT; USTXT; WOTXT; VEN; CNABS; CNTXT; CNKI: 东北大学, 知识图谱, 知识, 图谱, 传递关系, 关系, 距离, 梯度下降, 下降, 随机, 训练, 边缘, 最远, 传递, 梯度, 损失, 函数, SGD, transA, Knowledge, relationship?, descent?, gradient?, loss, optimiz+, edge+, map?, transfer, sample+, distance?, training+, graph?,

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 105824802 A (TSINGHUA UNIVERSITY) 03 August 2016 (2016-08-03) description, paragraphs [0008]-[0046] and [0052]-[0129], and figures 1-3	1, 2
A	CN 108959328 A (RICOH CO., LTD.) 07 December 2018 (2018-12-07) entire document	1, 2
A	CN 108446769 A (SHENZHEN AXMTEC CO., LTD.) 24 August 2018 (2018-08-24) entire document	1-2
A	US 2019057310 A1 (ROYAL BANK OF CANADA) 21 February 2019 (2019-02-21) entire document	1, 2
A	CN 109063021 A (ZHEJIANG UNIVERSITY) 21 December 2018 (2018-12-21) entire document	1-2
A	CN 107391677 A (SHENZHEN GOWILD ROBOTICS CO., LTD.) 24 November 2017 (2017-11-24) entire document	1, 2
A	CN 107784088 A (HANGZHOU BOSHI DATA NETWORK CO., LTD.) 09 March 2018 (2018-03-09) entire document	1, 2



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

11 September 2019

Date of mailing of the international search report

18 October 2019

Name and mailing address of the ISA/CN

**China National Intellectual Property Administration (ISA/
CN)**
**No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088**
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/077728

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	105824802	A	03 August 2016	CN	105824802	B	30 October 2018
CN	108959328	A	07 December 2018	US	2018341863	A1	29 November 2018
CN	108446769	A	24 August 2018	None			
US	2019057310	A1	21 February 2019	CA	3014309	A1	16 February 2019
CN	109063021	A	21 December 2018	None			
CN	107391677	A	24 November 2017	None			
CN	107784088	A	09 March 2018	None			

国际检索报告

国际申请号

PCT/CN2019/077728

A. 主题的分类

G06F 16/36 (2019.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

EPTXT;USTXT;WOTXT;VEN;CNABS;CNTXT;CNKI: 东北大学, 知识图谱, 知识, 图谱, 传递关系, 关系, 距离, 梯度下降, 下降, 随机, 训练, 边缘, 最远, 传递, 梯度, 损失, 函数, SGD, transA, Knowledge, relationship?, descent?, gradient?, loss, optimiz+, edge+, map?, transfer, sample+, distance?, training+, graph?,

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 105824802 A (清华大学) 2016年 8月 3日 (2016 - 08 - 03) 说明书第[0008]-[0046]、[0052]-[0129]段、附图1-3	1-2
A	CN 108959328 A (株式会社理光) 2018年 12月 7日 (2018 - 12 - 07) 全文	1-2
A	CN 108446769 A (深圳市阿西莫夫科技有限公司) 2018年 8月 24日 (2018 - 08 - 24) 全文	1-2
A	US 2019057310 A1 (ROYAL BANK OF CANADA) 2019年 2月 21日 (2019 - 02 - 21) 全文	1-2
A	CN 109063021 A (浙江大学) 2018年 12月 21日 (2018 - 12 - 21) 全文	1-2
A	CN 107391677 A (深圳狗尾草智能科技有限公司) 2017年 11月 24日 (2017 - 11 - 24) 全文	1-2
A	CN 107784088 A (杭州博世数据网络有限公司) 2018年 3月 9日 (2018 - 03 - 09) 全文	1-2

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2019年 9月 11日

国际检索报告邮寄日期

2019年 10月 18日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

梁玮

电话号码 (86-20)28950043

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/077728

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	105824802	A	2016年 8月 3日	CN 105824802 B	2018年 10月 30日
CN	108959328	A	2018年 12月 7日	US 2018341863 A1	2018年 11月 29日
CN	108446769	A	2018年 8月 24日	无	
US	2019057310	A1	2019年 2月 21日	CA 3014309 A1	2019年 2月 16日
CN	109063021	A	2018年 12月 21日	无	
CN	107391677	A	2017年 11月 24日	无	
CN	107784088	A	2018年 3月 9日	无	