(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 6 月 11 日 (11.06.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/114022 A1

(51) 国际专利分类号:
G06F 16/28 (2019.01)

(21) 国际申请号: PCT/CN2019/103487

(22) 国际申请日: 2019 年 8 月 30 日 (30.08.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811474699.X 2018年12月4日 (04.12.2018) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(72) 发明人: 吴壮伟 (WU, Zhuangwei); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市世联合知识产权代理有限公司 (SL INTELLECTUAL PROPERTY CO., LTD.); 中国广东省深圳市福田区泰然六路泰然科技园205栋二层203-206, Guangdong 518000 (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,

WO 2020/114022 A1

(54) **Title:** KNOWLEDGE BASE ALIGNMENT METHOD AND APPARATUS, COMPUTER DEVICE AND STORAGE MEDIUM

(54) **发明名称:** 一种知识库对齐方法、装置、计算机设备及存储介质

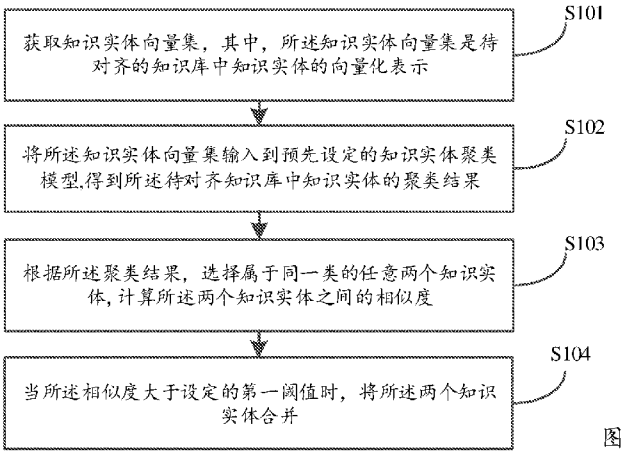


图 1

S101 Obtain a knowledge entity vector set, wherein the knowledge entity vector set is a vectorized representation of knowledge entities in a knowledge base to be aligned
S102 Input the knowledge entity vector set into a preset knowledge entity clustering model to obtain a clustering result of the knowledge entities in said knowledge base
S103 Select any two knowledge entities of the same type according to the clustering r esult, and calculate the similarity between the two knowledge entities
S104 If the similarity is greater than a set first threshold, combine the two knowledge entities

(57) **Abstract:** Disclosed in the embodiments of the present application are a knowledge base alignment method and apparatus, a computer device and a storage medium. The method comprises the following steps: obtaining a knowledge entity vector set, wherein the knowledge entity vector set is a vectorized representation of knowledge entities in a knowledge base to be aligned; inputting the knowledge entity vector set into a preset knowledge entity clustering model to obtain a clustering result of the knowledge entities in said knowledge base; selecting any two knowledge entities of the same type according to the clustering result, and calculating the similarity between the two knowledge entities; and if the similarity is greater than a set first threshold, combining the two knowledge entities.

[见续页]

SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(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)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

The comparison for the similarity between two knowledge entities is limited to the entities of the same type, so the calculation amount is greatly reduced; during clustering, the clustering is achieved by means of an artificial intelligence technology, so that the clustering result is more in line with the expectation; the similarity calculation integrates the attribute similarity and vector similarity of the entities, thus the similarity calculation is more reasonable, and redundant information can be more effectively found and removed.

(57) 摘要：本申请实施例公开了一种知识库对齐的方法、装置、计算机设备及存储介质，其中方法包括下述步骤：获取知识实体向量集，其中，所述知识实体向量集是待对齐的知识库中知识实体的向量化表示；将所述知识实体向量集输入到预先设定的知识实体聚类模型,得到所述待对齐知识库中知识实体的聚类结果；根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度；当所述相似度大于设定的第一阈值时，将所述两个知识实体合并。两个知识实体相似度的比较限于同一类实体中，大大减少了计算量，聚类时，通过人工智能技术实现，使聚类结果更符合预期，相似度的计算综合了实体的属性相似度和向量相似度，使相似度的计算更合理，可以更有效的发现和去除冗余信息。

一种知识库对齐方法、装置、计算机设备及存储介质

本申请以2018年12月4日提交的申请号为201811474699X，名称为“一种知识库对齐方法、装置、计算机设备及存储介质”的中国发明专利申请为基础，并要求其优先权。

【技术领域】

本申请涉及知识库处理技术领域，尤其涉及一种知识库对齐方法、装置、计算机设备及存储介质。

【背景技术】

随着互联网的发展，各个领域构建了越来越多的知识库，这些知识库也被广泛的应用于搜索服务、自动问答等互联网应用中。知识库对信息的共享和传播具有积极意义。然而，单个知识库的信息有限，在一些情况下不能满足用户的需求；且通常知识库是持续扩充的，占用的存储资源的规模也持续扩大，但持续扩充到知识库中的数据可能存在冗余，这种冗余造成存储资源的浪费，同时，也使搜索计算量增大，搜索结果信息重复，给用户带来不便。

知识库对齐(Knowledge Base Alignment)指对于不同来源的各个实体，找出属于现实中同一事物的实体。这里实体指客观存在并可相互区别的事物，包括具体的人、事、物、抽象的概念、关系。因此知识库对齐，即抽取实体信息，去除冗余，是构建高质量知识库的关键问题。

知识库对齐常用的方法是利用实体的属性信息判定不同源实体是否可进行对齐，由于不同实体数据属于用户原创内容（User Generated Content，UGC）类型，不同用户编辑的数据质量参差不齐，仅通过用户编辑的实体属性信息难以准确判定是否为同一实体。

【发明内容】

本申请实施例提供了一种知识库对齐方法、装置、计算机设备及存储介质。

一种知识库对齐方法，所述知识库对齐方法包括：

获取知识实体向量集，其中，所述知识实体向量集是待对齐的知识库中知识实体的向量化表示；

将所述知识实体向量集输入到预先设定的知识实体聚类模型,得到所述待对齐知识库中知识实体的聚类结果；

根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度；

当所述相似度大于设定的第一阈值时，将所述两个知识实体合并。

一种知识库对齐装置，所述知识库对齐装置包括：

获取模块，用于获取知识实体向量集，其中，所述知识实体向量集是待对齐的知识库中知识实体的向量化表示；

处理模块，用于将所述知识实体向量集输入到预先设定的知识实体聚类模型,得到所述待对齐知识库中知识实体的聚类结果；

计算模块，用于根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度；

执行模块，用于当所述相似度大于设定的第一阈值时，将所述两个知识实体合并。

一种计算机设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令，所述处理器执行所述计算机可读指令时实现上述任一种知识库对齐方法的步骤。

一种存储有计算机可读指令的可读存储介质，所述计算机可读指令被一种处理器执行时实现上述任一种知识库对齐方法的步骤。

本申请的一个或多个实施例的细节在下面的附图和描述中提出，本申请的其他特征和优点将从说明书、附图以及权利要求变得明显。

【附图说明】

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图1为本申请实施例一种知识库对齐方法基本流程示意图；

图2为本申请实施例基于IF-IDF算法对知识实体向量化的示意图；

图3为本申请实施例基于卷积神经网络的聚类模型训练流程示意图；

图4为本申请实施例知识实体相似度计算流程示意图；

图5为本申请实施例知识实体合并流程示意图；

图6为本申请实施例一种知识库对齐装置基本结构框图；

图7为本申请实施计算机设备基本结构框图。

【具体实施方式】

为了使本技术领域的人员更好地理解本申请方案，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。

在本申请的说明书和权利要求书及上述附图中的描述的一些流程中，包含了按照特定顺序出现的多个操作，但是应该清楚了解，这些操作可以不按照其在本文中出现的顺序来执行或并行执行，操作的序号如 101、102 等，仅仅是用于区分开各个不同的操作，序号本身不代表任何的执行顺序。另外，这些流程可以包括更多或更少的操作，并且这些操作可以按顺序执行或并行执行。需要说明的是，本文中的“第一”、“第二”等描述，是用于区分不同的消息、设备、模块等，不代表先后顺序，也不限定“第一”和“第二”是不同的类型。

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，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

实施例

本技术领域技术人员可以理解，这里所使用的“终端”、“终端设备”既包括无线信号接收器的设备，其仅具备无发射能力的无线信号接收器的设备，又包括接收和发射硬件的设备，其具有能够在双向通信链路上，执行双向通信的接收和发射硬件的设备。这种设备可以包括：蜂窝或其他通信设备，其具有单线路显示器或多线路显示器或没有多线路显示器的蜂窝或其他通信设备；PCS（Personal Communications Service，个人通信系统），其可以组合语音、数据处理、传真和/或数据通信能力；PDA（Personal Digital Assistant，个人数字助理），其可以包括射频接收器、寻呼机、互联网/内联网访问、网络浏览器、记事本、日历和/或 GPS（Global Positioning System，全球定位系统）接收器；常规膝上型

和/或掌上型计算机或其他设备，其具有和/或包括射频接收器的常规膝上型和/或掌上型计算机或其他设备。这里所使用的“终端”、“终端设备”可以是便携式、可运输、安装在交通工具（航空、海运和/或陆地）中的，或者适合于和/或配置为在本地运行，和/或以分布形式，运行在地球和/或空间的任何其他位置运行。这里所使用的“终端”、“终端设备”还可以是通信终端、上网终端、音乐/视频播放终端，例如可以是 PDA、MID（Mobile Internet Device，移动互联网设备）和/或具有音乐/视频播放功能的移动电话，也可以是智能电视、机顶盒等设备。

本实施方式中的终端即为上述的终端。

具体地，请参阅图 1，图 1 为本实施例一种知识库对齐方法的基本流程示意图。

如图 1 所示，一种知识库对齐方法，包括下述步骤：

S101、获取知识实体向量集，其中，所述知识实体向量集是待对齐的知识库中知识实体的向量化表示；

保存在知识库中的知识实体通常为文本或图片，在对知识实体进行对齐时，通常需要计算知识实体间的相似度，为了方便计算机处理和理解，需要用将知识实体转化为向量。例如文本的向量化表示通过向量空间模型也称为词袋模型(bag of words)实现，其中最简单的模式是基于词的独热编码（one-hot encoding），用每一个词作为维度 key，有单词对应的位置为 1，其他为 0，向量长度和词典大小相同。

S102、将所述知识实体向量集输入到预先设定的知识实体聚类模型,得到所述待对齐知识库中知识实体的聚类结果；

将表示知识实体的向量集输入到预先设定的知识实体聚类模型。其中知识实体的聚类模型采用基于密度的聚类算法，基于密度的聚类算法不需要事先确定簇类的数据，可以发现任意形状的簇类，能够识别出噪声点，对离群点有较好的鲁棒性，可以检测离群点。 DBSCAN 是该类方法中最典型的代表算法之一，其核心思想就是先发现密度较高的点，然后把相近的高密度点逐步都连成一片，进而生成各种簇。具体的算法实现：对每个数据点为圆心，以 eps 为半径画个圈（称为邻域 eps-neigbourhood），然后数有多少个点在这个圈内，这个数就是该点密度值。然后选取一个密度阈值 MinPts，如圈内点数小于 MinPts 的圆心点为低密度的点，而大于或等于 MinPts 的圆心点高密度的点（称为核心点 Core point）。如果有一个高密度的点在另一个高密度的点的圈内，我们就把这两个点

连接起来，这样我们可以把好多点不断地串联出来。之后，如果有低密度的点也在高密度的点的圈内，把它也连到最近的高密度点上，称之为边界点。这样所有能连到一起的点就成一了个簇，而不在任何高密度点的圈内的低密度点就是异常点。

在一些实施方式中，采用经过训练的卷积神经网络模型来实现聚类，通过对卷积神经网络进行训练学习人工对训练样本聚类的特征，使卷积神经网络模型可以按照预期对知识实体进行聚类。

S103、根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度；

通过步骤 S102，对知识库中的知识实体进行了聚类，再在同一类中，通过计算任意两个知识实体的相似度，来判断是否存在冗余的实体，这样缩小了知识实体比较的范围，减小了计算量，提高了判断是否存在冗余实体的效率。

两个知识实体的相似度通过计算表示两个知识实体的向量之间的相似度来得出。两个向量之间的相似度可以是余弦相似度。余弦相似度通过测量两个向量的夹角的余弦值来度量它们之间的相似性。0 度角的余弦值是 1，而其他任何角度的余弦值都不大于 1；并且其最小值是-1。从而两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。两个向量有相同的指向时，余弦相似度的值为 1；两个向量夹角为 90° 时，余弦相似度的值为 0；两个向量指向完全相反的方向时，余弦相似度的值为-1。这结果是与向量的长度无关的，仅仅与向量的指向方向相关。余弦相似度对于任何维度的向量空间都适用，且常用于高维正空间，所以适合用于文本文件的比较。

也可以通过计算向量之间的欧氏距离来衡量两个向量之间的相似度。为了避免尺度的影响，先对向量进行归一化，再按照以下公式求向量空间中两个点 X_1，X_2 之间的距离：

$$d = \sqrt{\sum_{i=1}^{N} (x_{1i} - x_{2i})^2}$$

其中 x_{1i}，x_{2i} 为 X_1，X_2 归一化后各维度的值。

S104、当所述相似度大于设定的第一阈值时，将所述两个知识实体合并。

预先设定一个阈值，这里称之为第一阈值，当两个知识实体的相似度大于设定的第一阈值时，认为两个知识实体部分内容重复，将两个知识实体合并为

一个实体。

如图 2 所示，在 S101 之前，还包括步骤：

S111、获取待对齐的知识库中的知识实体；

通过访问知识库所在服务器获取知识实体，知识实体可以属于同一知识所库，也可以来源于多个知识库。

S112、将所述知识实体基于 IF-IDF 算法向量化，得到所述知识实体向量集。

将知识实体向量化，除了前述的基于词袋模型向量化外，还可以在基于基于 IF-IDF 算法对知识实体向量化。TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF 实际上是：TF * IDF，TF(Term Frequency，词频)，IDF(Inverse Document Frequency，逆向文件频率)。TF 表示词条在文档 d 中出现的频率。使用 TF-IDF 对文本向量化，同样构建一个词典，用每个词的 TF-IDF 值作为该词的权重。

如图 3 所示，所述基于卷积神经网络的聚类模型的训练，包括下述步骤：

S121、获取标记有聚类判断信息的训练样本，所述训练样本的聚类判断信息为样本知识实体的类别；

本申请实施例中，卷积神经网络的训练目标是识别知识实体所属的类别，卷积神经网络模型通过训练学习样本中人工标注类别的特征，实现对知识实体聚类的功能。

S122、将所述训练样本输入卷积神经网络模型获取所述训练样本的模型聚类参照信息；

卷积神经网络模型由：卷积层、池化层、全连接和分类层组成。其中，卷积层被用于对知识实体向量局部进行感知，且卷积层通常以级联的方式进行连接，级联中位置越靠后的卷积层能够感知越全局化的信息。

全连接层在整个卷积神经网络中起到“分类器”的作用。如果说卷积层、池化层和激活函数层等操作是将原始数据映射到隐层特征空间的话，全连接层则起到将学到的“分布式特征表示”映射到样本标记空间的作用。全连接层连接在卷积层输出位置，能够感知知识实体向量的全局化特征。

将训练样本输入到卷积神经网络模型中，获取卷积神经网络模型输入聚类参照信息。

S123、通过损失函数比对所述训练样本内不同样本的模型聚类参照信息与所述聚类判断信息是否一致；

通过损失函数比对聚类参照信息和样本标注的聚类判断信息是否一致，本申请实施例中使用 softmax 交叉熵损失函数，具体为：

$$L_{crossEntropyLoss}=-\frac{1}{N}log\left(\frac{e^{h_{yi}}}{\sum_{j=1}^{c}e^{h_j}}\right)$$

假设共有 N 个训练样本，针对网络最后分层第 i 个样本的输入特征为 X_i，其对应的标记为 Y_i 是最终的分类结果，h=（h1,h2,...,hc）为网络的最终输出，即样本 i 的预测结果。其中 C 是最后所有分类的数量。

S124、当所述模型聚类参照信息与所述聚类判断信息不一致时，反复循环迭代的更新所述卷积神经网络模型中的权重，至所述模型聚类参照信息与所述聚类判断信息一致时结束。

在训练过程中，调整卷积神经网络模型中各节点的权重，使 Softmax 交叉熵损失函数尽可能收敛，也就是说继续调整权重，得到的损失函数的值不再缩小，反而增大时，认为卷积神经网络训练可以结束。各节点权重的调整采用梯度下降法，梯度下降法是一个最优化算法，用于机器学习和人工智能当中用来递归性地逼近最小偏差模型。

通过训练后的卷积神经网络模型对知识实体进行聚类，可以使聚类结果更接近用户的预期。

如图 4 所示，步骤 S103 还包括下述步骤：

S131、获取所述两个知识实体的属性，其中，所述知识实体的属性为描述对应知识实体的数据；

在一些情况下，虽然两个知识实体从内容来看相似度不高，但是两个知识实体都对应现实中的一个实体，也就是说，两个知识实体分别描述了现实中某个实体的两部分信息，为了使用的方便，也有必要将这两部分信息合在一起。所以，这里引入属性相似度。先获取知识实体的属性，属性是用来描述知识实体的数据，也可以称为标签。

S132、计算所述两个知识实体的属性相似度和向量相似度；

属性相似度，本申请实施例中采用编辑距离来衡量两个知识实体之间的相似度。编辑距离，是指利用字符操作，把字符串 A 转换成字符串 B 所需要的最少操作数。字符操作包括：删除一个字符、修改一个字符、插入一个字符。在

这里设置每次操作的代价为 1，属性相似度可以通过以下公式计算：

属性相似度=(1-编辑距离)/两个属性字符串的最大长度

向量相似度，即前述的衡量两个知识实体向量相似度的余弦相似度或欧氏距离。

S133、按照以下公式计算所述两个知识实体的属性相似度和向量相似度的加权和，得到所述两个知识实体之间的相似度，即：

$$S = aX + bY$$

其中，S 为所述两个知识实体之间的相似度，X 为所述属性相似度，Y 为所述向量相似度，a、b 分别为所述属性相似度和所述向量相似度的权重。

综合属性相似度和向量相似度，可以在内容相似度不高的情况下，发现描述同一现实实体的两个知识实体，并对描述同一现实实体的知识实体进行合并，方便用户的使用和知识库的维护。

步骤 S104 还包括下述步骤：

S141、当所述相似度大于设定的第二阈值时，其中，所述第二阈值大于所述第一阈值，从待对齐的知识库中删除所述两个知识实体中的任意一个。

当两个知识实体的相似度非常高，这里我们设定第二阈值，第二阈值大于前述的第一阈值，例如设定的第二阈值为 0.95，即认为两个知识实体基本相同，这时，从知识库中删除任意一个知识实体就是有效的去除冗余的方法。

如图 5 所示，步骤 S104 还包括下述步骤：

S151、将所述两个知识实体分割成若干个子实体；

当两个知识实体的相似度大于预设的第一阈值时，认为两个知识实体部分内容重复，为了将重复的内容剔出，可以先将两个知识实体按照一定的规则分割成若干个子实体，例如按照内容段落分割。

S152、选择所述若干个子实体中的任意两个子实体，计算所述两个子实体之间的相似度；

选择分割后的任意两个子实体，计算两个子实体间的相似度，即如前面所述，先将子实体向量化，然后计算表示子实体的向量之间的相似度，可以是余弦相似度、也可以是欧氏距离。

S153、当所述两个子实体之间的相似度大于预设的第三阈值时，删除所述两个子实体中的任意一个，其中，所述第三阈值大于所述第一阈值；

当两个子实体之间的相似度大于预设的阈值时，这里称为第三阈值，认为

两个子实体内容基本重复，删除其中任意一个。为避免删除过多的内容，第三阈值要求大于前述的第一阈值。

S154、重复步骤 S152 和步骤 S153，直到保留的子实体中任意两个子实体之间的相似度都小于或等于预设的第三阈值；

重复进行子实体之间相似度的比较，删除重合度高的子实体，使保留的子实体中任意两个子实体的相似度都小于或等于预设的第三阈值。

S155、将所述保留的子实体合并作为所述两个知识实体的对齐实体。

将保留的子实体合并作为之前待对齐的两个知识实体的对齐结果。

为解决上述技术问题本申请实施例还提供一种知识库对齐装置。具体请参阅图 6，图 6 为本实施例知识库对齐装置的基本结构框图。

如图 6 所示，一种知识库对齐装置，包括：获取模块 210、处理模块 220、计算模块 230 和执行模块 240。其中，获取模块 210，用于获取知识实体向量集，其中，所述知识实体向量集是待对齐的知识库中知识实体的向量化表示；处理模块 220，用于将所述知识实体向量集输入到预先设定的知识实体聚类模型,得到所述待对齐知识库中知识实体的聚类结果；计算模块 230，用于根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度；执行模块 240，用于当所述相似度大于设定的第一阈值时，将所述两个知识实体合并。

本申请实施例通过获取知识实体向量集，将所述知识实体向量集输入到预先设定的知识实体聚类模型,得到所述待对齐知识库中知识实体的聚类结果，根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度，当所述相似度大于设定的第一阈值时，将所述两个知识实体合并。两个知识实体相似度的比较限于同一类实体中，大大减少了计算量，其中，相似度的计算综合了实体的属性相似度和向量相似度，使相似度的计算更合理，可以更有效的发现和去除冗余信息。

在一些实施方式中，所述知识库对齐装置还包括：第一获取子模块和第一处理子模块。其中，第一获取子模块，用于获取待对齐的知识库中的知识实体；第一处理子模块，用于将所述知识实体基于 IF-IDF 算法向量化，得到所述知识实体向量集。

在一些实施方式中，所述知识库对齐装置中预先设定的知识实体聚类模型采用DBSCAN密度聚类算法。

在一些实施方式中，所述知识库对齐装置中预先设定的知识实体聚类模型

采用基于卷积神经网络的聚类模型。

在一些实施方式中，所述计算模块230包括：第二获取子模块、第一计算子模块和第二计算子模块。其中，第二获取子模块，用于获取所述两个知识实体的属性，其中，所述知识实体的属性为描述对应知识实体的数据；第一计算子模块，用于计算所述两个知识实体的属性相似度和向量相似度；第二计算子模块，用于按照以下公式计算所述两个知识实体的属性相似度和向量相似度的加权和，得到所述两个知识实体之间的相似度，即：

$$S = aX + bY$$

其中，S为所述两个知识实体之间的相似度，X为所述属性相似度，Y为所述向量相似度，a、b分别为所述属性相似度和所述向量相似度的权重。

在一些实施方式中，所述执行模块240包括：第一执行子模块，用于当所述相似度大于设定的第二阈值时，其中，所述第二阈值大于所述第一阈值，从待对齐的知识库中删除所述两个知识实体中的任意一个。

在一些实施方式中，所述执行模块240包括：第一分割子模块、第三计算子模块、第二执行子模块、第一循环子模块和第三执行子模块。其中，第一分割子模块，用于将所述两个知识实体分割成若干个子实体；第三计算子模块，用于选择所述若干个子实体中的任意两个子实体，计算所述两个子实体之间的相似度；第二执行子模块，用于当所述两个子实体之间的相似度大于预设的第三阈值时，删除所述两个子实体中的任意一个，其中，所述第三阈值大于所述第一阈值；第一循环子模块，用于使第三计算子模块和第二执行子模块重复运行，直到保留的子实体中任意两个子实体之间的相似度都小于或等于预设的第三阈值；第三执行子模块，用于将所述保留的子实体合并作为所述两个知识实体的对齐实体。

为解决上述技术问题，本申请实施例还提供计算机设备。具体请参阅图 7，图 7 为本实施例计算机设备基本结构框图。

如图 7 所示，计算机设备的内部结构示意图。如图 7 所示，该计算机设备包括通过系统总线连接的处理器、非易失性存储介质、存储器和网络接口。其中，该计算机设备的非易失性存储介质存储有操作系统、数据库和计算机可读指令，数据库中可存储有控件信息序列，该计算机可读指令被处理器执行时，可使得处理器实现一种知识库对齐的方法。该计算机设备的处理器用于提供计算和控制能力，支撑整个计算机设备的运行。该计算机设备的存储器中可存储

有计算机可读指令，该计算机可读指令被处理器执行时，可使得处理器执行一种知识库对齐的方法。该计算机设备的网络接口用于与终端连接通信。本领域技术人员可以理解，图 7 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

本实施方式中处理器用于执行图 6 中获取模块 210、处理模块 220、计算模块 230 和执行模块 240 的具体内容，存储器存储有执行上述模块所需的计算机可读指令和各类数据。网络接口用于向用户终端或服务器之间的数据传输。本实施方式中的存储器存储有知识库对齐方法中执行所有子模块所需的计算机可读指令及数据，服务器能够调用服务器的计算机可读指令及数据执行所有子模块的功能。

计算机设备通过获取知识实体向量集，将所述知识实体向量集输入到预先设定的知识实体聚类模型,得到所述待对齐知识库中知识实体的聚类结果，根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度，当所述相似度大于设定的第一阈值时，将所述两个知识实体合并。两个知识实体相似度的比较限于同一类实体中，大大减少了计算量，其中，相似度的计算综合了实体的属性相似度和向量相似度，使相似度的计算更合理，可以更有效的发现和去除冗余信息

本申请还提供一种存储有计算机可读指令的存储介质，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行上述任一实施例所述知识库对齐方法的步骤。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机可读指令来指令相关的硬件来完成，该计算机可读指令可存储于一计算机可读取存储介质中，该计算机可读指令在执行时，可包括如上述各方法的实施例的流程。其中，前述的存储介质可为磁碟、光盘、只读存储记忆体（Read-Only Memory，ROM）等非易失性存储介质，或随机存储记忆体（Random Access Memory，RAM）等。

应该理解的是，虽然附图的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，其可以以其他的顺序执行。而

且，附图的流程图中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，其执行顺序也不必然是依次进行，而是可以与其他步骤或者其他步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

以上所述仅是本申请的部分实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本申请的保护范围。

权 利 要 求 书

1、一种知识库对齐方法,其特征在于,包括下述步骤:

获取知识实体向量集，其中，所述知识实体向量集是待对齐的知识库中知识实体的向量化表示;

将所述知识实体向量集输入到预先设定的知识实体聚类模型,得到所述待对齐知识库中知识实体的聚类结果;

根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度;

当所述相似度大于设定的第一阈值时，将所述两个知识实体合并。

2、根据权利要求 1 所述的知识库对齐方法，其特征在于，在所述获取知识实体向量集的步骤之前，还包括下述步骤:

获取待对齐的知识库中的知识实体;

将所述知识实体基于 IF-IDF 算法向量化，得到所述知识实体向量集。

3、根据权利要求 1 所述的知识库对齐方法，其特征在于，所述预先设定的知识实体聚类模型采用 DBSCAN 密度聚类算法。

4、根据权利要求 1 所述的知识库对齐方法，其特征在于，所述预先设定的知识实体聚类模型采用基于卷积神经网络的聚类模型，所述基于卷积神经网络的聚类模型的训练包含下述步骤:

获取标记有聚类判断信息的训练样本，所述训练样本的聚类判断信息为样本知识实体的类别;

将所述训练样本输入卷积神经网络模型获取所述训练样本的模型聚类参照信息;

通过损失函数比对所述训练样本内不同样本的模型聚类参照信息与所述聚类判断信息是否一致;

当所述模型聚类参照信息与所述聚类判断信息不一致时，反复循环迭代的更新所述卷积神经网络模型中的权重，至所述模型聚类参照信息与所述聚类判断信息一致时结束。

5、根据权利要求 1 所述的知识库对齐方法，其特征在于，在所述根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度的步骤具体包括下述步骤:

获取所述两个知识实体的属性，其中，所述知识实体的属性为描述对应知识实体的数据；

计算所述两个知识实体的属性相似度和向量相似度；

按照以下公式计算所述两个知识实体的属性相似度和向量相似度的加权和，得到所述两个知识实体之间的相似度，即：

$$S = aX + bY$$

其中，S 为所述两个知识实体之间的相似度，X 为所述属性相似度，Y 为所述向量相似度，a、b 分别为所述属性相似度和所述向量相似度的权重。

6、根据权利要求 1 所述的知识库对齐方法，其特征在于，在所述当所述相似度大于设定的第一阈值时，将所述两个知识实体合并的步骤中，还包括下述步骤：

当所述相似度大于设定的第二阈值时，其中，所述第二阈值大于所述第一阈值，从待对齐的知识库中删除所述两个知识实体中的任意一个。

7、根据权利要求 1 所述的知识库对齐方法，其特征在于，在所述当所述相似度大于设定的第一阈值时，将所述两个知识实体合并的步骤中，还包括下述步骤：

a.将所述两个知识实体分割成若干个子实体；

b.选择所述若干个子实体中的任意两个子实体，计算所述两个子实体之间的相似度；

c.当所述两个子实体之间的相似度大于预设的第三阈值时，删除所述两个子实体中的任意一个，其中，所述第三阈值大于所述第一阈值；

d.重复步骤 b 和步骤 c，直到保留的子实体中任意两个子实体之间的相似度都小于或等于预设的第三阈值；

e.将所述保留的子实体合并作为所述两个知识实体的对齐实体。

8、一种知识库对齐装置，其特征在于，包括：

获取模块，用于获取知识实体向量集，其中，所述知识实体向量集是待对齐的知识库中知识实体的向量化表示；

处理模块，用于将所述知识实体向量集输入到预先设定的知识实体聚类模型，得到所述待对齐知识库中知识实体的聚类结果；

计算模块，用于根据所述聚类结果，选择属于同一类的任意两个知识实体，计算所述两个知识实体之间的相似度；

执行模块，用于当所述相似度大于设定的第一阈值时，将所述两个知识实体合并。

9、根据权利要求 8 所述的知识库对齐装置，其特征在于，还包括：

第一获取子模块，用于获取待对齐的知识库中的知识实体；

第一处理子模块，用于将所述知识实体基于 IF-IDF 算法向量化，得到所述知识实体向量集。

10、根据权利要求 8 所述的知识库对齐装置，其特征在于，所述计算模块包括：

第二获取子模块，用于获取所述两个知识实体的属性，其中，所述知识实体的属性为描述对应知识实体的数据；

第一计算子模块，用于计算所述两个知识实体的属性相似度和向量相似度；

第二计算子模块，用于按照以下公式计算所述两个知识实体的属性相似度和向量相似度的加权和，得到所述两个知识实体之间的相似度，即：

$$S = aX + bY$$

其中，S 为所述两个知识实体之间的相似度，X 为所述属性相似度，Y 为所述向量相似度，a、b 分别为所述属性相似度和所述向量相似度的权重。

11、一种计算机设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令，其特征在于，所述处理器执行所述计算机可读指令时实现如下知识库对齐方法的步骤：

获取知识实体向量集，其中，所述知识实体向量集是待对齐的知识库中知识实体的向量化表示；

将所述知识实体向量集输入到预先设定的知识实体聚类模型,得到所述待对齐知识库中知识实体的聚类结果；

根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度；

当所述相似度大于设定的第一阈值时，将所述两个知识实体合并。

12、根据权利要求 11 所述的计算机设备，其特征在于，在所述获取知识实体向量集的步骤之前，还包括下述步骤：

获取待对齐的知识库中的知识实体；

将所述知识实体基于 IF-IDF 算法向量化，得到所述知识实体向量集。

13、根据权利要求 11 所述的计算机设备，其特征在于，所述预先设定的知

识实体聚类模型采用 DBSCAN 密度聚类算法。

14、根据权利要求 11 所述的计算机设备，其特征在于，所述预先设定的知识实体聚类模型采用基于卷积神经网络的聚类模型，所述基于卷积神经网络的聚类模型的训练包含下述步骤：

获取标记有聚类判断信息的训练样本，所述训练样本的聚类判断信息为样本知识实体的类别；

将所述训练样本输入卷积神经网络模型获取所述训练样本的模型聚类参照信息；

通过损失函数比对所述训练样本内不同样本的模型聚类参照信息与所述聚类判断信息是否一致；

当所述模型聚类参照信息与所述聚类判断信息不一致时，反复循环迭代的更新所述卷积神经网络模型中的权重，至所述模型聚类参照信息与所述聚类判断信息一致时结束。

15、根据权利要求 11 所述的计算机设备，其特征在于，在所述根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度的步骤具体包括下述步骤：

获取所述两个知识实体的属性，其中，所述知识实体的属性为描述对应知识实体的数据；

计算所述两个知识实体的属性相似度和向量相似度；

按照以下公式计算所述两个知识实体的属性相似度和向量相似度的加权和，得到所述两个知识实体之间的相似度，即：

$$S = aX + bY$$

其中，S 为所述两个知识实体之间的相似度，X 为所述属性相似度，Y 为所述向量相似度，a、b 分别为所述属性相似度和所述向量相似度的权重。

16、一种存储有计算机可读指令的可读存储介质，其特征在于，所述计算机可读指令被一种处理器执行时，使得所述一种处理执行如下步骤：

获取知识实体向量集，其中，所述知识实体向量集是待对齐的知识库中知识实体的向量化表示；

将所述知识实体向量集输入到预先设定的知识实体聚类模型,得到所述待对齐知识库中知识实体的聚类结果；

根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知

识实体之间的相似度；

当所述相似度大于设定的第一阈值时，将所述两个知识实体合并。

17、根据权利要求 16 所述的可读存储介质，其特征在于，在所述获取知识实体向量集的步骤之前，还包括下述步骤：

获取待对齐的知识库中的知识实体；

将所述知识实体基于 IF-IDF 算法向量化，得到所述知识实体向量集。

18、根据权利要求 16 所述的可读存储介质，其特征在于，所述预先设定的知识实体聚类模型采用 DBSCAN 密度聚类算法。

19、根据权利要求 16 所述的可读存储介质，其特征在于，所述预先设定的知识实体聚类模型采用基于卷积神经网络的聚类模型，所述基于卷积神经网络的聚类模型的训练包含下述步骤：

获取标记有聚类判断信息的训练样本，所述训练样本的聚类判断信息为样本知识实体的类别；

将所述训练样本输入卷积神经网络模型获取所述训练样本的模型聚类参照信息；

通过损失函数比对所述训练样本内不同样本的模型聚类参照信息与所述聚类判断信息是否一致；

当所述模型聚类参照信息与所述聚类判断信息不一致时，反复循环迭代的更新所述卷积神经网络模型中的权重，至所述模型聚类参照信息与所述聚类判断信息一致时结束。

20、根据权利要求 16 所述的可读存储介质，其特征在于，在所述根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度的步骤具体包括下述步骤：

获取所述两个知识实体的属性，其中，所述知识实体的属性为描述对应知识实体的数据；

计算所述两个知识实体的属性相似度和向量相似度；

按照以下公式计算所述两个知识实体的属性相似度和向量相似度的加权和，得到所述两个知识实体之间的相似度，即：

$$S = aX + bY$$

其中，S 为所述两个知识实体之间的相似度，X 为所述属性相似度，Y 为所述向量相似度，a、b 分别为所述属性相似度和所述向量相似度的权重。

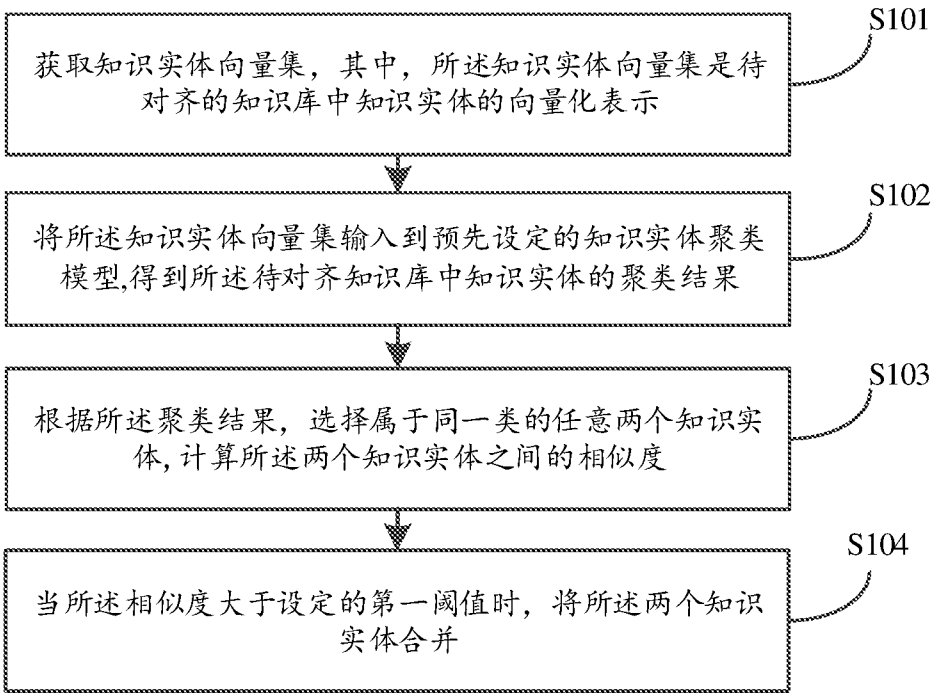

S101
获取知识实体向量集，其中，所述知识实体向量集是待对齐的知识库中知识实体的向量化表示
S102
将所述知识实体向量集输入到预先设定的知识实体聚类模型,得到所述待对齐知识库中知识实体的聚类结果
S103
根据所述聚类结果，选择属于同一类的任意两个知识实体,计算所述两个知识实体之间的相似度
S104
当所述相似度大于设定的第一阈值时，将所述两个知识实体合并

图 1

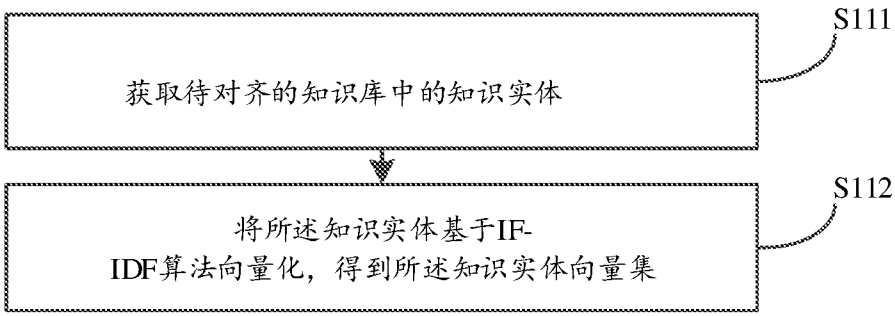

S111
获取待对齐的知识库中的知识实体
S112
将所述知识实体基于IF-IDF算法向量化，得到所述知识实体向量集

图 2

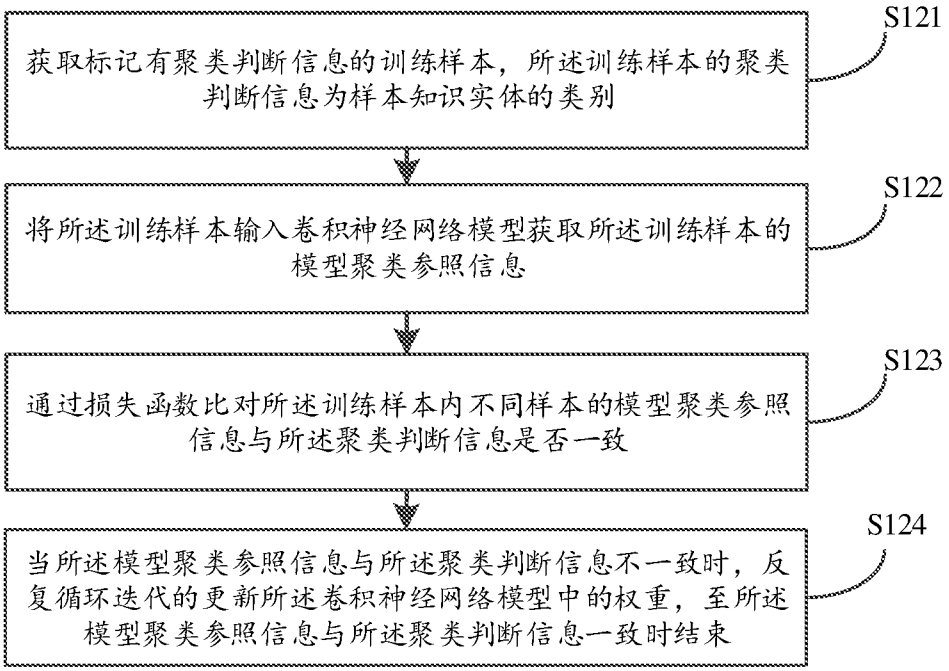

获取标记有聚类判断信息的训练样本，所述训练样本的聚类判断信息为样本知识实体的类别
S121
将所述训练样本输入卷积神经网络模型获取所述训练样本的模型聚类参照信息
S122
通过损失函数比对所述训练样本内不同样本的模型聚类参照信息与所述聚类判断信息是否一致
S123
当所述模型聚类参照信息与所述聚类判断信息不一致时，反复循环迭代的更新所述卷积神经网络模型中的权重，至所述模型聚类参照信息与所述聚类判断信息一致时结束
S124

图 3

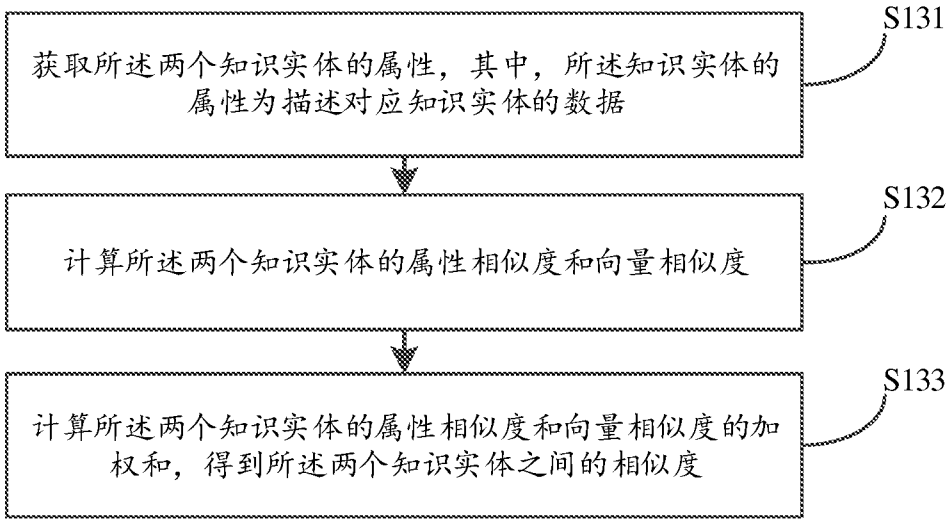

获取所述两个知识实体的属性，其中，所述知识实体的属性为描述对应知识实体的数据
S131
计算所述两个知识实体的属性相似度和向量相似度
S132
计算所述两个知识实体的属性相似度和向量相似度的加权和，得到所述两个知识实体之间的相似度
S133

图 4

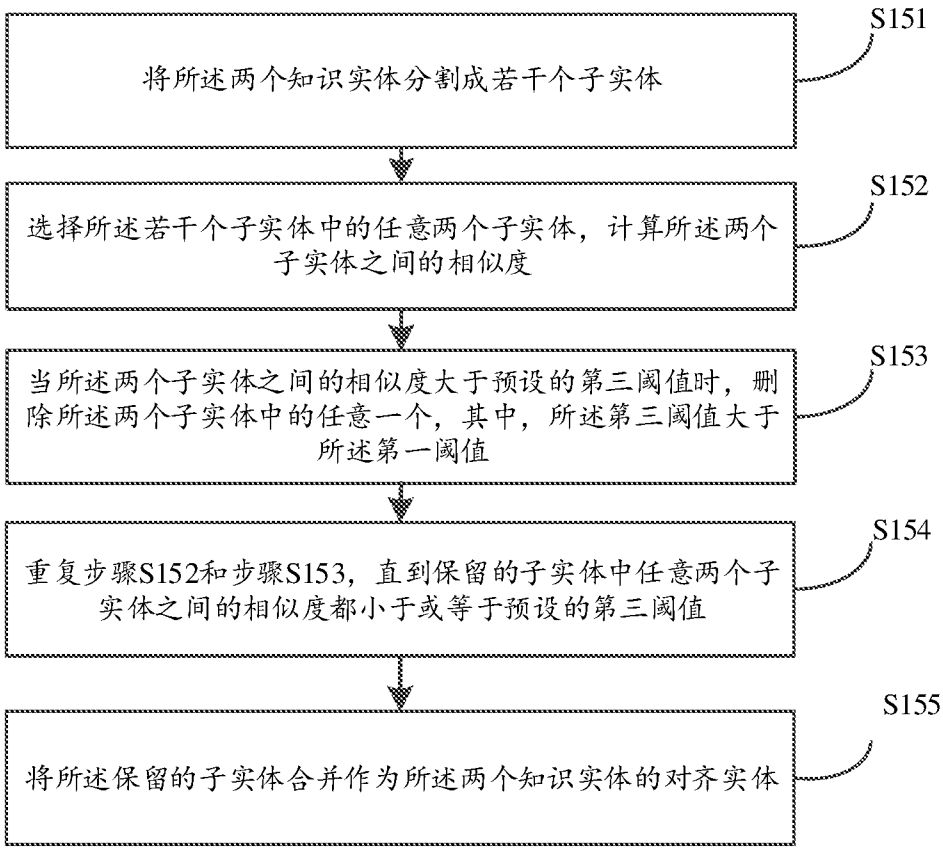

S151
将所述两个知识实体分割成若干个子实体
S152
选择所述若干个子实体中的任意两个子实体，计算所述两个子实体之间的相似度
S153
当所述两个子实体之间的相似度大于预设的第三阈值时，删除所述两个子实体中的任意一个，其中，所述第三阈值大于所述第一阈值
S154
重复步骤S152和步骤S153，直到保留的子实体中任意两个子实体之间的相似度都小于或等于预设的第三阈值
S155
将所述保留的子实体合并作为所述两个知识实体的对齐实体

图 5

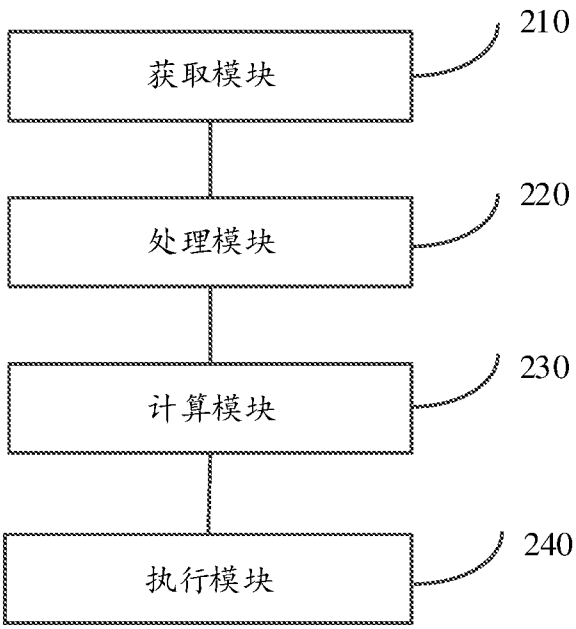

210
获取模块
220
处理模块
230
计算模块
240
执行模块

图 6

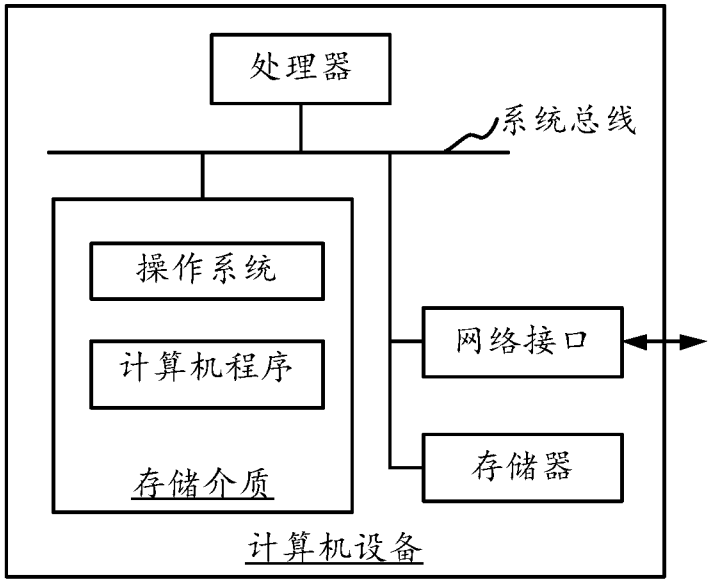


图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/103487

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/28(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)

CNPAT, CNKI, WPI, EPODOC: 知识库, 知识实体, 向量化, 聚类, 相似度, 对齐, 合并, 融合, 归一, knowledge base, knowledge entity, vector, cluster, similarity, align, fusion, normalize

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109783582 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 21 May 2019 (2019-05-21) claims 1-10, and description, paragraphs 0005-0151	1-20
PX	CN 109739939 A (MIOTECH INFORMATION TECH SHANGHAI CO., LTD.) 10 May 2019 (2019-05-10) abstract, and description, paragraphs 0007-0010	1-20
X	CN 108154198 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 12 June 2018 (2018-06-12) description, paragraphs 0059-0153	1-20
A	CN 103699663 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES) 02 April 2014 (2014-04-02) entire document	1-20
A	US 9430738 B1 (MASHWORK, INC.) 30 August 2016 (2016-08-30) entire document	1-20

☐ 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	Date of mailing of the international search report
24 October 2019	**27 November 2019**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/103487

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 109783582 A	21 May 2019	None	
CN 109739939 A	10 May 2019	None	
CN 108154198 A	12 June 2018	EP 3528180 A1	21 August 2019
		US 2019228320 A1	25 July 2019
CN 103699663 A	02 April 2014	None	
US 9430738 B1	30 August 2016	None	

国际检索报告

国际申请号
PCT/CN2019/103487

A. 主题的分类

G06F 16/28(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））

CNPAT,CNKI,WPI,EPODOC：知识库，知识实体，向量化，聚类，相似度，对齐，合并，融合，归一，knowledge base, knowledge entity, vector, cluster, similarity, align, fusion, normalize

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
PX	CN 109783582 A (平安科技深圳有限公司) 2019年 5月 21日 (2019-05-21) 权利要求1-10，说明书第0005-0151段	1-20
PX	CN 109739939 A (颖投信息科技上海有限公司) 2019年 5月 10日 (2019-05-10) 摘要，说明书第0007-0010段	1-20
X	CN 108154198 A (北京百度网讯科技有限公司) 2018年 6月 12日 (2018-06-12) 说明书第0059-0153段	1-20
A	CN 103699663 A (中国科学院自动化研究所) 2014年 4月 2日 (2014-04-02) 全文	1-20
A	US 9430738 B1 (MASHWORK, INC.) 2016年 8月 30日 (2016-08-30) 全文	1-20

☐ 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	国际检索报告邮寄日期
2019年 10月 24日	2019年 11月 27日

ISA/CN的名称和邮寄地址	受权官员
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	王燕花
传真号 (86-10)62019451	电话号码 86-（10）-53961656

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/103487

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 109783582 A	2019年 5月 21日	无	
CN 109739939 A	2019年 5月 10日	无	
CN 108154198 A	2018年 6月 12日	EP 3528180 A1	2019年 8月 21日
		US 2019228320 A1	2019年 7月 25日
CN 103699663 A	2014年 4月 2日	无	
US 9430738 B1	2016年 8月 30日	无	