(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106909622 A
(43)申请公布日 2017.06.30

(21)申请号 201710041593.X

(22)申请日 2017.01.20

(71)申请人 中国科学院计算技术研究所
地址 100190 北京市海淀区中关村科学院南路6号

(72)发明人 程学旗 贾岩涛 李曼玲 王元卓 靳小龙 苏佳林

(74)专利代理机构 北京泛华伟业知识产权代理有限公司 11280
代理人 王勇

(51)Int.Cl.
G06F 17/30(2006.01)

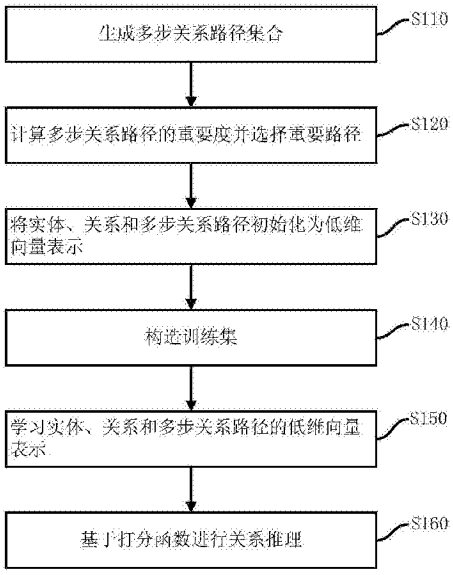
权利要求书1页 说明书8页 附图1页

(54)发明名称

知识图谱向量表示方法、知识图谱关系推理方法及系统

(57)摘要

本发明提供了一种知识图谱的向量表示方法。该方法包括:将知识图谱中的实体对、关系和实体对之间的多步关系路径表示为初始低维向量;利用间隔可变的损失函数训练实体对、关系和实体对之间的多步关系路径的低维向量表示的模型。利用本发明学习到的模型进行关系推理能够提高不同的知识图谱中的推理的精确度。



CN 106909622 A

1.一种知识图谱的向量表示方法，包括：

步骤1：将知识图谱中的实体对、关系和所述实体对之间的多步关系路径表示为初始低维向量；

步骤2：利用间隔可变的损失函数训练所述实体对、关系和所述实体对之间的多步关系路径的低维向量表示的模型。

2.根据权利要求1所述的方法，其中，所述损失函数包括实体对和多步关系路径的损失。

3.根据权利要求2所述的方法，其中，所述实体对和多步关系路径的损失定义为：

$L_{p,r}=[||p-r||+M_{path}(p)-||p-r'||]_+$

其中，$M_{path}(p)$为多步关系路径正反例之间的间隔，定义为$M_{path}(p)=\min_{r,r'}|||p-r'||-||p-r|||$，p是多步关系路径的向量表示；r′为负例关系$r'\in N_{h,t}$中的低维向量表示向量，$N_{h,t}$为知识图谱中负例三元组(h,r′,t)的集合；||·||表示L_1或L_2范式。

4.根据权利要求2所述的方法，其中，所述损失函数还包括实体对和关系的损失。

5.根据权利要求4所述的方法，其中，所述实体对和关系的损失定义为：

$$E_{h,r,t}=\frac{1}{Z}\sum_{(h',r',t')\in\Delta'}[\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|+\gamma-\|\boldsymbol{h}'+\boldsymbol{r}'-\boldsymbol{t}'\|]_+$$

其中，Δ为三元组(h,r,t)构成的训练集，h为头实体，t为尾实体，r表示两者间的关系；Z表示训练集Δ的模数；Δ′表示实体对和关系的负例三元组，三元组$(h',r',t')\in\Delta'$，是将$(h,r,t)\in\Delta$中的h,r,t替换为h′,r′,t′所获得的；$h,r,t,h',r',t'\in R^d$，R^d表示维数为d的低维向量空间；$[x]_+$返回x与0两者中的较大值；||·||表示L_1或L_2范式；γ是三元组的正例和负例之间的间隔。

6.根据权利要求1所述的方法，其中，所述多步关系路径的长度低于阈值。

7.根据权利要求1所述的方法，其中，在步骤2的训练过程中，采用梯度下降法来更新所述损失函数。

8.一种知识图谱的关系推理方法，包括：

步骤11：根据目标实体从知识图谱中找出所有与该目标实体之间有目标关系的所有候选实体作为候选实体集；

步骤12：根据权利要求1-7中的任一项获得的低维向量表示的模型来推断所述候选实体集中可能与该目标实体存在所述目标关系的候选实体。

9.根据权利要求8所述的方法，其中，在步骤12中，采用下式来推断所述候选集中的候选实体与目标实体存在所述目标关系的可能性：

$f(h,r,t)=||h+r-t||$

其中，粗体h,r,t表示h,r,t在低维向量空间的向量表示，||·||表示L_1或L_2范式。

10.一种知识图谱的关系推理系统，包括：

用于根据目标实体从知识图谱中找出所有与该目标实体之间有目标关系的所有候选实体作为候选实体集的装置；

用于根据权利要求1-7中的任一项所获得的低维向量表示的模型来推断所述候选实体集中可能与该目标实体存在所述目标关系的候选实体的装置。

知识图谱向量表示方法、知识图谱关系推理方法及系统

技术领域

[0001] 本发明涉及信息处理技术领域，具体涉及一种基于特定路径翻译的知识图谱的关系推理方法和系统。

背景技术

[0002] 知识图谱是知识工程中以图的形式组织的知识集群，其由不同类型的实体作为节点、关系作为连接节点的边所构成的。在知识图谱中，实体指真实世界中的客观物体(例如，贝拉克·奥巴马)，或者人类思想中的抽象概念(例如，美国的第44届总统)。关系则是描述两个实体之间的实际关系(例如，贝拉克·奥巴马是美国的第44届总统，即贝拉克·奥巴马与美国的第44届总统之间存在“是”的关系)。

[0003] 在已知的知识图谱中，实体类型有人物、事件、组织机构、地点等，而它们之间的关系类型也十分多样化。不同的实体类型所关注的关系也是不同的。例如，对于人物实体之间，常见关系有亲人及朋友关系；对于人与组织机构之间，常见关系有工作单位、毕业院校等。这些已知的实体间的关系在原始的知识图谱中比较稀疏，而实际上实体间还存在大量的隐含关系，可以通过知识图谱中已有的知识和关系，来发掘或推理这些隐含关系。

[0004] 最常用的推理方法是基于规则的方法，即通过对已有知识的分析，制定合适的推断规则，最终由这些规则推出实体间的关系。但这种方法由人工来制定规则，工作量很大且能制定的规则数量有限，涵盖范围较小，具有较大的局限性。为了减少规则的人工标注量，另一个常用的知识图谱的关系推理方法是根据已有知识通过机器学习自动地获得规则，例如，利用现有的transE、transR、transH等基于翻译的模型，但这种方法的效果较为依赖于对特征的选择和模型的参数的选择，在不同领域的知识图谱中迁移需要花费较多精力，例如，对于学术领域的关系推理，如合作关系等，更侧重于研究热点的内容相似度特征，且有效路径长度通常较短；而人物关系领域的关系推理更侧重于结构相似度特征，且有效路径长度可能较长，因此，在实际应用中，不同领域的知识图谱之间的迁移具有局限性。此外，在传统的自动学习方法的模型中，通常是通过基于间隔的损失函数来衡量学习的精确度，间隔通常从候选值中预先选择，而且，该间隔在学习过程中是固定不变的。这种固定不变的间隔不能自适应的调节不同知识图谱、不同的实体和关系的学习的精确性。

发明内容

[0005] 本发明的目的在于克服上述现有技术中的缺陷，提供一种改进的知识图谱的关系推理方法。

[0006] 根据本发明的第一方面，提供了一种知识图谱的向量表示方法，包括：

[0007] 步骤1：将知识图谱中的实体对、关系和所述实体对之间的多步关系路径表示为初始低维向量；

[0008] 步骤2：利用间隔可变的损失函数训练所述实体对、关系和所述实体对之间的多步关系路径的低维向量表示的模型。

[0009] 优选地，所述损失函数包括实体对和关系的损失以及实体对和多步关系路径的损失。

[0010] 优选地，所述实体对和关系的损失定义为：

[0011] $$E_{h,r,t}=\frac{1}{Z}\sum_{(h',r',t')\in\Delta'}[\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|+\gamma-\|\boldsymbol{h}'+\boldsymbol{r}'-\boldsymbol{t}'\|]_+$$

[0012] 其中，Δ为三元组(h,r,t)构成的训练集，h为头实体，t为尾实体，r表示两者间的关系；Z表示训练集Δ的模数；Δ′表示实体对和关系的负例三元组，三元组$(h',r',t')\in\Delta'$，是将$(h,r,t)\in\Delta$中的h,r,t替换为h′,r′,t′所获得的；$h,r,t,h',r',t'\in R^d$，R^d表示维数为d的低维向量空间；$[x]_+$返回x与0两者中的较大值；||·||表示L_1或L_2范式；γ是三元组的正例和负例之间的间隔。

[0013] 优选地，所述实体对和多步关系路径的损失定义为：

[0014] $L_{p,r}=[||p-r||+M_{path}(p)-||p-r'||]_+$

[0015] 其中，$M_{path}(p)$为多步关系路径正反例之间的间隔，定义为$M_{path}(p)=\min_{r,r'}|||p-r'||-||p-r|||$，p是多步关系路径的向量表示；r′为负例关系$r'\in N_{h,t}$中的低维向量表示向量，$N_{h,t}$为知识图谱中负例三元组(h,r′,t)的集合；||·||表示L_1或L_2范式。

[0016] 优选地，所述多步关系路径的长度低于阈值。

[0017] 优选地，在步骤2中的训练过程中，采用梯度下降法来更新所述损失函数。

[0018] 根据本发明的第二方面，提供了一种知识图谱的关系推理方法。该方法包括：

[0019] 步骤11：根据目标实体从知识图谱中找出所有与该目标实体之间有目标关系的所有候选实体作为候选实体集；

[0020] 步骤12：根据知识图谱的向量表示方法来推断所述候选实体集中可能与该目标实体存在所述目标关系的候选实体。

[0021] 优选地，在步骤12中，采用下式来推断所述候选集中的候选实体与目标实体存在所述目标关系的可能性：

[0022] $f(h,r,t)=||h+r-t||$

[0023] 其中，粗体h,r,t表示h,r,t在低维向量空间的向量表示，||·||表示L_1或L_2范式。

[0024] 根据本发明的第三方面，提供了一种知识图谱的关系推理系统。该系统包括：用于根据目标实体从知识图谱中找出所有与该目标实体之间有目标关系的所有候选实体作为候选实体集的装置；用于根据本发明的知识图谱的向量表示方法来推断所述候选实体集中可能与该目标实体存在所述目标关系的候选实体的装置。

[0025] 与现有技术相比，本发明的优点在于：

[0026] 利用知识图谱中已有的实体关系和实体自动学习推断规则，利用间隔可变的损失函数进行学习可以自适应地建立实体对间的关系和多步关系路径之间的联系，提高学习的精确性。对于不同的知识图谱，本发明能够自适应地计算损失函数的最优间隔值，不需要提前定义任何候选间隔值。与其他模型相比，复杂度相同，但关系推理效果得到了明显提高。

附图说明

[0027] 被结合在说明书中并构成说明书的一部分的附图示出了本发明的实施例，并且连

同其说明一起用于解释本发明的原理。

[0028] 图1示出了根据本发明的一个实施例的知识图谱的关系推理方法的流程图。

[0029] 图2示出了根据本发明的一个实施例的特定路径间隔的示意图。

具体实施方式

[0030] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图通过具体实施例对本发明进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。

[0031] 简言之，根据发明的实施例的方法包括：将实体对、关系和实体对之间的多步关系路径表示为初始低维向量；设计学习目标(损失函数)，基于特定路径翻译的模型来训练学习，通过不断调整，更新实体、关系和向量的取值使损失函数最小；利用训练好的这些向量进行实体预测或关系推理等。

[0032] 在本文中，所述的特定路径翻译是指，在基于现有的路径翻译的知识图谱表示学习过程中，对每个多步关系路径计算特定的间隔，以更好地学习多步关系路径的低维向量，即每个多步关系路径的间隔是特定的，因此，可以称为基于特定路径翻译的模型。本发明采用了间隔可变的损失函数进行训练学习，进一步增加了学习过程的自适应性和学习的精确度。具体地，图1示出了根据本发明的一个实施例的知识图谱的关系推理方法的流程图。

[0033] 1)步骤S110，生成多步关系路径集合。

[0034] 在本文中，多步关系路径是指实体对之间存在目标关系的路径，路径的长度可以取值大于或等于1。例如，$A \xrightarrow{\text{工作于}} v_0 \xleftarrow{\text{工作于}} B$的，其路径的长度是2，通常用“-工作于-工作于$^{-1}$-”来表示。当路径的长度超过阈值时，认为该实体对之间的关联关系较弱，可以不予考虑。

[0035] 在一个实施例中，获取多步路径关系的集合的方式是根据给定的目标关系从现有的知识库中找出所有存在该关系的实体对；对于每个实体对，从头实体开始，直至到达尾实体或超过设定的路径阈值，获得多步关系路径的集合。现有的知识库包括但不限于Freebase、Google Knowledge Graph、GeneOntology等。

[0036] 例如，如果查找的目标关系是“coauthor(合著)”，则遍历所有存在该关系的实体对，可能存在“<A-B>”，“<C-D>”等。例如，对于<A-B>实体对，可使用宽度优先搜索遍历方法来获得它们之间的所有路径集合。通过这种方式可以最终获得所有实体对的多步关系路径的集合。

[0037] 2)步骤S120，计算多步关系路径的重要度并选择重要路径。

[0038] 在此步骤中，通过计算每对实体之间的每条多步关系路径的得分R(p|h,t)来衡量其重要度。R(p|h,t)是从头实体h(head)到达尾实体t(tail)的资源量，用于衡量实体对之间的关系路径的重要度，分值越高表示该条路径越重要。

[0039] 例如，对于实体对(h,t)，如果该实体对之间的关系路径p是$h \xrightarrow{r_1} S_1 \xrightarrow{r_2} \ldots \xrightarrow{r_l} t$，令$R_p(h)=1$，$R([|h,t)=R_p(t)$，且对于实体$m \in S_i$，m的连接关系$r_i$的前置节点表示为$S_{i-1}(\cdot, m)$，则该条关系路径的得分计算为：

[0040] $$R(p|h,t)=R_p(t)=\sum_{n\in S_{i-1}(\cdot,t)}\frac{1}{|S_i(n,\cdot)|}R_p(n) \qquad (1)$$

[0041] 其中，得分的取值范围是0到1之间；•表示任意节点。例如，$S_i(n,\bullet)$表示以n为起点的任意点的集合。

[0042] 具体而言，如果知识图谱中，对于某一实体对，有路径$h\xrightarrow{r1}a\xrightarrow{r2}t$和$h\xrightarrow{r1}b\xrightarrow{r2}t$，及$a\xrightarrow{r3}t'$，

[0043] 那么，$S_1(\bullet,t)=\{a,b\}$，并且，$S_1(h,\bullet)=\{a,b\}$，$S_2(a,\bullet)=\{t,t'\}$，$S_2(b,\bullet)=\{t\}$，则：

[0044] $$R_p(a)=\sum_{n\in\{h\}}\frac{1}{|S_1(n,\cdot)|}R_p(n)=\frac{1}{|S_1(h,\cdot)|}R_p(h)=\frac{1}{2}R_p(h)=\frac{1}{2}$$

[0045] $$R_p(b)=\sum_{n\in\{h\}}\frac{1}{|S_1(n,\cdot)|}R_p(n)=\frac{1}{|S_1(h,\cdot)|}R_p(h)=\frac{1}{2}R_p(h)=\frac{1}{2}$$

[0046]

$$R(p|h,t)=R_p(t)=\sum_{n\in\{a,b\}}\frac{1}{|S_2(n,\cdot)|}R_p(n)=\frac{1}{|S_2(a,\cdot)|}R_p(a)+\frac{1}{|S_2(b,\cdot)|}R_p(b)$$

$$=\frac{1}{2}R_p(a)+\frac{1}{1}R_p(b)=\frac{1}{4}+\frac{1}{2}=\frac{3}{4}$$

[0047] 通过该公式(1)可以计算出每条关系路径的得分，在本实施例中，可选地可以仅仅选择重要的路径(例如，对于每对实体对，选择得分最高的三条路径)来执行以下过程。通过这种方式，可以降低计算量，提高学习的效率。

[0048] 3)步骤S130，将实体、关系和多步关系路径初始化为低维向量表示。

[0049] 在此步骤中将已经获得的知识图谱中的实体、关系、多步关系路径，初始化为用向量来表示，向量的维度一般在0-300之间。在一个实施例中，采用平均分布或伯努利分布来进行初始化。用三元组(h,r,t)来表示头尾实体和关系，用p来表示路径。

[0050] 通过这种低维向量表示的方法可以将所有对象(实体、关系等)映射到这个低维空间中。

[0051] 4)步骤S140，构造训练集合。

[0052] 在此实施例中，构造两个训练集合用于模型的训练，第一个训练集合由三元组(h,r,t)的正例和负例组成；第二个训练集合由是实体对和多步关系路径的正例和负例组成。

[0053] 例如，已知李某和张某存在“合著”关系，则三元组(h,r,t)是(李某,合著,张某)，这个三元组是真实存在的，因此，是正例三元组。通过随机替换h、r、t中的至少一项可以得到对应的负例三元组，例如，(李某,合著,沈某)。该实体对对应的多步关系路径可能有p_1=“-工作于-工作于$^{-1}$-”、p_2=“-导师-导师$^{-1}$-”、p3=“-发表于-发表于$^{-1}$-”等，通过替换r可以获得实体对和多步关系路径对应负例，例如，(李某,夫妻,张某)，该实体对的多步关系路径应该与“合著”对应，现在与“夫妻”对应。

[0054] 5)步骤S150，学习实体、关系和多步关系路径的低维向量表示。

[0055] 此步骤的目的是通过不断调整、更新实体、关系、多步关系路径等向量的取值使其更接近于真实结果。用损失函数来衡量与真实结果的接近程度，即学习目标是通过优化使损失函数取值最小。

[0056] 定义损失函数为：

[0057] $$L(E,z)=\sum_{(h,r,t)\in\Delta}\left[E_{h,r,t}+\frac{1}{Z}\sum_{p\in P_{h,t}}R(p|h,t)\sum_{(h,r',t)\in\Delta''}L_{p,r}\right] \quad (2)$$

[0058] 其中，L(E,z)表示低维向量学习过程的损失函数；E表示低维向量表示；z表示三元组(h,r,t)，h为头实体，t为尾实体，r为两者间的关系；Δ为知识图谱中的已知三元组的集合，作为学习过程中的训练集；Z表示训练集Δ的模数，即Δ中三元组的个数；$P_{h,t}=\{p_1,p_2,\cdots,p_m\}$表示从h到t的多步关系路径的集合；$L_{p,r}$表示多步关系路径p和关系r的损失；Δ″表示实体对和多步关系路径的负例，用于优化$\frac{1}{Z}\sum_{p\in P_{h,t}}R(p|h,t)\sum_{(h,r',t)\in\Delta''}L_{p,r}$。

[0059] 在公式(2)中，第一部分$E_{h,r,t}$表示实体对和关系的损失；第二部分$\frac{1}{Z}\sum_{p\in P_{h,t}}R(p|h,t)\sum_{(h,r',t)\in\Delta''}L_{p,r}$表示实体对和多步关系路径的损失，其中，R(p|h,t)为衡量实体对(h,t)之间的关系路径p重要度的得分。

[0060] 计算公式(2)的具体步骤包括：

[0061] a)计算实体对和关系的损失$E_{h,r,t}$。

[0062] 实体对和关系的损失函数定义为：

[0063] $$E_{h,r,t}=\frac{1}{Z}\sum_{(h',r',t')\in\Delta'}[\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|+\gamma-\|\boldsymbol{h}'+\boldsymbol{r}'-\boldsymbol{t}'\|]_+ \quad (3)$$

[0064] 其中，Δ为三元组构成的训练集；Z表示训练集Δ的模数，即Δ中三元组的个数；Δ′表示实体对和关系的负例三元组，三元组$(h',r',t')\in\Delta'$，是将$(h,r,t)\in\Delta$中的h,r,t替换为h′,r′,t′所获得的，h为头实体，t为尾实体，r表示两者间的关系；$h,r,t,h',r',t'\in R^d$，R^d表示维数为d的低维向量空间；$[x]_+$返回x与0两者中的较大值；$\|\cdot\|$表示L_1或L_2范式；γ为间隔，即区分正例三元组$(h,r,t)\in\Delta$和负例三元组$(h',r',t')\in\Delta'$的一个非负值，γ的取值可以预先设定。

[0065] 由公式(3)可知，对于已知李某和张某存在“合著”关系的示例，其优化目标是正例(李某,合著,张某)的损失加上间隔γ大于负例(李某,合著,沈某)的损失。

[0066] b)计算多步关系路径的间隔。

[0067] 可以计算每条关系路径的最优间隔，也可以根据上述的多步关系路径重要度的得分仅计算其中一部分重要的关系路径的最优间隔。

[0068] 在本发明中，关系路径的最优间隔定义为：

[0069] $$M_{path}(p)=\min_{r,r'}|\|p-r'\|-\|p-r\||| \quad (4)$$

[0070] 其中，正例关系$r\in R_{h,t}$，$R_{h,t}$为知识图谱中正例三元组(h,r,t)的关系r的集合；负例关系$r'\in N_{h,t}$，$N_{h,t}$为知识图谱中负例三元组(h,r′,t)的关系r′的集合，(h,r′,t)是将(h,r,t)中的r替换为r′得到的；$\|\cdot\|$表示L_1或L_2范式。

[0071] 更具体地说，在理想的知识图谱低维向量表示中，对于特定的多步关系路径p，与其相关的正例关系r在低维向量空间中应尽可能相近，而与其相关的负例关系r′在低维向量空间中应尽可能相远。因此，关系路径p的向量表示使得所对应的正例关系$r\in R_{h,t}$聚簇在

一起，并与负例关系$r' \in N_{h,t}$之间具有一定的间隔。

[0072] 如图2所示(以二维图形示出)，特定路径的最优间隔$M_{path}(p)$等于两个同心超球面体的超半径模长的差，正例关系r(用空心圆表示)均位于内侧球体，负例关系r'(用空心正方形表示)均位于外部球体以外，最优间隔$M_{path}(p)$等于内外超球面体之间的距离间隔。因此，通过对每条多步关系路径或重要的多步关系路径设置适当的间隔能够将正例和负例尽可能的分隔开，从而可以提高低维向量学习的精确度。

[0073] c)多步关系路径和关系之间的损失。

[0074] $L_{p,r}=[||p-r||+M_{path}(p)-||p-r'||]_+$ (5)

[0075] 其中，$M_{path}(p)$为多步关系路径的最优间隔；p是由关系$r_1,r_2,\cdots,r_l$组成的关系路径$p=\{r_1,r_2,\cdots,r_l\}$的低维向量表示向量，$p\in R^d$，且$p=r_1+r_2+\cdots+r_l$，d为低维向量空间的维度；r'为负例关系$r' \in N_{h,t}$中的低维向量表示向量，$N_{h,t}$为知识图谱中负例三元组(h，r'，t)的集合；$||\cdot||$表示L_1或L_2范式。

[0076] 由公式(5)可知，通过定义$M_{path}(p)$，对于不同的知识图谱或不同实体对的关系路径，其间隔是可变的，即损失函数是可变的。因此，这种方法可以自适应地学习知识向量的低维表示。

[0077] 综上所述，在本发明的实施例中，定义的低维向量学习过程的损失函数L(E，z)相当于有两个约束：一是使得实体对和关系组成的三元组最接近真实情况；二是使得实体对的多步关系路径最接近真实情况。其中，实体对和关系的损失最小是指头实体向量加关系向量应该与尾实体向量接近。实体对和关系路径之间的损失的最小通过实体对之间的关系来衡量，如“-父亲-父亲-”这条路径的向量应该与“爷爷”的向量在低维空间中距离接近。

[0078] 在学习过程中，可以通过更新向量来优化损失函数。例如，可以采用梯度下降方法来进行更新，向量更新方式如下：

[0079] $\forall i \in \{0,1,2,\ldots \dim\}, p \in P_{h,t},$

[0080] $h_i=h_i+\mu*2*|t_i-h_i-r_i|$

[0081] $r_i=r_i+\mu*2*|t_i-h_i-r_i|$

[0082] $t_i=t_i-\mu*2*|t_i-h_i-r_i|$

[0083] $h'_i=h'_i-\mu*2*|t'_i-h'_i-r'_i|$

[0084] $r'_i=r'_i-\mu*2*|t'_i-h'_i-r'_i|$

[0085] $t'_i=t'_i+\mu*2*|t'_i-h'_i-r'_i|$

[0086] $$p_i = p_i + \mu * R(p|h,t) * |r_i - \sum_{pp \in P_{h,t}} pp_i|$$

[0087] $$r_i = r_i - \mu * R(p|h,t) * |r_i - \sum_{pp \in P_{h,t}} pp_i|$$

[0088] $$p_i = p_i - \mu * R(p|h,t) * |r'_i - \sum_{pp \in P_{h,t}} pp_i|$$

[0089] $$r'_i = r'_i + \mu * R(p|h,t) * |r'_i - \sum_{pp \in P_{h,t}} pp_i|$$

[0090] 其中，dim是向量空间的维度，h_i代表h的第i维向量。μ为学习率，一般在{0.1，0.01，.0.001}中选择。

[0091] 通过迭代执行步骤S150，直到收敛，即可获得训练后的实体、关系、多步关系路径的向量。

[0092] 6)步骤S160，基于打分函数进行关系推理。

[0093] 本步骤的目的在于根据上述学习获得的实体、关系、关系路径的低维向量表示，来进行关系推理。具体包括：

[0094] a)对于待推断的目标实体和关系，选取所有符合候选实体类型的实体作为候选实体集合。

[0095] b)、根据以下打分函数和学习获得的实体、关系的低维向量表示，计算候选实体和目标实体、目标关系组成的三元组的分值。

[0096] f(h,r,t)＝||h+r-t|| (6)

[0097] 其中，粗体h,r,t表示h,r,t在低维向量空间的向量表示，||·||表示L_1或L_2范式。

[0098] c)、根据计算的分值对候选实体排序，获得有序的候选实体列表。例如，候选实体列表中的第一个是最可能与目标实体产生目标关系的实体。

[0099] 例如，待推断的目标关系r＝出生地点，待推断的目标头实体h＝李某，候选实体集合可能是{北京，杭州，青岛，海南…}，计算分值后获得的有序的实体列表{杭州，北京}。

[0100] 为了进一步说明利用本发明进行关系推理的效果，发明人在Freebase的数据集FB15K上进行了验证。Freebase数据集是现有知识库的典型代表，是一个开源且不断更新的开放的知识库。发明人采用实体预测技术和本发明提供的方法，采用正实体的平均序值(Mean Rank)作为评价指标来进行验证。实验参数如下：数据集FB15K中，存在1345种关系和14951个实体。对FB15K数据集，学习过程使用的学习率λ＝0.001，向量可选的维度d＝50，批处理大小B＝4800，参数γ＝1，路径长度为2，选用L_1范式衡量相似度。

[0101] 经过实验，根据本发明的方法在“Filter”条件下平均序值为85，在“Raw”条件下平均序值为24。(其中，“Filter”表示在对候选实体进行排序之前，正三元组已经被过滤掉(例如，李某有两个孩子A和B，对实体对(李某，孩子，B)进行预测，给定“李某”和关系“孩子”，候选集把正例A去掉，使得候选集合只有B一个正确实体，称为filter；相反，候选集不做处理，使得候选集合中有A、B两个正确实体，排序时可能A排在B前面，使得mean rank变差，称为raw)。采用现有技术中流行的TransR方法得到的平均序值为226(Filter)，78(Raw)；采用另一种现有的HOLE方法得到的平均序值为259(Filter)，116(Raw)。因此，而采用本发明提供的方法与TransR技术相比，序值显著降低了50～110；与采用HOLE技术相比，序值显著降低了90～180。

[0102] 综上所述，根据本发明的方法，首先模型化目标实体对之间的关系和多步关系路径的相互联系，将其用低维向量表示，再定义自适应的最优的间隔值，可以提高模型学习的精确度，从而能够提高了知识图谱的关系推断的准确性。

[0103] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也

不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。本发明的范围由所附权利要求来限定。

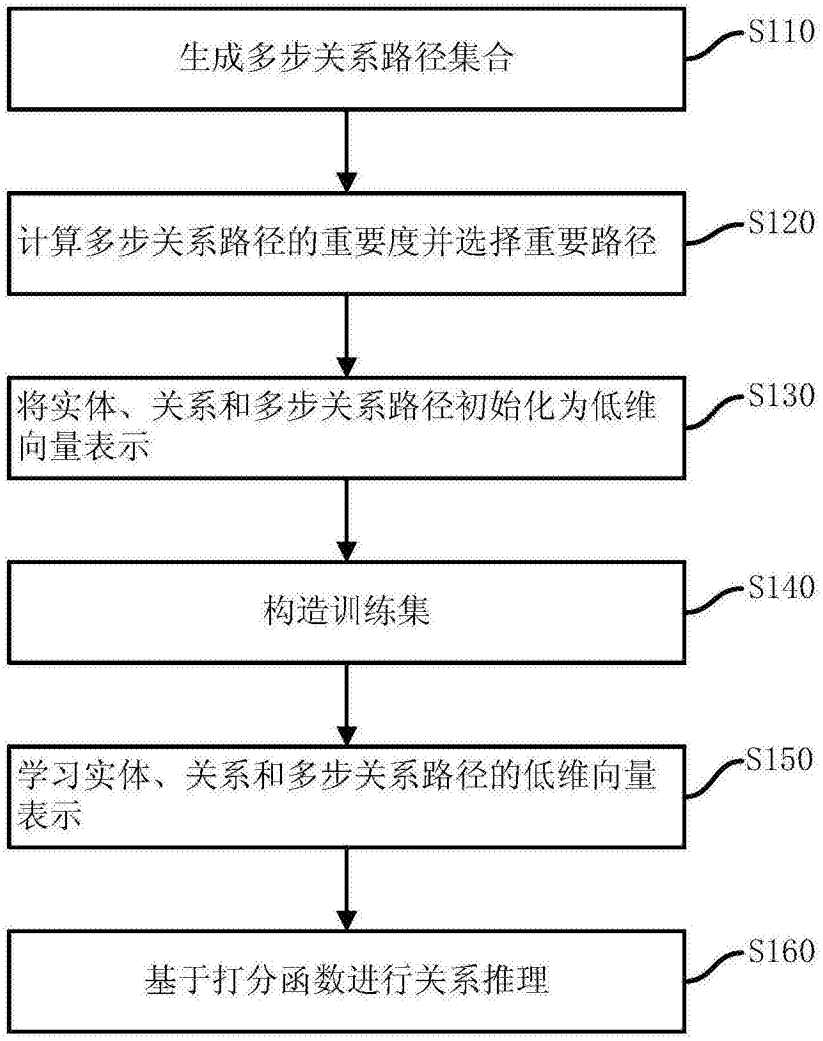

生成多步关系路径集合
S110
计算多步关系路径的重要度并选择重要路径
S120
将实体、关系和多步关系路径初始化为低维向量表示
S130
构造训练集
S140
学习实体、关系和多步关系路径的低维向量表示
S150
基于打分函数进行关系推理
S160

图1

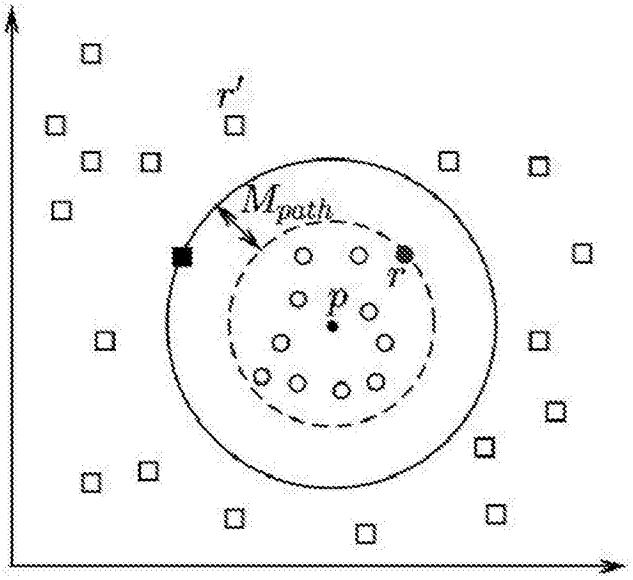

r'
M_{path}
r
p

图2