(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 108694201 A
(43)申请公布日 2018.10.23

(21)申请号 201710230135.0

(22)申请日 2017.04.10

(71)申请人 华为软件技术有限公司
地址 210012 江苏省南京市雨花台区软件大道101号华为南京基地

(72)发明人 贾岩涛 李曼玲 刘诗凯 邓拯宇

(74)专利代理机构 北京中博世达专利商标代理有限公司 11274
代理人 申健

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

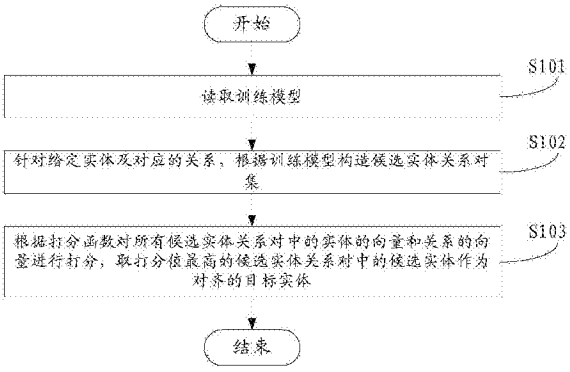
权利要求书3页 说明书10页 附图8页

(54)发明名称

一种实体对齐方法和装置

(57)摘要

本申请公开了一种实体对齐方法和装置，涉及大数据领域，用于提高基于向量空间表示的实体对齐效率。实体对齐方法包括：读取训练模型，其中，训练模型中包括实体的向量和关系的向量；针对给定实体及对应的关系，根据训练模型构造候选实体关系对集合；根据打分函数对所有候选实体关系对中的实体的向量和关系的向量进行打分，取打分值最高的候选实体关系对中的候选实体作为对齐的目标实体，其中，打分函数中包括给定实体的向量与候选实体的向量之间的属性相似度，当属性相似度值越高时打分函数打分值越高。本申请实施例应用于实体对齐。



CN 108694201 A

1.一种实体对齐方法，其特征在于，包括：

读取训练模型，其中，所述训练模型中包括实体的向量和关系的向量；

针对给定实体及对应的关系，根据所述训练模型构造候选实体关系对集合，其中，所述实体关系对集合中包括至少一个候选实体关系对，每个候选实体关系对包括给定实体、关系和候选实体，并且所述候选实体与所述给定实体的类型相同；

根据打分函数对所有候选实体关系对中的实体的向量和关系的向量进行打分，取打分值最高的候选实体关系对中的候选实体作为对齐的目标实体，其中，所述打分函数中包括所述给定实体的向量与候选实体的向量之间的属性相似度，当所述属性相似度值越高时所述打分函数打分值越高。

2.根据权利要求1所述的方法，其特征在于，当所述给定实体为尾实体t，对应的关系为r，候选实体为头实体h'时，

所述打分函数为$f_{predict}(h',r,t)=(1+w\times Dist(h',t))\|h'+r-t\|_{l_1}$，其中，$\|h'+r-t\|_{l_1}$表示h'和t的向量相似度，Dist(h',t)表示h'和t的属性相似度，w表示惩罚力度，取值范围为0到1，

其中，$Dist(h',t)=|t_t-h'_t|+EditDist(t_{attribute},h'_{attribute})$

其中，t_t表示t的时间，h'_t表示h'的时间，$t_{attribute}$表示t的属性，$h'_{attribute}$表示h'的属性，$EditDist(t_{attribute},h'_{attribute})$表示属性之间的编辑距离。

3.根据权利要求1所述的方法，其特征在于，在所述读取训练模型之前，所述方法还包括：

根据知识图谱的至少一个实体关系对(h,r,t)得到正例实体关系对集合Δ、负例实体关系对集合Δ'、与头实体h按照关系r构成的正例集$P_r=\{t|(h,r,t)\in\Delta\}$以及与头实体h按照关系r构成的负例集$N_r=\{t|(h,r,t)\notin\Delta,(h,r'',t)\in\Delta,\exists r''\in R\}$，其中，R表示关系集合，所述实体关系对(h,r,t)包括头实体h、关系r和尾实体t，所述正例实体关系对集合Δ表示所述知识图谱中存在的实体关系对(h,r,t)的集合，所述负例实体关系对集合Δ'表示所述知识图谱中不存在的实体关系对(h',r',t')的集合；

根据给定维度，初始化所述知识图谱的实体关系对(h,r,t)中的头实体向量、关系向量和尾实体向量，其中，每个头实体h对应一个头实体向量，每个关系r对应一个关系向量，每个尾实体t对应一个尾实体向量；

针对特定实体h及对应关系r，根据所述正例集P_r以及负例集N_r，计算所述特定实体h的实体间隔M_h；

根据所述正例实体关系对集合Δ、所述负例实体关系对集合Δ'和所述实体间隔M_h计算损失函数；

对实体关系对的头实体向量、关系向量和尾实体向量迭代进行更新，当所述损失函数满足预设条件时，更新得到的头实体向量、关系向量和尾实体向量作为所述训练模型。

4.根据权利要求3所述的方法，其特征在于，所述针对特定实体h，根据所述正例集P_r以及负例集N_r，计算所述特定实体的实体间隔M_h，包括：

针对特定实体h及其对应的关系r，选择$\forall t\in P_r$和$\forall t''\in N_r$，计算实体间隔$M_h=\min_{t,t''}\delta(||h-t''||-||h-t||)$，其中，$\delta(x)=\begin{cases}x & x\geq 0\\ -x & x<0\end{cases}$，$||\cdot||$表示$L_1$或$L_2$范式，$\min_{t,t''}$表示从所有根

据t或t”计算的结果中取最小值。

5.根据权利要求3所述的方法，其特征在于，所述损失函数为：

$$L=\sum_{(h,r,t)\in\Delta}\left[\sum_{(h',r',t')\in\Delta'}[\|h+r-t\|+M_h-\|h'+r'-t'\|]_+\right]$$

其中，M_h表示与头实体h对应的实体间隔，$[x]_+$返回x与0两者中的较大值，$||\cdot||$表示L_1或L_2范式。

6.根据权利要求3所述的方法，其特征在于，所述对所有实体关系对的头实体向量、关系向量和尾实体向量迭代进行更新，包括：

采用梯度下降法进行更新：$\forall i\in\{0,1,2,...,\dim\}$，其中，dim是向量空间的维度，$h_i$表示头实体h向量的第i维向量，μ为学习率，

$h_i=h_i-\mu*2*|t_i-h_i-r_i|$，

$r_i=r_i-\mu*2*|t_i-h_i-r_i|$，

$t_i=t_i+\mu*2*|t_i-h_i-r_i|$，

$h'_i=h'_i-\mu*2*|t'_i-h'_i-r'_i|$，

$r'_i=r'_i-\mu*2*|t'_i-h'_i-r'_i|$，

$t'_i=t'_i-\mu*2*|t'_i-h'_i-r'_i|$。

7.一种实体对齐装置，其特征在于，包括：

读取单元，用于读取训练模型，其中，所述训练模型中包括实体的向量和关系的向量；

构造单元，用于针对给定实体及对应的关系，根据所述训练模型构造候选实体关系对集合，其中，所述实体关系对集合中包括至少一个候选实体关系对，每个候选实体关系对包括给定实体、关系和候选实体，并且所述候选实体与所述给定实体的类型相同；

打分单元，用于根据打分函数对所有候选实体关系对中的实体的向量和关系的向量进行打分，取打分值最高的候选实体关系对中的候选实体作为对齐的目标实体，其中，所述打分函数中包括所述给定实体的向量与候选实体的向量之间的属性相似度，当所述属性相似度值越高时所述打分函数打分值越高。

8.根据权利要求7所述的装置，其特征在于，当所述给定实体为尾实体t，对应的关系为r，候选实体为头实体h’时，

所述打分函数为$f_{predict}(h',r,t)=(1+w\times Dist(h',t))\|h'+r-t\|_{l_i}$，其中，$\|h'+r-t\|_{l_i}$表示h’和t的向量相似度，Dist(h’,t)表示h’和t的属性相似度，w表示惩罚力度，取值范围为0到1，

其中，$Dist(h',t)=|t_t-h'_t|+EditDist(t_{attribute},h'_{attribute})$

其中，t_t表示t的时间，h'_t表示h’的时间，$t_{attribute}$表示t的属性，$h'_{attribute}$表示h’的属性，$EditDist(t_{attribute},h'_{attribute})$表示属性之间的编辑距离。

9.根据权利要求7所述的装置，其特征在于，所述装置还包括：

获取单元，用于在所述读取单元读取训练模型之前，根据知识图谱的至少一个实体关系对(h,r,t)得到正例实体关系对集合Δ、负例实体关系对集合Δ’、与头实体h按照关系r构成的正例集$P_r=\{t|(h,r,t)\in\Delta\}$以及与头实体h按照关系r构成的负例集$N_r=\{t|(h,r,t)\notin\Delta,(h,r'',t)\in\Delta,\exists r''\in R\}$，其中，R表示关系集合，所述实体关系对(h,r,t)包括头实体h、关系r和尾实体t，所述正例实体关系对集合Δ表示所述知识图谱中存在的实体

关系对(h,r,t)的集合,所述负例实体关系对集合Δ'表示所述知识图谱中不存在的实体关系对(h',r',t')的集合;

初始化单元,用于根据给定维度,初始化所述知识图谱的实体关系对(h,r,t)中的头实体向量、关系向量和尾实体向量,其中,每个头实体h对应一个头实体向量,每个关系r对应一个关系向量,每个尾实体t对应一个尾实体向量;

计算单元,用于针对特定实体h及对应关系r,根据所述正例集P_r以及负例集N_r,计算所述特定实体h的实体间隔M_h;

所述计算单元,还用于根据所述正例实体关系对集合Δ、所述负例实体关系对集合Δ'和所述实体间隔M_h计算损失函数;

所述获取单元,还用于对实体关系对的头实体向量、关系向量和尾实体向量迭代进行更新,当所述损失函数满足预设条件时,更新得到的头实体向量、关系向量和尾实体向量作为所述训练模型。

10.根据权利要求9所述的装置,其特征在于,所述计算单元具体用于:

针对特定实体h及其对应的关系r,选择$\forall t \in P_r$和$\forall t'' \in N_r$,计算实体间隔$M_h=\min_{t,t''}\delta(||h-t''||-||h-t||)$,其中,$\delta(x)=\begin{cases}x & x\geq 0\\ -x & x<0\end{cases}$,$||\cdot||$表示$L_1$或$L_2$范式,$\min_{t,t''}$表示从所有根据t或t"计算的结果中取最小值。

11.根据权利要求9所述的装置,其特征在于,所述损失函数为:

$$L=\sum_{(h,r,t)\in\Delta}\left[\sum_{(h',r',t')\in\Delta'}[\|h+r-t\|+M_h-\|h'+r'-t'\|]_+\right]$$

其中,M_h表示与头实体h对应的实体间隔,$[x]_+$返回x与0两者中的较大值,$||\cdot||$表示L_1或L_2范式。

12.根据权利要求9所述的装置,其特征在于,所述获取单元具体用于:

采用梯度下降法进行更新:$\forall i \in \{0,1,2,...,dim\}$,其中,dim是向量空间的维度,$h_i$表示头实体h向量的第i维向量,μ为学习率,

$h_i=h_i-\mu*2*|t_i-h_i-r_i|$,

$r_i=r_i-\mu*2*|t_i-h_i-r_i|$,

$t_i=t_i+\mu*2*|t_i-h_i-r_i|$,

$h'_i=h'_i-\mu*2*|t'_i-h'_i-r'_i|$,

$r'_i=r'_i-\mu*2*|t'_i-h'_i-r'_i|$,

$t'_i=t'_i-\mu*2*|t'_i-h'_i-r'_i|$。

一种实体对齐方法和装置

技术领域

[0001] 本申请涉及大数据领域，尤其涉及一种实体对齐方法和装置。

背景技术

[0002] 网络大数据时代的到来，使得网络上的数据呈爆炸式的增长。这些数据包含大量有价值的实体相关的信息，这里的实体指的是具体的某个现实社会中的对象，例如张艺谋、十面埋伏、巩俐等。根据其来源的不同，可以分为三类：垂直服务网站的实体数据、在线百科中的实体页面数据、开放新闻网页中实体相关的数据。不同的数据来源对同一个实体的名称表述可能不同。例如，《X战警：天启》这部电影，在不同的视频网站描述名称不同，例如《变种特攻：天启灭世战》、《X战警：启示录》等。这就需要对不同数据来源的视频信息进行对齐，即确定是否描述的是同一个实体。因此衍生出了实体对齐技术。

[0003] 目前比较有效的实体对齐技术是基于向量空间表示的实体对齐技术，即将实体的知识图谱表示成向量空间中的向量，通过各实体在向量空间中的位置等信息，预测各实体间的对齐关系。

[0004] 具体地，首先，通过优化一个基于间隔的损失函数，将知识图谱的点(实体)和边(实体间关系)表示成向量空间中的向量；然后，针对给定实体生成候选实体集合；最后，根据实体关系对打分函数对候选实体进行打分，取分数最高的作为对齐的实体，实现实体对齐的工作。常用的向量空间表示方法是TransE方法等。

[0005] 现有技术的打分函数将所有实体关系对统一进行打分，实际上打分效率并不高。比如，对上映时间差距较大的两个视频实体进行打分；将视频实体与人物类型实体进行打分等。

发明内容

[0006] 本申请的实施例提供一种实体对齐方法和装置，用于提高基于向量空间表示的实体对齐效率。

[0007] 为达到上述目的，本申请的实施例采用如下技术方案：

[0008] 第一方面，提供了一种实体对齐方法，包括：读取训练模型，其中，训练模型中包括实体的向量和关系的向量；针对给定实体及对应的关系，根据训练模型构造候选实体关系对集合，其中，实体关系对集合中包括至少一个候选实体关系对，每个候选实体关系对包括给定实体、关系和候选实体，并且候选实体与给定实体的类型相同；根据打分函数对所有候选实体关系对中的实体的向量和关系的向量进行打分，取打分值最高的候选实体关系对中的候选实体作为对齐的目标实体，其中，打分函数中包括给定实体的向量与候选实体的向量之间的属性相似度，当属性相似度值越高时打分函数打分值越高。本申请实施例提供的实体对齐方法通过在打分函数中加入属性相似度，并且当属性相似度值越高时打分函数打分值越高，无须经过人工对属性相似度进行阈值调整以及复杂的特征选取，提高基于向量空间表示的实体对齐效率。

[0009] 在一种可能的设计中，当给定实体为尾实体t，对应的关系为r，候选实体为头实体h'时，打分函数为$f_{predict}(h',r,t)=(1+w\times Dist(h',t))\|h'+r-t\|_{l_1}$，其中，$\|h'+r-t\|_{l_1}$表示h'和t的向量相似度，Dist(h',t)表示h'和t的属性相似度，w表示惩罚力度，取值范围为0到1，其中，Dist(h',t)＝|t_t-h'$_t$|+EditDist($t_{attribute}$,h'$_{attribute}$)，其中，t_t表示t的时间，h'$_t$表示h'的时间，$t_{attribute}$表示t的属性，h'$_{attribute}$表示h'的属性，EditDist($t_{attribute}$,h'$_{attribute}$)表示属性之间的编辑距离。该设计提供了一种打分函数的具体实现方式。

[0010] 在一种可能的设计中，在读取训练模型之前，方法还包括：根据知识图谱的至少一个实体关系对(h,r,t)得到正例实体关系对集合Δ、负例实体关系对集合Δ'、与头实体h按照关系r构成的正例集P_r＝{t|(h,r,t)∈Δ}以及与头实体h按照关系r构成的负例集$N_r=\{t|(h,r,t)\notin\Delta,(h,r'',t)\in\Delta,\exists r''\in R\}$，其中，R表示关系集合，实体关系对(h,r,t)包括头实体h、关系r和尾实体t，正例实体关系对集合Δ表示知识图谱中存在的实体关系对(h,r,t)的集合，负例实体关系对集合Δ'表示知识图谱中不存在的实体关系对(h',r',t')的集合；根据给定维度，初始化知识图谱的实体关系对(h,r,t)中的头实体向量、关系向量和尾实体向量，其中，每个头实体h对应一个头实体向量，每个关系r对应一个关系向量，每个尾实体t对应一个尾实体向量；针对特定实体h及对应关系r，根据正例集P_r以及负例集N_r，计算特定实体h的实体间隔M_h；根据正例实体关系对集合Δ、负例实体关系对集合Δ'和实体间隔M_h计算损失函数；对实体关系对的头实体向量、关系向量和尾实体向量迭代进行更新，当损失函数满足预设条件时，更新得到的头实体向量、关系向量和尾实体向量作为训练模型。该设计提供了一种得到训练模型的具体实现方式。

[0011] 在一种可能的设计中，针对特定实体h，根据正例集P_r以及负例集N_r，计算特定实体的实体间隔M_h，包括：针对特定实体h及其对应的关系r，选择$\forall t\in P_r$和$\forall t''\in N_r$，计算实体间隔M_h＝$\min_{t,t''}\delta(||h-t''||-||h-t||)$，其中，$\delta(x)=\begin{cases}x & x\geq 0\\ -x & x<0\end{cases}$，||·||表示$L_1$或$L_2$范式，$\min_{t,t''}$表示从所有根据t或t"计算的结果中取最小值。该设计提供了一种实体间隔M_h的具体实现方式。

[0012] 在一种可能的设计中，损失函数为：$L=\sum_{(h,r,t)\in\Delta}\left[\sum_{(h',r',t')\in\Delta'}[\|h+r-t\|+M_h-\|h'+r'-t'\|]_+\right]$

其中，M_h表示与头实体h对应的实体间隔，[x]$_+$返回x与0两者中的较大值，||·||表示L_1或L_2范式。该设计提供了一种损失函数的具体实现方式。

[0013] 在一种可能的设计中，对所有实体关系对的头实体向量、关系向量和尾实体向量迭代进行更新，包括：采用梯度下降法进行更新：$\forall i\in\{0,1,2,...,\dim\}$，其中，dim是向量空间的维度，$h_i$表示头实体h向量的第i维向量，μ为学习率。$h_i=h_i-\mu*2*|t_i-h_i-r_i|$，$r_i=r_i-\mu*2*|t_i-h_i-r_i|$，$t_i=t_i+\mu*2*|t_i-h_i-r_i|$，$h'_i=h'_i-\mu*2*|t'_i-h'_i-r'_i|$，$r'_i=r'_i-\mu*2*|t'_i-h'_i-r'_i|$，$t'_i=t'_i-\mu*2*|t'_i-h'_i-r'_i|$。该设计提供了一种对所有实体关系对的头实体向量、关系向量和尾实体向量迭代进行更新的具体实现方式。

[0014] 第二方面，本申请实施例提供了一种实体对齐装置，包括：读取单元，用于读取训练模型，其中，训练模型中包括实体的向量和关系的向量；构造单元，用于针对给定实体及对应的关系，根据训练模型构造候选实体关系对集合，其中，实体关系对集合中包括至少一

个候选实体关系对，每个候选实体关系对包括给定实体、关系和候选实体，并且候选实体与给定实体的类型相同；打分单元，用于根据打分函数对所有候选实体关系对中的实体的向量和关系的向量进行打分，取打分值最高的候选实体关系对中的候选实体作为对齐的目标实体，其中，打分函数中包括给定实体的向量与候选实体的向量之间的属性相似度，当属性相似度值越高时打分函数打分值越高。基于同一发明构思，由于该装置解决问题的原理以及有益效果可以参见上述第一方面和第一方面的各可能的方法实施方式以及所带来的有益效果，因此该装置的实施可以参见上述第一方面和第一方面的各可能的方法的实施方式，重复之处不再赘述。

[0015] 第三方面，本申请实施例提供一种实体对齐装置，包括：处理器、存储器、总线和通信接口；该存储器用于存储计算机执行指令，该处理器与该存储器通过该总线连接，当该设备运行时，该处理器执行该存储器存储的该计算机执行指令，以使该设备执行上述第一方面中任意一项的方法；基于同一发明构思，处理器调用存储在存储器中的指令以实现上述第一方面的方法设计中的方案，由于该设备解决问题的实施方式以及有益效果可以参见上述第一方面和第一方面的各可能的方法的实施方式以及有益效果，因此该设备的实施可以参见上述方法的实施，重复之处不再赘述。

[0016] 第四方面，本申请实施例提供了一种计算机存储介质，包括指令，当其在计算机上运行时，使得计算机执行如第一方面的实体对齐方法。

[0017] 第五方面，本申请实施例提供了一种包含指令的计算机程序产品，当其在计算机上运行时，使得该计算机执行如第一方面的实体对齐方法。

[0018] 另外，第三方面至第五方面中任一种设计方式所带来的技术效果可参见第一方面中不同设计方式所带来的技术效果，此处不再赘述。

附图说明

[0019] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍。

[0020] 图1为本申请的实施例提供的实体对齐装置的硬件结构示意图；

[0021] 图2为现有技术中TransE方法的示意图；

[0022] 图3为本申请的实施例提供的一种实体对齐方法的流程示意图；

[0023] 图4为本申请的实施例提供的另一种实体对齐方法的流程示意图；

[0024] 图5为本申请的实施例提供的实体集内容的示例示意图；

[0025] 图6为本申请的实施例提供的关系集内容的示例示意图；

[0026] 图7为本申请的实施例提供的一种训练集内容的示例示意图；

[0027] 图8为本申请的实施例提供的另一种训练集内容的示例示意图；

[0028] 图9为本申请的实施例提供的实体间隔的示意图；

[0029] 图10为本申请的实施例提供的一种实体向量表示的示意图；

[0030] 图11为本申请的实施例提供的另一种实体向量表示的示意图；

[0031] 图12为本申请的实施例提供的实体对齐结果的示意图；

[0032] 图13为本申请的实施例提供的一种实体对齐装置的结构示意图；

[0033] 图14为本申请的实施例提供的另一种实体对齐装置的结构示意图；

[0034] 图15为本申请的实施例提供的又一种实体对齐装置的结构示意图。

具体实施方式

[0035] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述。

[0036] 参照图1中所示，为本申请实施例提供的一种实体对齐装置的硬件结构示意图，该实体对齐装置100包括至少一个处理器101，通信总线102，存储器103以及至少一个通信接口104。

[0037] 处理器101可以是一个通用中央处理器(central processing unit，CPU)，微处理器，特定应用集成电路(application-specific integrated circuit，ASIC)，或一个或多个用于控制本申请方案程序执行的集成电路。

[0038] 通信总线102可包括一通路，在上述组件之间传送信息。

[0039] 通信接口104，使用任何收发器一类的装置，用于与其他设备或通信网络通信，如以太网，无线接入网(radio access network，RAN)，无线局域网(wireless local area networks，WLAN)等。

[0040] 存储器103可以是只读存储器(read-only memory，ROM)或可存储静态信息和指令的其他类型的静态存储设备，随机存取存储器(random access memory，RAM)或者可存储信息和指令的其他类型的动态存储设备，也可以是电可擦可编程只读存储器(electrically erasable programmable read-only memory，EEPROM)、只读光盘(compact disc read-only memory，CD-ROM)或其他光盘存储、光碟存储(包括压缩光碟、激光碟、光碟、数字通用光碟、蓝光光碟等)、磁盘存储介质或者其他磁存储设备、或者能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质，但不限于此。存储器可以是独立存在，通过总线与处理器相连接。存储器也可以和处理器集成在一起。

[0041] 其中，存储器103用于存储执行本申请方案的应用程序代码，并由处理器101来控制执行。处理器101用于执行存储器103中存储的应用程序代码，从而实现本申请实施例中所述的方法。

[0042] 在具体实现中，作为一种实施例，处理器101可以包括一个或多个CPU，例如图中的CPU0和CPU1。

[0043] 在具体实现中，作为一种实施例，实体对齐装置100可以包括多个处理器，例如图中的处理器101和处理器108。这些处理器中的每一个可以是一个单核(single-CPU)处理器，也可以是一个多核(multi-CPU)处理器。这里的处理器可以指一个或多个设备、电路、和/或用于处理数据(例如计算机程序指令)的处理核。

[0044] 在具体实现中，作为一种实施例，实体对齐装置100还可以包括输出设备105和输入设备106。输出设备105和处理器101通信，可以以多种方式来显示信息。例如，输出设备105可以是液晶显示器(liquid crystal display，LCD)，发光二级管(light emitting diode，LED)显示设备，阴极射线管(cathode ray tube，CRT)显示设备，或投影仪(projector)等。输入设备106和处理器101通信，可以以多种方式接受用户的输入。例如，输入设备106可以是鼠标、键盘、触摸屏设备或传感设备等。

[0045] 上述的实体对齐装置100可以是一个通用设备或者是一个专用设备。在具体实现

中，实体对齐装置100可以是台式机、便携式电脑、网络服务器、掌上电脑(personal digital assistant，PDA)、移动手机、平板电脑、无线用户设备、嵌入式设备或有图中类似结构的设备。本申请实施例不限定实体对齐装置100的类型。

[0046] 本申请实施例所述的实体对齐是指：对于第一数据源的某个实体，在其他数据源(包括第一数据源或者另一个数据源)中找到与其相同的实体，称为实体对齐。本申请实施例所述的实体通常指具体的某个个体，如张艺谋、十面埋伏、巩俐等。本申请实施例所述的知识图谱是由实体及实体间关系构成的网络，其中，网络中的点是实体，网络中的边是实体间的关联关系。

[0047] 参照图2中所示，TransE方法认为知识图谱由实体关系对(h,r,t)组成，h表示头实体，t表示尾实体，r表示头实体与尾实体之间的关系实体，例如小明的爸爸是大明，表示成实体关系对即为(小明，爸爸，大明)。

[0048] 在建立训练模型时，首先随机初始化头实体h的向量、尾实体t的向量和关系实体r的向量。然后通过迭代运算对上述向量进行优化，使得最终生成的向量应满足：在向量空间中，头实体h的向量加上关系实体r的向量与尾实体t的向量非常相近，即$h+r\approx t$。具体的，通过损失函数$L=f(h,r,t)+M-f(h',r',t')$取值最小使得正例实体关系对尽量满足上述假设，负例实体关系对尽量不满足上述假设来实现。其中，损失函数是基于间隔的函数；(h,r,t)是正例实体关系对，表示知识图谱中存在的实体关系对；(h',r',t')是负例实体关系对，表示知识图谱中不存在的实体关系对；M是非负实数的间隔，在TransE方法中是一个常量，例如$M=4$。

[0049] 在针对特定实体进行实体对齐时，根据打分函数$f(h,r,t)=||h+r-t||$取值最高从训练模型中选出对齐的候选实体。

[0050] 首先，现有技术在对基于间隔的损失函数进行优化时，损失函数中的间隔对知识图谱中所有实体关系对均相同，使得学习效果受到制约。不同实体和关系是具有结构差异的，相关关系稀疏的实体对应的实体关系对的间隔应该较大，而相关关系稠密的实体对应的实体关系对的间隔应该较小；同时在优化过程中间隔应该随着优化效果变化，迭代轮数少的时候，向量学习不充分，间隔应该较小，迭代轮数较多的时候，向量学习充分，间隔应该较大，使得进行更充分的学习。因此统一设定一个间隔，使得向量的学习效果受限，不能很好地反映实体向量之间的关系，实体对齐的效果受到影响。

[0051] 其次，损失函数中间隔的取值选择较为复杂。损失函数中的间隔的取值是在预先给定的候选值集合中选取，通过在验证数据集上验证来选取最优间隔值，确定最优损失函数。至于损失函数各间隔的取值为何在事先给定的有限集合中选择，却没有一种有力的解释。显而易见的是，间隔取值为非负。因此，遍历整个非负集合的方式去寻找最优间隔值是一件不可能的事情，而在事先给定的有限集合中遍历选取最优间隔值，使得调整参数的工作量大，实体对齐耗费时间长。

[0052] 再次，现有技术没有考虑实体对齐任务的特殊性，在对给定实体的候选实体集打分时，仅利用知识图谱的实体关系对局部结构特性进行学习打分，选择候选集中分数最高的实体作为对齐实体，但没有考虑实体对齐任务要求给定实体和对齐的实体之间的类型约束和内容的高度相似性，使得现有技术的实体对齐效果受到制约。

[0053] 本申请实施例提供的实体对齐方法和装置，对上述TransE方法进行改进，一方面，

调整损失函数中的间隔M，使得生成的训练模型更好拟合知识图谱；另一方面，针对打分函数引入惩罚机制，使得打分效率更高。另外，本申请实施例虽然示例性的以视频主体为例进行说明，但是本领域技术人员可以理解，本申请实施例还可以应用于其他主体对齐场景，例如音乐领域的歌曲名之间的对齐，旅游领域的相关地点的对齐，等等。

[0054] 本申请实施例提供了一种实体对齐方法，参照图3中所示，包括：

[0055] S101、读取训练模型。

[0056] 其中，训练模型中包括实体的向量和关系的向量，具体的，训练模型中包括头实体h的向量、关系r的向量和尾实体t的向量，头实体h、关系r和尾实体t构成实体关系对(h,r,t)。

[0057] S102、针对给定实体及对应的关系，根据训练模型构造候选实体关系对集合。

[0058] 其中，实体关系对集合中包括至少一个候选实体关系对，每个候选实体关系对包括给定实体、关系和候选实体，并且候选实体与给定实体的类型相同。

[0059] 选取候选实体时进行类型约束，仅选取与给定实体类型相同的实体。

[0060] 示例性的可以将给定实体作为尾实体，候选实体作为头实体；或者，候选实体作为尾实体，给定实体作为头实体。

[0061] S103、根据打分函数对所有候选实体关系对中的实体的向量和关系的向量进行打分，取打分值最高的候选实体关系对中的候选实体作为对齐的目标实体。

[0062] 打分函数中包括给定实体向量与候选实体向量的属性相似度，当属性相似度值越高时打分函数打分值越高。

[0063] 当给定实体作为尾实体t，候选实体作为头实体h'时，打分函数为$f_{predict}(h',r,t)=(1+w\times Dist(h',t))\|h'+r-t\|_{l_1}$。

[0064] 其中，$\|h'+r-t\|_{l_1}$表示h′和t的向量相似度，Dist(h',t)表示h′和t的属性相似度，w表示惩罚力度，取值范围为0到1，由数据集属性的可信度决定。

[0065] 其中，$Dist(h',t)=|t_t-h'_t|+EditDist(t_{attribute},h'_{attribute})$。

[0066] 其中，t_t表示t的时间，h'_t表示h'的时间，如果没有时间属性则这两个值为0。$t_{attribute}$表示t的属性，$h'_{attribute}$表示h'的属性，$EditDist(t_{attribute},h'_{attribute})$表示属性之间的编辑距离，例如属性可以为名称、数值。示例性的，当应用于音乐领域的歌曲名之间的对齐时，可以调整属性约束，例如对于歌曲而言，可以基于对歌曲更为重要的时间和歌手属性对实体关系对进行惩罚。

[0067] 通过一种类型约束的、带有惩罚项的打分函数，替换了原来TransE的打分函数。对于时间等数值类型的属性，进行统一的处理，无须经过人工对属性相似度进行阈值调整以及复杂的特征选取。

[0068] 本申请实施例提供的实体对齐方法通过在打分函数中加入属性相似度，并且当属性相似度值越高时打分函数打分值越高，无须经过人工对属性相似度进行阈值调整以及复杂的特征选取，提高基于向量空间表示的实体对齐效率。

[0069] 参照图4中所示，在步骤S101之前，所述方法还包括：

[0070] S201、根据知识图谱的至少一个实体关系对(h,r,t)得到正例实体关系对集合Δ、负例实体关系对集合Δ'、与头实体h按照关系r构成的正例集$P_r=\{t|(h,r,t)\in\Delta\}$以及与

头实体h按照关系r构成的负例集$N_r=\{t|(h,r,t)\notin\Delta,(h,r'',t)\in\Delta,\exists r''\in R\}$。

[0071] 其中，R表示关系集合，以视频实体为例，该关系集合包括导演、演员、上映时间、类型、对等、名称等关系；实体关系对(h,r,t)包括头实体h、关系r和尾实体t；正例实体关系对集合Δ表示知识图谱中存在的实体关系对(h,r,t)的集合，即(h,r,t)∈Δ；负例实体关系对集合Δ'表示知识图谱中不存在的实体关系对(h',r',t')的集合，即(h',r',t')∈Δ'，其通过对每个实体关系对(h,r,t)随机替换其头实体h、尾实体t、关系r而得到。

[0072] 知识图谱的具体形式包括文本文件。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。在本申请实施例中，示例性的，以实体为视频主体为例，假设存在视频来源1和视频来源2，需要将视频来源2中视频主体与视频来源1中的视频主体对齐。实体包括视频、人物、时间等，关系包括视频网络中的对等、导演等关系。以音乐领域的歌曲名之间的对齐为例，知识图谱中实体包括歌曲、人物、公司、时间、唱片等，关系包括歌曲、歌手、唱片、发行公司、发行时间、简介等。

[0073] 每个知识图谱包括3份数据文件：实体集(例如entity2id.txt)、关系集(例如relation2id.txt)、训练集(例如train.txt)。各数据文件均由行组成，每个数据文件中行的格式说明如下：

[0074] 1)实体集：每行数据包括两列，第一列数据为实体，第二列为标识(ID)编号，列与列之间用Tab隔开。示例性的如图5中所示

[0075] 2)关系集：每行数据包括两列，第一列数据为关系，第二列为标识(ID)编号，列与列之间用Tab隔开。示例性的如图6中所示

[0076] 3)训练集：包括反映主体属性的实体关系对和反映主体对齐信息的实体关系对。

[0077] 其中，反映主体属性实体关系对中，每行数据包括三列，第一列数据为头实体，第二列为关系，第三列为尾实体，列与列之间用Tab隔开。示例性的如图7中所示。

[0078] 反映主体对齐信息的实体关系对中，每行数据包括三列，第一列数据为视频数据源1的键值key(头实体)，第二列为关系(例如“对等”指的是头实体尾实体对应同一视频)，第三列数据为视频数据源2的键值key(尾实体)，列与列之间用Tab隔开。示例性的如图8中所示。

[0079] S202、根据给定维度，初始化知识图谱的实体关系对(h,r,t)中的头实体向量、关系向量和尾实体向量，其中，每个头实体h对应一个头实体向量，每个关系r对应一个关系向量，每个尾实体t对应一个尾实体向量。

[0080] 具体的，可以采用平均分布初始化、伯努利分布初始化等，将知识图谱的实体关系对初始化为给定维度的头实体的向量、尾实体的向量和关系的向量。

[0081] S203、针对特定实体h及对应关系r，根据正例集P_r以及负例集N_r，计算特定实体h的实体间隔M_h；

[0082] 针对特定实体h及其对应的关系r，选择$\forall t\in P_r$和$\forall t''\in N_r$，计算实体间隔$M_h=\min_{t,t''}\delta(||h-t''||-||h-t||)$，其中，$\delta(x)=\begin{cases}x & x\geq 0\\ -x & x<0,\end{cases}$ $||\cdot||$表示L_1或L_2范式，$\min_{t,t''}$表示从所有根据t或t”计算的结果中取最小值。

[0083] 通过这个公式，可以实现根据知识图谱中特定实体的结构特性，计算特定实体的间隔。事实上，对于特定的实体(包括头实体或尾实体)，当实体的向量表示使得所对应正例

的尾(或头)实体聚簇在一起,那些与其为负例的实体之间具有一定的间隔,此时取得最优值。正例实体与头实体(或尾实体)具有相同的关系,负例实体与头实体(或尾实体)之间存在不同的关系。从这个角度来讲,最优的实体间隔M等于两个同心超球面体的超半径模长的差,具体如图9所示(这里以二维图形来表示)。对于头实体h,与其具有正例关系的实体(空心圆)均位于内侧球体,与其具有负例关系的实体(空心矩形)均在外部球体以外。最优的实体间隔M等于内外超球面体之间的半径距离,该实体间隔把属于N_r的元素与头实体h分开,同时能够使得属于P的元素距离头实体h更近。

[0084] 该方法可以根据知识图谱的结构特点,自动的选择优化损失函数中的间隔,克服了传统向量表示学习中无法快速获得最优间隔的问题,同时通过在自适应间隔中融合各个实体的结构特性,根据向量的即时训练状态决定间隔,提升了模型对知识图谱的拟合程度,使得实体对齐的效果得到了提升。

[0085] S204、根据正例实体关系对集合Δ、负例实体关系对集合Δ'和实体间隔M_h计算损失函数。

[0086] 根据已初始化的实体、关系,建模表示知识图谱。基于假设实体关系对在向量空间中满足h+r≈t,损失函数使得正例实体关系对尽量满足此假设,负例实体关系对尽量不满足此假设,其中,损失函数如下:

[0087] $$L=\sum_{(h,r,t)\in\Delta}\left[\sum_{(h',r',t')\in\Delta'}[\|h+r-t\|+M_h-\|h'+r'-t'\|]_+\right]$$

[0088] 其中,M_h表示与头实体h对应的实体间隔,$[x]_+$返回x与0两者中的较大值,$\|\cdot\|$表示L_1或L_2范式。

[0089] S205、对实体关系对的头实体向量、关系向量和尾实体向量迭代进行更新,当损失函数满足预设条件时,更新得到的头实体向量、关系向量和尾实体向量作为训练模型。

[0090] 损失函数满足预设条件包括:损失函数计算达到最大迭代次数,或者,损失函数的结果值在各次迭代中保持不变。

[0091] 具体的,可以采用梯度下降法来进行更新,更新方式如下:$\forall i\in\{0,1,2,...,\dim\}$,其中,dim是向量空间的维度,$h_i$表示头实体h向量的第i维向量,μ为学习率。

[0092] $h_i=h_i-\mu*2*|t_i-h_i-r_i|$,

[0093] $r_i=r_i-\mu*2*|t_i-h_i-r_i|$,

[0094] $t_i=t_i+\mu*2*|t_i-h_i-r_i|$,

[0095] $h'_i=h'_i-\mu*2*|t'_i-h'_i-r'_i|$,

[0096] $r'_i=r'_i-\mu*2*|t'_i-h'_i-r'_i|$,

[0097] $t'_i=t'_i-\mu*2*|t'_i-h'_i-r'_i|$。

[0098] 为验证本发明提供的实体对齐方法,以视频实体对齐任务为例,采用本发明提供的方法,在真实百度视频、豆瓣视频数据集上进行了实验,采用对齐的正确率(accuracy)作为评价指标,实验参数如下:

[0099] 数据集百度视频-豆瓣视频中,存在770个豆瓣视频,770个百度视频。包括6种关系(上映时间、导演、类型、演员、名称、对等),包括视频、人物等在内的8179个实体,训练实体关系对为28920个,其中,对等关系670个,测试对齐关系对为100个。

[0100] 学习过程使用的学习率μ＝0.001，向量的维度d＝60，参数w＝1，选用L_1范式衡量相似度。

[0101] 得到实体对齐的正确率为93.02％。具体如表1中所示：

[0102] 表1

[0103]

数据集	维度	迭代次数	正确率
豆瓣和百度视频	60	100	0.930233

[0104] 学习得到的向量表示为60维向量，例如，百度视频中“X战警(天启)”视频的向量表示参照图10中所示。豆瓣视频中的“X战警:天启”视频的向量表示参照图11中所示。给定豆瓣视频“X战警:天启”，根据打分函数计算出百度视频的分数排位参照图12中所示，从中可以看出X战警(天启)是对齐结果。

[0105] 本申请实施例提供一种实体对齐装置，用于执行上述通信系统间移动方法。本申请实施例可以根据上述方法示例对实体对齐装置进行功能模块的划分，例如，可以对应各个功能划分各个功能模块，也可以将两个或两个以上的功能集成在一个处理模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。需要说明的是，本申请实施例中对模块的划分是示意性的，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

[0106] 在采用对应各个功能划分各个功能模块的情况下，图13示出了上述实施例中所涉及的实体对齐装置的一种可能的结构示意图，实体对齐装置20包括：读取单元2011、构造单元2012、打分单元2013、获取单元2014、初始化单元2015、计算单元2016。读取单元2011用于支持实体对齐装置20执行图3中的过程S101；构造单元2012用于支持实体对齐装置20执行图3中的过程S102；打分单元2013用于支持实体对齐装置20执行图3中的过程S103；获取单元2014用于支持实体对齐装置20执行图4中的过程S201、S205；初始化单元2015用于支持实体对齐装置20执行图4中的过程S202；计算单元2016用于支持实体对齐装置20执行图4中的过程S203、S204。其中，上述方法实施例涉及的各步骤的所有相关内容均可以援引到对应功能模块的功能描述，在此不再赘述。

[0107] 在采用集成的单元的情况下，图14示出了上述实施例中所涉及的实体对齐装置的一种可能的结构示意图。实体对齐装置20包括：处理模块2022和通信模块2023。处理模块2022用于对实体对齐装置20的动作进行控制管理，例如，处理模块2022用于支持实体对齐装置20执行图3中的过程S101-S103、图4中的过程S20。通信模块2023用于支持实体对齐装置与其他实体的通信，例如与图1中示出的功能模块或网络实体之间的通信。实体对齐装置20还可以包括存储模块2021，用于存储实体对齐装置的程序代码和数据。

[0108] 其中，处理模块2022可以是处理器或控制器，例如可以是中央处理器(central processing unit,CPU)，通用处理器，数字信号处理器(digital signal processor,DSP)，专用集成电路(application-specific integrated circuit,ASIC)，现场可编程门阵列(field programmable gate array,FPGA)或者其他可编程逻辑器件、晶体管逻辑器件、硬件部件或者其任意组合。其可以实现或执行结合本申请公开内容所描述的各种示例性的逻辑方框，模块和电路。所述处理器也可以是实现计算功能的组合，例如包含一个或多个微处理器组合，DSP和微处理器的组合等等。通信模块2023可以是收发器、收发电路或通信接口

等。存储模块2021可以是存储器。

[0109] 当处理模块2022为处理器，通信模块2023为收发器，存储模块2021为存储器时，本申请实施例所涉及的实体对齐装置可以为如下所述的实体对齐装置。

[0110] 参照图15所示，该实体对齐装置20包括：处理器2032、收发器2033、存储器2031、总线2034。其中，收发器2033、处理器2032、存储器2031通过总线2034相互连接；总线2034可以是外设部件互连标准(peripheral component interconnect，PCI)总线或扩展工业标准结构(extended industry standard architecture，EISA)总线等。所述总线可以分为地址总线、数据总线、控制总线等。为便于表示，图中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

[0111] 以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

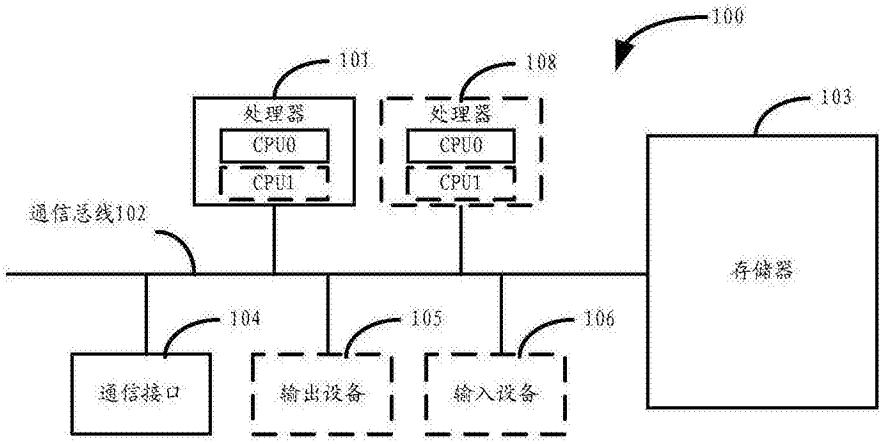

100
101
处理器
CPU0
CPU1
108
处理器
CPU0
CPU1
103
存储器
通信总线102
104
通信接口
105
输出设备
106
输入设备

图1

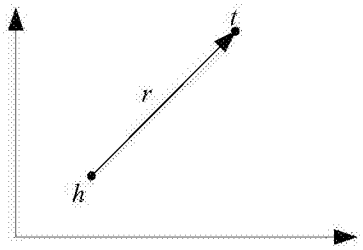

t
r
h

图2

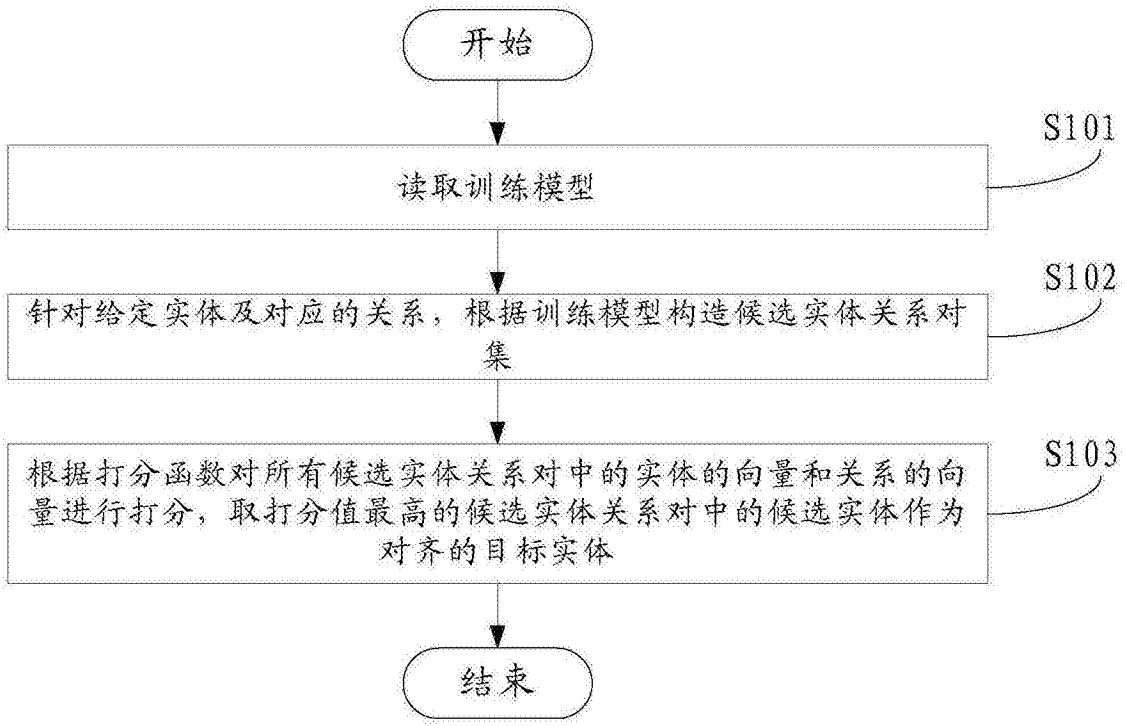

开始
S101
读取训练模型
S102
针对给定实体及对应的关系，根据训练模型构造候选实体关系对集
S103
根据打分函数对所有候选实体关系对中的实体的向量和关系的向量进行打分，取打分值最高的候选实体关系对中的候选实体作为对齐的目标实体
结束

图3

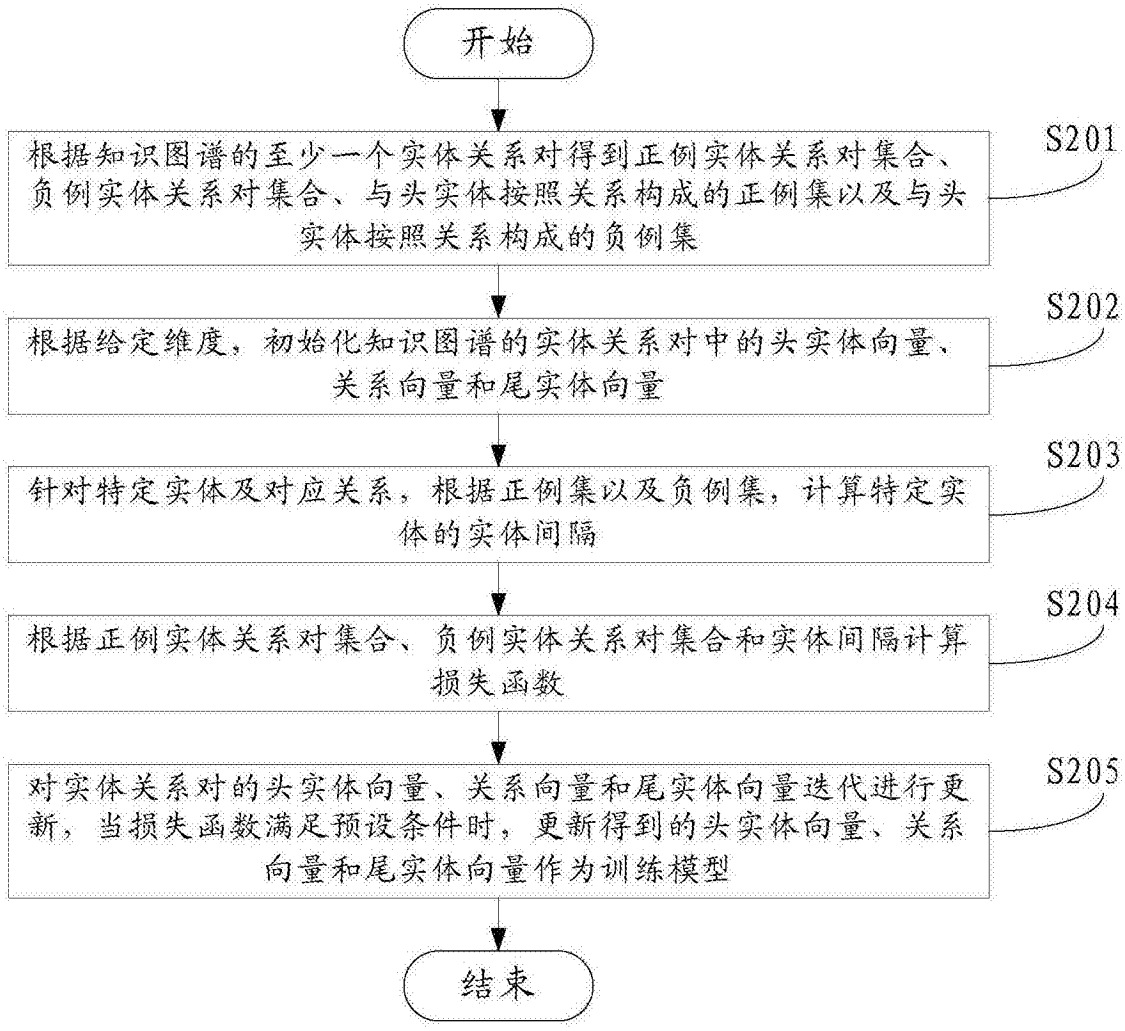

开始
根据知识图谱的至少一个实体关系对得到正例实体关系对集合、负例实体关系对集合、与头实体按照关系构成的正例集以及与头实体按照关系构成的负例集
S201
根据给定维度，初始化知识图谱的实体关系对中的头实体向量、关系向量和尾实体向量
S202
针对特定实体及对应关系，根据正例集以及负例集，计算特定实体的实体间隔
S203
根据正例实体关系对集合、负例实体关系对集合和实体间隔计算损失函数
S204
对实体关系对的头实体向量、关系向量和尾实体向量迭代进行更新，当损失函数满足预设条件时，更新得到的头实体向量、关系向量和尾实体向量作为训练模型
S205
结束

图4


张达明 0
大卫·圣.詹姆斯 1
樱花大战(活动写真) 剧场版 2
亚当·戈德堡 3
千娇百媚 4
刘谦益 5
徐广林 6
Amy Madigan 7
史蒂夫·麦奎因 8
0581000001303913 9
0561000000025102 10

图5

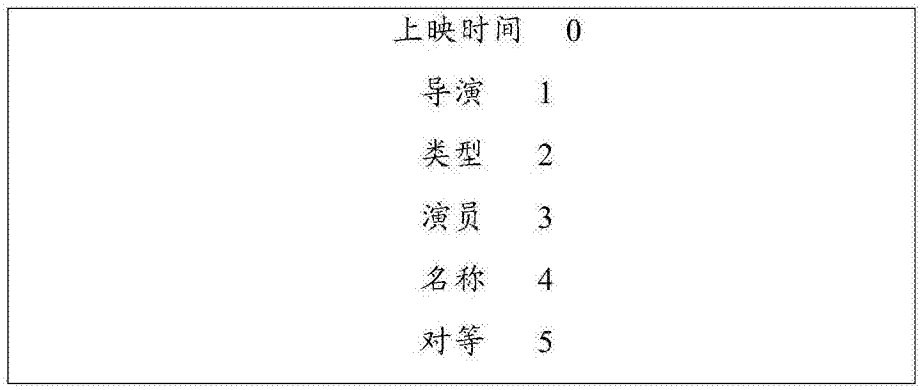

上映时间 0
导演 1
类型 2
演员 3
名称 4
对等 5

图6

0561000000014614 类型 电影

0561000000014614 名称 龙虎门

2006 类型 时间

0561000000014614 上映时间 2006

甄子丹 类型 人物

0561000000014614 演员 甄子丹

谢霆锋 类型 人物

0561000000014614 演员 谢霆锋

余文乐 类型 人物

0561000000014614 演员 余文乐

董洁 类型 人物

0561000000014614 演员 董洁

李小冉 类型 人物

0561000000014614 演员 李小冉

叶伟信 类型 人物

0561000000014614 导演 叶伟信

0561000000016068 类型 电影

0561000000016068 名称 大追捕

2012 类型 时间

0561000000016068 上映时间 2012

图7

0561000000015242	对等	0581000003546019
0562000000011024	对等	0581000005395263
0561000000022674	对等	0581000003649049
0561000000112922	对等	0581000001293708
0561000000021353	对等	0581000001550445
0561000000014141	对等	0581000001306249
0562000000012837	对等	0581000001478186
0562000000011409	对等	0581000002156667
0561000000021693	对等	0581000006893711
0562000000013001	对等	0581000006893711
0562000000014427	对等	0581000003014957
0562000000011190	对等	0581000003732699

图8

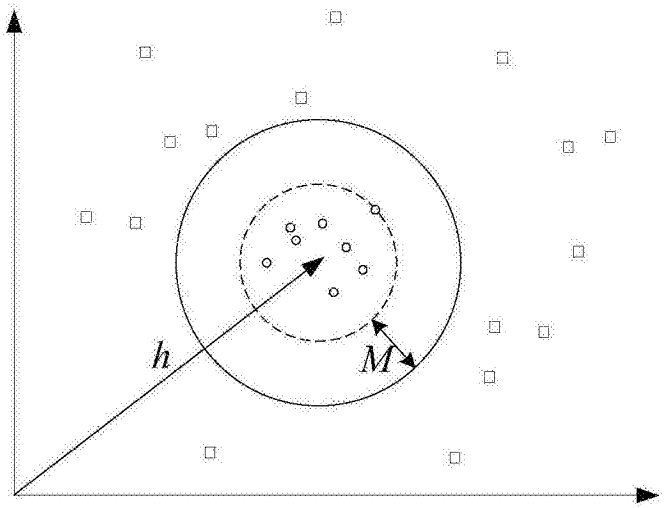


图9

0.165963 0.141895 -0.046573 -0.030492 -0.181973 -0.155722 -0.008115 0.139305 -0.122368 0.008246 0.178326 0.149461 -0.096322 -0.136397 0.126709 -0.160746 -0.107846 -0.152673 -0.169473 -0.151452 0.115710 -0.185473 0.169529 0.002271 -0.096940 0.082594 -0.157725 0.148325 -0.113182 -0.158990 0.013055 -0.063735 -0.093295 0.114467 -0.093022 0.147141 -0.065504 0.161225 0.081945 0.190091 -0.138557 0.183582 0.012448 0.164524 -0.119500 -0.143366 0.109497 -0.179581 -0.113213 -0.026855 0.183149 0.029990 -0.154405 0.037277 -0.133470 0.128393 0.096617 -0.114441 0.175726 0.135902

图10

0.080719 0.006505 -0.041561 0.184700 0.081253 -0.057214 -0.091525 0.181425 0.016547 0.120036 -0.010313 -0.121201 0.178466 0.046792 0.034523 0.152666 -0.030591 -0.130452 -0.022529 -0.164799 -0.147954 0.133757 -0.135171 0.028159 -0.028209 0.111333 -0.059675 0.035682 -0.029829 0.064682 -0.071957 -0.206229 0.158262 0.031891 -0.103002 0.185844 -0.045473 -0.102753 0.046084 0.119942 -0.055354 -0.102351 0.039817 0.072848 0.059110 -0.083732 0.106675 0.052241 0.039380 -0.169498 0.144741 0.009944 -0.090671 -0.104482 0.037474 -0.095683 0.028326 0.058688 -0.083203 -0.010289

图11

X战警（天启） -0.261229
-0.261229
巨人捕手杰克 -10.4117
江湖论剑实录 -14.4094
飞虎出征 -14.4905
逃出生天 -14.5627

图12

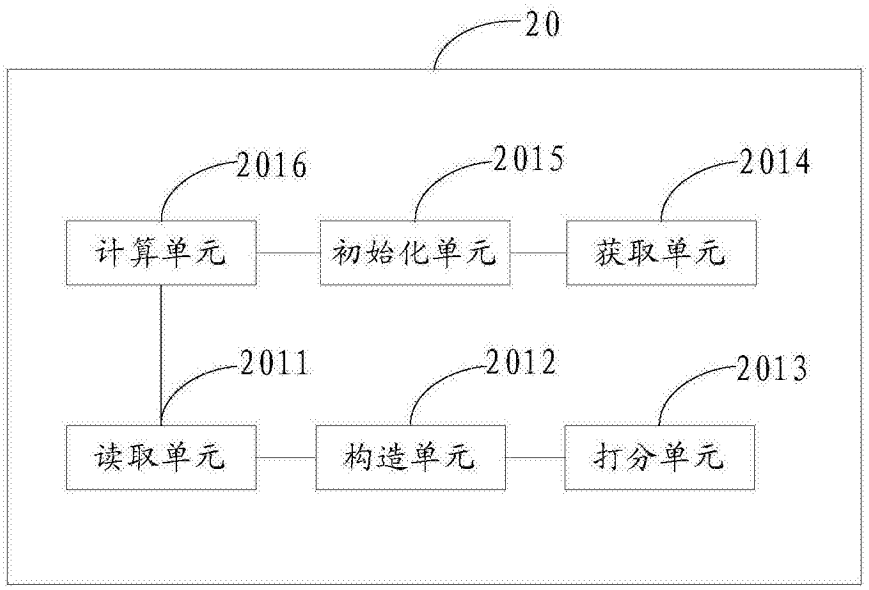

20
2016
2015
2014
计算单元
初始化单元
获取单元
2011
2012
2013
读取单元
构造单元
打分单元

图13

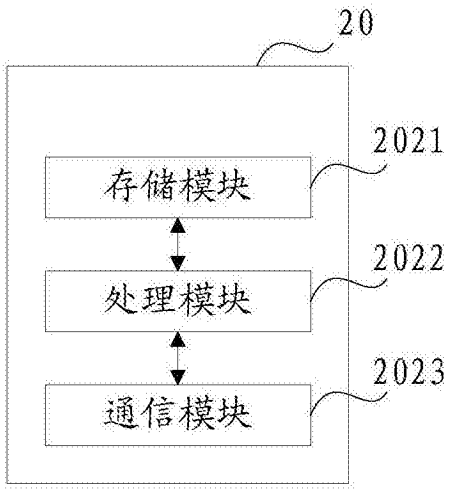

20
2021
存储模块
2022
处理模块
2023
通信模块

图14

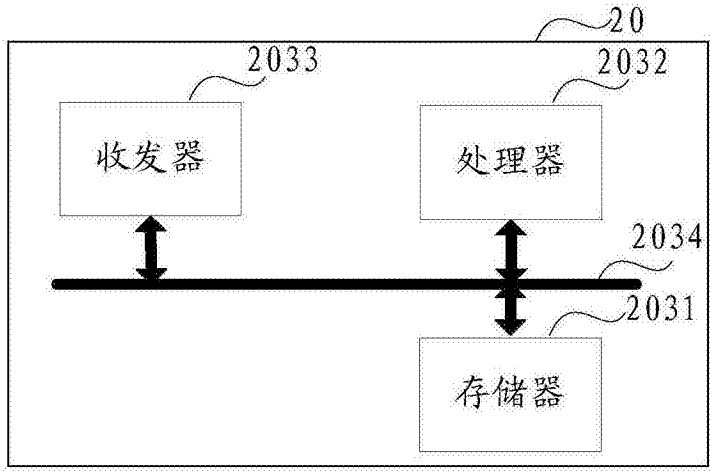


图15