(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109919316 B
(45)授权公告日 2021.03.12

(21)申请号 201910159408.6

(22)申请日 2019.03.04

(65)同一申请的已公布的文献号
申请公布号 CN 109919316 A

(43)申请公布日 2019.06.21

(73)专利权人 腾讯科技(深圳)有限公司
地址 518044 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 郑博 陈培炫 陈谦

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 郭润湘 李娟

(51)Int.Cl.
G06N 7/00(2006.01)
G06N 3/08(2006.01)

(56)对比文件
CN 109063041 A,2018.12.21
CN 109194707 A,2019.01.11
CN 108805291 A,2018.11.13
CN 108880846 A,2018.11.23
Petar Velickovic,et al..GRAPH ATTENTION NETWORKS.《arXiv》.2018,第1-12页.
涂存超,等..网络表示学习综述.《中国科学: 信息科学》.2017,第47卷(第8期),第980-996页.

审查员 马聪聪

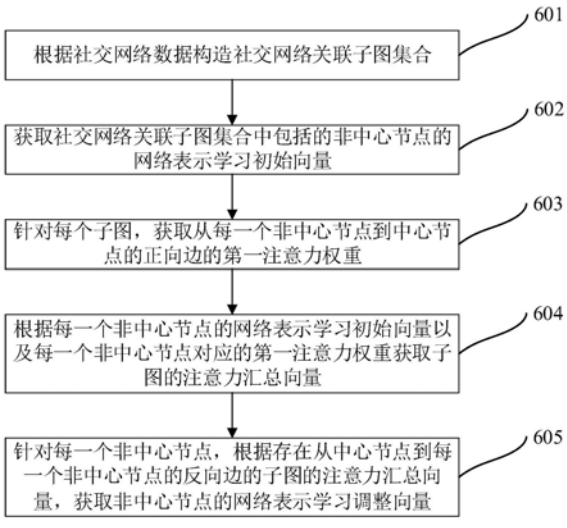
权利要求书3页 说明书19页 附图6页

(54)发明名称

获取网络表示学习向量的方法、装置和设备及存储介质

(57)摘要

本发明公开了一种获取网络表示学习向量的方法、装置和设备及存储介质,属于计算机技术领域,用于在计算代价尽量小的基础上,使得表示学习向量能够表达更多层次的信息。该方法包括:根据社交网络数据构造社交网络关联子图集合;获取所述社交网络关联子图集合中包括的非中心节点的网络表示学习初始向量;针对每个子图,获取从每一个非中心节点到中心节点的正向边的第一注意力权重,并根据每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重获取子图的注意力汇总向量;针对每一个非中心节点,根据存在从中心节点到所述每一个非中心节点的反向边的子图的注意力汇总向量,获取非中心节点的网络表示学习调整向量。



CN 109919316 B

1.一种获取网络表示学习向量的方法,其特征在于,所述方法包括:

根据社交网络数据构造社交网络关联子图集合,所述社交网络关联子图集合中的每一个子图包括一个中心节点以及至少一个非中心节点,用于表征各非中心节点与中心节点之间的关联关系;获取所述社交网络关联子图集合中包括的非中心节点的网络表示学习初始向量;

针对每个子图,获取从每一个非中心节点到中心节点的正向边的第一注意力权重,并根据每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重获取子图的注意力汇总向量,所述第一注意力权重用于表征每一个非中心节点对中心节点的关联重要程度,所述注意力汇总向量用于表征每一个子图的中心节点的特征描述;

针对每一个非中心节点,根据存在从中心节点到所述每一个非中心节点的反向边的子图的注意力汇总向量,获取非中心节点的网络表示学习调整向量;

其中,所述获取从每一个非中心节点到中心节点的正向边的第一注意力权重,包括:

获取每一个子图中任意两个非中心节点的网络表示学习初始向量之间的距离,并根据距离的大小构建每个子图对应的距离直方图;并,根据所述距离直方图与预设参考向量得到每个子图的第一注意力权重矩阵,所述第一注意力权重矩阵中的每一个值为每个子图中的一个非中心节点到中心节点的正向边的第一注意力权重;其中,所述预设参考向量为通过标注了目标值的多组标注数据训练得到的,每组标注数据包括一个非中心节点关联的一个子图以及表征非中心节点与子图之间的关联关系的目标值;或者,

获取每一个子图中每一个非中心节点的网络表示学习初始向量与预设参考向量之间的相似度;其中,所述预设参考向量为各非中心节点的网络表示学习初始向量的平均向量,或者,所述预设参考向量为各非中心节点的网络表示学习初始向量中长度最长的向量;并,基于所述相似度为每一个非中心节点分配所述第一注意力权重。

2.如权利要求1所述的方法,其特征在于,所述根据获取的社交网络数据构造社交网络关联子图集合,包括:

根据获取的所述社交网络数据的类型,获取各个社交网络关联子图的中心节点;

基于所述社交网络数据中非中心节点与各个中心节点之间的关联关系,构造非中心节点与各个中心节点之间的连边,以形成社交网络关联子图。

3.如权利要求1所述的方法,其特征在于,所述根据每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重获取子图的注意力汇总向量,包括:

将每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重进行加权求和,以得到所述注意力汇总向量。

4.如权利要求1所述的方法,其特征在于,所述第一注意力权重包括多组第一子注意力权重,则所述根据每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重获取子图的注意力汇总向量,包括:

将每一个非中心节点的网络表示学习初始向量分别与每一组第一子注意力权重进行加权求和,得到多组子注意力汇总向量;

将多组子注意力汇总向量进行拼接,以得到所述注意力汇总向量。

5.如权利要求1所述的方法，其特征在于，所述根据存在从中心节点到所述每一个非中心节点的反向边的子图的注意力汇总向量，获取非中心节点的网络表示学习调整向量，包括：

获取从每一个中心节点到非中心节点的反向边的第二注意力权重，所述第二注意力权重用于表征每一个中心节点对非中心节点的关联重要程度；

根据每一个子图的注意力汇总向量以及每一个子图的中心节点对应的第二注意力权重，得到所述网络表示学习调整向量。

6.如权利要求1～5任一所述的方法，其特征在于，在根据存在从中心节点到所述每一个非中心节点的反向边的子图的注意力汇总向量，获取非中心节点的网络表示学习调整向量之后，所述方法还包括：

将所述网络表示学习调整向量存储至指定的向量存储服务器中，以使得在业务服务器请求从所述向量存储服务器获取所述网络表示学习调整向量时，所述向量存储服务器能够将所述网络表示学习调整向量提供给所述业务服务器。

7.一种获取网络表示学习向量的装置，其特征在于，所述装置包括：

图构造单元，用于根据社交网络数据构造社交网络关联子图集合，所述社交网络关联子图集合中的每一个子图包括一个中心节点以及至少一个非中心节点，用于表征各非中心节点与中心节点之间的关联关系；

初始向量获取单元，用于获取所述社交网络关联子图集合中包括的非中心节点的网络表示学习初始向量；

向量汇总单元，用于针对每个子图，获取从每一个非中心节点到中心节点的正向边的第一注意力权重，并根据每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重获取子图的注意力汇总向量，所述第一注意力权重用于表征每一个非中心节点对中心节点的关联重要程度，所述注意力汇总向量用于表征每一个子图的中心节点的特征描述；

向量调整单元，用于针对每一个非中心节点，根据存在从中心节点到所述每一个非中心节点的反向边的子图的注意力汇总向量，获取非中心节点的网络表示学习调整向量；

其中，所述向量汇总单元获取从每一个非中心节点到中心节点的正向边的第一注意力权重，包括：

获取每一个子图中任意两个非中心节点的网络表示学习初始向量之间的距离，并根据距离的大小构建每个子图对应的距离直方图；并，根据所述距离直方图与预设参考向量得到每个子图的第一注意力权重矩阵，所述第一注意力权重矩阵中的每一个值为每个子图中的一个非中心节点到中心节点的正向边的第一注意力权重；其中，所述预设参考向量为通过标注了目标值的多组标注数据训练得到的，每组标注数据包括一个非中心节点关联的一个子图以及表征非中心节点与子图之间的关联关系的目标值；或者，

获取每一个子图中每一个非中心节点的网络表示学习初始向量与预设参考向量之间的相似度；其中，所述预设参考向量为各非中心节点的网络表示学习初始向量的平均向量，或者，所述预设参考向量为各非中心节点的网络表示学习初始向量中长度最长的向量；并，基于所述相似度为每一个非中心节点分配所述第一注意力权重。

8.如权利要求7所述的装置，其特征在于，所述图构造单元，具体用于：

根据获取的所述社交网络数据的类型，获取各个社交网络关联子图的中心节点；

基于所述社交网络数据中非中心节点与各个中心节点之间的关联关系，构造非中心节点与各个中心节点之间的连边，以形成社交网络关联子图。

9.如权利要求7所述的装置，其特征在于，所述向量调整单元，具体用于：

获取从每一个中心节点到非中心节点的反向边的第二注意力权重，所述第二注意力权重用于表征每一个中心节点对非中心节点的关联重要程度；

根据每一个子图的注意力汇总向量以及每一个子图的中心节点对应的第二注意力权重，得到所述网络表示学习调整向量。

10.一种计算机设备，其特征在于，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行如权利要求1-6中任一权利要求所述的方法。

11.一种计算机可读存储介质，其特征在于，

所述计算机可读存储介质存储有计算机指令，当所述计算机指令在计算机上运行时，使得计算机能够执行如权利要求1-6中任一权利要求所述的方法。

获取网络表示学习向量的方法、装置和设备及存储介质

技术领域

[0001] 本发明涉及计算机技术领域，特别涉及一种获取网络表示学习向量的方法、装置和设备及存储介质。

背景技术

[0002] 网络的分布式表示是一种网络数据的表示方法，其是将网络中的每个节点映射到一个K维的向量空间，用K维向量来对每个节点进行表示，并且节点向量还能包含一定的语义信息，例如连接紧密的节点向量的距离一般在向量空间中的距离就会很相近，这样就将一个高维度向量表示为低维稠密的实值向量。表示学习则是得到网络表示学习向量的一种方式，其是通过机器学习方法来学习得到网络表示学习向量的过程，或者说是利用网络数据自身表现的相似性或相关性，在向量空间中以距离的形式表达出来的过程，较为常见的表示学习方法有奇异值分解(Singular Value Decomposition，SVD分解)以及Node2Vec等方法。

[0003] 目前，多数的表示学习方法均为目标固定的计算模型，例如Node2Vec算法，其目标是预测节点语料，或称节点袋(Bag of Nodes)中共现节点出现的概率。而在网络数据中，以社交网络为例，信息的层次有多种，不仅包括简单的好友和关注关系，还包括了复杂的社群、讨论组、交互以及点赞等关系，不同的层次包含了不同的信息，而现有的表示学习算法得到的表示学习向量往往只能表达单一层次的信息。当然，也存在将多种信息融合在同一个网络结构中，进行异构网络的学习，但是这种异构网络本质上相当于一个扩充的大网络，对于社交网络而言，这样的网络规模增长在计算上是难以承受的，因此在同样的计算量下，基于异构网络学习算法得到的结果往往还不如一般的表示学习算法。

[0004] 因此，如何在计算代价尽量小的基础上，使得表示学习向量能够表达更多层次的信息是目前亟待解决的技术问题。

发明内容

[0005] 本发明实施例提供一种获取网络表示学习向量的方法、装置和设备及存储介质，用于在计算代价尽量小的基础上，使得表示学习向量能够表达更多层次的信息。

[0006] 一方面，提供一种获取网络表示学习向量的方法，所述方法包括：

[0007] 根据社交网络数据构造社交网络关联子图集合，所述社交网络关联子图集合中的每一个子图包括一个中心节点以及至少一个非中心节点，用于表征各非中心节点与中心节点之间的关联关系；

[0008] 获取所述社交网络关联子图集合中包括的非中心节点的网络表示学习初始向量；

[0009] 针对每个子图，获取从每一个非中心节点到中心节点的正向边的第一注意力权重，并根据每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重获取子图的注意力汇总向量，所述第一注意力权重用于表征每一个非中心节点对中心节点的关联重要程度，所述注意力汇总向量用于表征每一个子图的中心节点的特征

描述；

[0010] 针对每一个非中心节点，根据存在从中心节点到所述每一个非中心节点的反向边的子图的注意力汇总向量，获取非中心节点的网络表示学习调整向量。

[0011] 一方面，提供一种获取网络表示学习向量的装置，所述装置包括：

[0012] 图构造单元，用于根据社交网络数据构造社交网络关联子图集合，所述社交网络关联子图集合中的每一个子图包括一个中心节点以及至少一个非中心节点，用于表征各非中心节点与中心节点之间的关联关系；

[0013] 初始向量获取单元，用于获取所述社交网络关联子图集合中包括的非中心节点的网络表示学习初始向量；

[0014] 向量汇总单元，用于针对每个子图，获取从每一个非中心节点到中心节点的正向边的第一注意力权重，并根据每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重获取子图的注意力汇总向量，所述第一注意力权重用于表征每一个非中心节点对中心节点的关联重要程度，所述注意力汇总向量用于表征每一个子图的中心节点的特征描述；

[0015] 向量调整单元，用于针对每一个非中心节点，根据存在从中心节点到所述每一个非中心节点的反向边的子图的注意力汇总向量，获取非中心节点的网络表示学习调整向量。

[0016] 一方面，提供一种计算机设备，包括：

[0017] 至少一个处理器；以及

[0018] 与所述至少一个处理器通信连接的存储器；其中，

[0019] 所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行上述方面所述的方法。

[0020] 一方面，提供一种计算机可读存储介质，

[0021] 所述计算机可读存储介质存储有计算机指令，当所述计算机指令在计算机上运行时，使得计算机能够执行上述方面所述的方法。

[0022] 本发明实施例中，可以通过获取的社交网络数据构造社交关联子图，根据各子图中各非中心节点已有的网络表示学习初始向量获取该子图的汇总向量，在通过各非中心节点关联的汇总向量反向对非中心节点的向量进行调整，以得到各非中心节点的网络表示学习调整向量。其中，本发明实施例中利用已有的网络学习结果作为前提数据，再通过附加新的网络结构，对之前的学习结果进行调整，这样，则能够将新的信息能够融合到调整向量中，使得调整向量能够表达更多层次的信息。此外，由于本发明实施例采用了注意力机制，其参数规模相比表示学习方法要少很多，因此计算开销大幅下降，节约了成本。

附图说明

[0023] 为了更清楚地说明本发明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，显而易见地，下面描述中的附图仅仅是本发明实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据提供的附图获得其他的附图。

[0024] 图1为本发明实施例提供的应用场景示意图；

[0025] 图2为本发明实施例提供的获取网络表示学习向量的方法的流程示意图；

[0026] 图3为本发明实施例提供的数据库的结构示意图；

[0027] 图4为本发明实施例提供的随机游走采样的示意图；

[0028] 图5为本发明实施例提供的基于Node2Vec算法的表示学习过程的示意图；

[0029] 图6为本发明实施例提供的向量调整过程的流程示意图；

[0030] 图7为本发明实施例提供的构建群组关系子图的示意图；

[0031] 图8为本发明实施例提供的向量调整过程的示意图；

[0032] 图9为本发明实施例提供的向量调整模型的架构示意图；

[0033] 图10为本发明实施例提供的获取网络表示学习向量的装置的一种结构示意图；

[0034] 图11为本发明实施例提供的计算机设备的一种结构示意图。

具体实施方式

[0035] 为使本发明的目的、技术方案和优点更加清楚明白，下面将结合本发明实施例中的附图，对本发明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发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。在不冲突的情况下，本发明中的实施例及实施例中的特征可以相互任意组合。并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0036] 为便于理解本发明实施例提供的技术方案，这里先对本发明实施例使用的一些关键名词进行解释：

[0037] 图表示学习算法：图表示学习算法是将图中的每一个节点利用边的邻接关系学习得到一个表示向量的算法，本发明实施例中主要涉及以Skip-gram框架的图表示学习算法，尤其以Node2Vec、LINE以及DeepWalk等算法为代表，这三者学习过程相似，区别在于邻接关系相似度的定义。

[0038] 星状图：在星状图中，包括一个中心节点和多个非中心节点，其中，仅存在唯一节点是所有其他节点的邻接节点，即中心节点，并且其他节点仅有中心节点这一个邻接节点。

[0039] 迁移学习：迁移学习是运用已存有的知识对不同但相关领域问题进行求解的新的一种机器学习方法，其目的是迁移已有的知识来解决目标领域中仅有少量有标签样本数据甚至没有的学习问题。

[0040] 多层感知器(Multilayer Perceptron，MLP)：多层感知器是一种前向结构的人工神经网络，映射一组输入向量到一组输出向量，MLP可以被看作是一个由多个的节点层所组成的有向图，每一层都全连接到下一层，除了输入节点，每个节点都是一个带有非线性激活函数的神经元(或称处理单元)。一般而言，可以通过基于反向传播算法(Backpropagation algorithm，BP)的监督学习方法来训练MLP。

[0041] 注意力(Attension)机制：即人工神经网络领域所使用的注意力模型。注意力机制的本质来自于人类视觉注意力机制，人们视觉在感知东西的时候一般不会是一个场景从到头看到尾每次全部都看，而往往是根据需求观察注意特定的一部分，而且当人们发现一个场景经常在某部分出现自己想观察的东西时，人们会进行学习在将来再出现类似场景时把

注意力放到该部分上。因此，注意力机制实质上是从大量信息中筛选出高价值信息的手段，在大量信息中，不同信息对于结果的重要性是不同的，这种重要性可以通过赋予不同大小的权值来体现，换言之，注意力机制可以理解成对多个来源进行合成时分配权重的一种规则。具体而言，注意力机制实现了从query和一系列键值对(key-value pair)到输出结果(output)的映射，其中query、key和value都是向量，输出结果的计算通过对value进行加权求和来得到，而每一个value所对应的权值是由query和key通过一个相容性函数来计算获取。即：

[0042] $$\mathrm{Attention}(\mathrm{Query},\ \mathrm{Source}) = \sum_{i=1}^{L_x} Similarity(Query, Key_i) * Value_i$$

[0043] Attention(Query,Source)表示得到的注意力汇总向量，也就是输出结果，Similarity可以是基于Softmax归一化函数计算得到的权值，Softmax归一化函数的参数即为query向量，Source为键值对集合，L_x为Source的长度，或为Source中键值对的数量。上述描述是针对单一注意力机制而言的，在单一注意力机制中采用一组query即可，在多头注意力机制中包含多组query，分别针对每组query进行加权平均得到计算结果之后，最后将多组结果再进行拼接等操作得到最终结果。

[0044] 另外，本文中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。另外，本文中字符“/”，在不做特别说明的情况下，一般表示前后关联对象是一种“或”的关系。

[0045] 目前，多数的表示学习方法均为目标固定的计算模型，例如Node2Vec算法，其目标是预测节点袋中共现节点出现的概率。而在网络数据中，以社交网络为例，信息的层次有多种，不仅包括简单的好友和关注关系，还包括了复杂的社群、讨论组、交互以及点赞等关系，不同的层次包含了不同的信息，而在现有拥有特定目标的算法模型中，很难再加入其他的信息，尤其对于社交网络这种异构且信息丰富的网络而言，大部分信息无法加入网络的表示学习当中。当然，也存在将多种信息融合在同一个网络结构中，进行异构网络的学习，但是这种异构网络本质上相当于一个扩充的大网络，对于社交网络而言，这样的网络规模增长在计算上是难以承受的，因此在同样的计算量下，基于异构网络学习算法得到的结果往往还不如一般的表示学习算法。

[0046] 现有的技术方案的问题主要集中在信息整合和计算代价的矛盾上，本质上而言，这两点的平衡是算法设计偏重所决定的，因此在一个单一的算法框架下，很难做到二者兼得。鉴于此，本发明实施例提供了一种获取网络表示学习向量的方法，在该方法中，可以通过获取的社交网络数据构造社交关联子图，根据各子图中各非中心节点已有的网络表示学习初始向量获取该子图的汇总向量，在通过各非中心节点关联的汇总向量反向对非中心节点的向量进行调整，以得到各非中心节点的网络表示学习调整向量。其中，本发明实施例中利用已有的网络学习结果作为前提数据，再通过附加新的网络结构，对之前的学习结果进行调整，这样，则能够将新的信息能够融合到调整向量中，使得调整向量能够表达更多层次的信息。此外，由于本发明实施例采用了注意力机制，其参数规模相比表示学习方法要少很多，因此计算开销大幅下降，节约了成本。

[0047] 此外，在本发明实施例中，所获取的网络表示学习向量可以存储至指定的向量存储服务器中，这样，在后续有业务需要用到网络表示学习向量时，就可以直接从向量存储服

务器中获取,更加方便快捷。

[0048] 在介绍完本发明实施例的设计思想之后,下面对本发明实施例的技术方案能够适用的应用场景做一些简单介绍,需要说明的是,以下介绍的应用场景仅用于说明本发明实施例而非限定。在具体实施过程中,可以根据实际需要灵活地应用本发明实施例提供的技术方案。

[0049] 请参见图1所示,为发明实施例中的技术方案能够适用的一种应用场景,在该场景中,可以包括业务服务器10、向量存储服务器20、计算集群30、数据库40以及用户终端50。

[0050] 业务服务器10是为实际业务提供方服务的服务器,例如可以为应用服务器或者网站服务器,应用服务器例如可以为微博应用、微信应用或者今日头条等应用(Application,APP)的后台服务器,或者为各应用中内嵌的轻应用的后台服务器,例如微信小程序或者微信的看一看等;网站服务器例如可以为各新闻网站的后台服务器。

[0051] 向量存储服务器20用于存储网络表示学习向量,并向业务服务器10提供向量查询功能,以及业务服务器10可以将自身对网络表示学习向量的要求或者需要对网络表示学习向量进行的操作提供给向量存储服务器20,这样,计算集群30则可以从向量存储服务器20获取到相关参数,并进行相应的计算。

[0052] 计算集群30则用于为计算得到存储网络表示学习向量,以及对网络表示学习向量进行的操作提供计算资源。其中,计算集群30是具备能够承担大规模社交网络表示学习计算能力的计算集群,例如可以是Hadoop或者Spark等分布式计算解决方案,或者还可以是基于消息传递接口(Message-Passing-Interface,MPI)或者共享存储并行编程(Open Multi-Processing,OpenMP)等计算框架构建的并行计算解决方案。

[0053] 数据库40可以用于存储社交网络数据,在计算存储网络表示学习向量时,可以根据实际需求从数据库40中获取相应的社交网络数据。其中,数据库40可以采用结构化查询语言(Structured Query Language,SQL)数据库、基于HIVE的Hadoop分布式文件系统(Hadoop Distributed File System/HIVE,HDFS/HIVE)或者键值(Key-Value,KV)非关系数据存储等多种数据存储方案.

[0054] 用户终端50中可以安装与业务服务器10对应的应用,或者可以通过用户终端50中的浏览器可以打开新闻网站的网页。

[0055] 在实际应用中,例如微信的看一看、微信公众号广告展示和信息红点推送等中,均可利用到本发明实施例的网络表示学习调整向量。例如通过本发明实施例的获取网络表示学习向量的方法得到网络表示学习调整向量并存储至向量存储服务器20中之后,业务服务器10可通过各用户的网络表示学习调整向量之间的相似度,查询相似用户,进而针对相似用户进行广告推荐;或者,基于文章与用户的网络表示学习调整向量之间的相似度高低,对用户客户端界面上的文章进行推荐排序;或者,还可以基于各用户的网络表示学习调整向量,进而社群信息挖掘,在社群有新消息时,则可以向社群包括的用户进行红点推送,以使得用户知道有新消息。具体以微信看一看为例,计算集群30可以根据看一看业务的需求,计算看一看中各文章的网络表示学习向量或者各微信用户的网络表示学习向量,并存储至向量存储服务器20中,业务服务器10则可以从向量存储服务器20中获取所需的网络表示学习向量,例如可以根据获取的微信用户A的网络表示学习向量为该用户推荐文章,以及对推荐的文章进行排序。当然,网络表示学习向量并不限于上述的应用,还可以有其他可能的应

用，本发明实施例对此不做限制。

[0056] 在实际应用中，上述的业务服务器10、向量存储服务器20、计算集群30以及数据库40可以是相互不同的设备，也可以是部分或者全部设备通过相同的设备实现，例如业务服务器10和向量存储服务器20可以是相同的设备。

[0057] 当然，本发明实施例提供的方法并不限用于图1所示的应用场景中，还可以用于其他可能的应用场景，本发明实施例并不进行限制。对于图1所示的应用场景的各个设备所能实现的功能将在后续的方法实施例中一并进行描述，在此先不过多赘述。

[0058] 请参见图2，为本发明实施例提供的获取网络表示学习向量的方法中，主要包括两大过程，即如下的步骤201和步骤202。

[0059] 步骤201：计算网络表示学习初始向量。

[0060] 本发明实施例中，计算网络表示学习初始向量可以采用如现有技术中相同的表示学习方法进行，例如可以表示学习方法可以为SVD矩阵分解及其衍生算法，或者也可以为Node2Vec、LINE或者DeepWalk等表示学习方法。

[0061] 具体的，在计算网络表示学习初始向量时，可以从存储社交网络数据的数据库中获取计算所需的社交网络数据，提交给计算集群后调用相应的表示学习算法，再将计算获得的网络表示学习初始向量存储至向量存储服务器中。

[0062] 其中，在数据库中，社交网络数据可以按照数据的分类进行存储，以便在查询数据，能够尽快的找到所需的数据。具体的，如图3所示，为数据库的结构示意图，其中，数据可以包括但不限于如下类型的子数据库：

[0063] (1)社交关系子数据库

[0064] 社交关系子数据库通常用于存储社交网络中各用户之间的社交关系，例如用户之间的好友关系或者用户之间的关注关系等。

[0065] (2)内容子数据库

[0066] 内容子数据库中通常用于存储社交网络中所产生的内容，内容例如可以包括发布的文章或者动态等内容。

[0067] (3)行为子数据库

[0068] 行为子数据库中通常用于存储各用户在社交网络中的行为记录，用户的行为主要包括用户之间的交互行为，例如用户向好友发送信息或者对好友发布的信息进行点赞或者评论等交互行为，用户的行为还可以包括用户与非用户节点进行的交互，例如用户查看文章或者转发文章等操作行为等。

[0069] (4)用户子数据库

[0070] 用户子数据库中通常用于存储用户信息，例如可以存储用户在社交网络中的社交账号信息以及社交角色相关信息等。

[0071] 实质上，各个子数据库中的数据之间是存在相互联系的，例如行为子数据库中所记录的行为一般都会产生相应的内容，且用户在进行操作时均是在登录了账号的前提下的，此时也会伴随着该账号信息的利用，因此虽然数据虽然分类进行存储，但是数据之间是存在相互联系的。

[0072] 下面以Node2Vec算法为例，对计算网络表示学习初始向量的过程进行介绍。

[0073] 首先，构造节点关系图，并通过对图中节点的邻居进行随机游走采样获得节点袋，

如图4所示，以节点v为起点，随机游走至邻居节点，再从邻居节点出发继续随机游走。其中，在Node2Vec算法中，在随机游走时加入了深度优先搜索(Depth First Search，DFS)和广度优先搜索(Breadth First Search，BFS)的参数α调节结构信息的提取，参数α即为随机游走概率，参数α的取值通过返回概率参数(Return parameter)p和离开概率参数(In outparameter)q控制。v的邻居节点即t、x1、x2和x_3，假设上一步游走的边为(t，v)，即t为游走的上一个节点，x_1、x_2和x_3为v可能游走到的下一个节点，对于节点v的不同邻居，在node2vec算法中根据p和q定义了不同的邻居的跳转概率，p控制跳向上一个节点的邻居的概率，q控制跳向上一个节点的非邻居的概率，换言之，p是控制访问走过的节点，即往回走；q是控制访问还没有走过的节点，即向外走。如图4中所示，当下一个节点x_1与前一个节点t和当前节点v等距时，α＝1；当下一个节点是上一个节点时，α＝1/p，也就是图4中所示的节点t；在其他情况下，α＝1/q。

[0074] 如图5所示，从提供的多层网络的网络数据中游走采样获得节点袋之后，通过Skip-Gram模型的采样-分类过程进行表示学习，Skip-Gram模型的目标是预测节点袋中共现节点出现的概率，共现节点是指两个节点同时出现在一个节点袋中，实质上即用于表征两个节点可能为邻居节点。其中，Skip-Gram模型一般包括输入层、隐藏层和输出层，输入层输入单个节点，输出层输出该节点的相邻节点的概率。Skip-Gram模型将训练一个带有单个隐藏层的神经网络来完成某个任务，但是Skip-Gram模型的目标实际上只是为了学习隐藏层的权重矩阵，该权重矩阵实际上也可被理解为所要学习的网络表示学习初始向量，因此在训练完成后，则可以将输出层舍弃。由于Skip-Gram模型属于现有技术的范畴，因此具体可参见现有技术中Skip-Gram模型的采样-分类过程，在此不再对其过多赘述。

[0075] 步骤202：采用向量表示调整算法，通过社交网络数据对网络表示学习初始向量进行调整，得到网络表示学习调整向量。

[0076] 本发明实施例中，为使得网络表示学习向量能够表达更多的信息，在得到网络表示学习初始向量之后，会采用向量表示调整算法，通过社交网络数据对网络表示学习初始向量进行调整，使得新的社交网络数据融合至网络表示学习初始向量中，进而得到的网络表示学习调整向量则能够表达更多的信息。

[0077] 一般而言，向量调整时所利用的社交网络数据的类型，是与在计算网络表示学习初始向量时所利用的社交网络数据的类型是不同的，这样，最终所得到的网络表示学习调整向量能够表达的信息层次更多；或者，向量调整时所利用的社交网络数据，相较在计算网络表示学习初始向量时所利用的社交网络数据的时新性更强，这样，最终所得到的网络表示学习调整向量所表达的信息时新性更强。

[0078] 下面，将具体针对上述的向量表示调整算法进行详细的描述。请参见图6所示，为本发明实施例所提供的向量表示调整算法的流程示意图。

[0079] 步骤601：根据社交网络数据构造社交网络关联子图集合。

[0080] 本发明实施例中，可以从数据库中读取向量调整所需的社交网络数据，进而根据读取的社交网络数据构造社交网络关联子图集合。其中，在社交网络关联子图集合中的每一个子图中，可以包括一个中心节点以及至少一个非中心节点，每一个子图均可用于表征各非中心节点与中心节点之间的关联关系。

[0081] 在实际应用中，向量调整所需要整合的社交网络数据也可以是上述数据库中的数

据，具体所需的数据可以根据实际应用环境的需求进行设置。社交网络数据具体可以是用户的群关系、活跃群关系、讨论组、收发消息、发表内容(文章、状态、社交图片等)、为他人或内容点赞/评论、发送/接收红包、用户关联地理位置、用户的年龄或者职业等。

[0082] 具体的，在构造社交网络关联子图时，可以根据获取的社交网络数据的类型，获取各个社交网络关联子图的中心节点，再基于社交网络数据中非中心节点与各个中心节点之间的关联关系，构造非中心节点与各个中心节点之间的连边，以形成社交网络关联子图，社交网络关联子图可以表示如下：

[0083] $$G_s = \{G_s^i\} = \{< V_s^i,\ E_s^i >\}$$

[0084] 其中，G_s用于表示社交网络关联子图集合，G_s^i用于表示第i个社交网络关联子图，V_s^i，E_s^i分别为第i个社交网络关联子图的构成元素，V_s^i为第i个社交网络关联子图中的顶点的集合，顶点即为子图涉及的节点，包括中心节点以及非中心节点，E_s^i为第i个社交网络关联子图中的边的集合。

[0085] 下面，以几种不同数据类型的子图构造方法为例，对子图构造过程进行介绍，当然，下述的构造方法仅为一种可行的构造方法，在实际应用中，还可以有其他可能的构造方法，本发明实施例对此不做限制。

[0086] (1)群关系

[0087] 当社交网络数据的数据类型为群关系时，则每个群节点可以构成一个子图G_s^i。具体的，可以将群抽象为一个虚拟节点g，群成员节点u与群节点之间存在双向边，群关系子图可以表示如下：

[0088] $$V_s^i = \{g_i, u_i^j\},\quad E_s^i = \{(g_i, u_i^j), (u_i^j, g_i)\}$$

[0089] 其中，g_i表示第i个群节点，u_i^j则表示第i个群中第j个群成员，(g_i, u_i^j)表示从群节点g_i到第j个群成员u_i^j之间的连边，相反的，(u_i^j, g_i)则表示从第j个群成员u_i^j到群节点g_i之间的连边。

[0090] (2)活跃群关系

[0091] 当社交网络数据的数据类型为活跃群关系时，活跃群关系则更为侧重于管新群中的活跃群成员，与群关系类似的，每个群节点同样可以构成一个子图G_s^i。具体的，可以将群抽象为一个虚拟节点g，活跃群成员节点u_{ia}^j与群节点g之间存在双向边，非活跃群成员只存在从群到成员的单向边，活跃群关系子图可以表示如下：

[0092] $$V_s^i = \{g_i, u_i^j\},\qquad E_s^i = \{(g_i, u_i^j)\} \cup \{(u_{ia}^j, g_i)\}$$

[0093] 其中，g_i表示第i个群节点，u_i^j表示第i个群中第j个群成员，u_{ia}^j表示第i个群中第j个活跃群成员，(g_i, u_i^j)表示从群节点g_i到第j个群成员u_i^j之间的连边，相反的，(u_{ia}^j, g_i)则

表示从活跃群成员u_{ia}^{j}到群节点g_i之间的连边。

[0094] (3)讨论组/邮件组

[0095] 当社交网络数据的数据类型为讨论组或者邮件组时，具体以讨论组为例，则每个讨论组可以构成一个子图G_s^i。具体的，每个讨论组可以抽象为一个虚拟节点t，发消息成员节点u_{is}^{j}与讨论组节点之间存在由讨论组成员到讨论组节点的单向边(u_{is}^{j}, t_i)，收消息群成员u_{ir}^{j}存在从组到成员的单向边(t_i, u_{ir}^{j})，讨论组子图可以表示如下：

[0096] $$V_s^i = \{t_i, u_i^j\}, \qquad E_s^i = \{(t_i, u_{ir}^j)\} \cup \{(u_{is}^j, t_i)\}$$

[0097] 其中，t_i表示第i个讨论组节点，u_i^j表示第i个讨论组中第j个群成员。

[0098] (4)收发消息关系

[0099] 当社交网络数据的数据类型为收发消息关系时，则每个用户构成一个子图G_s^i。具体的，将一段时间内与该用户有收发信行为的用户，取收/发信次数或者收/发信天数最多的K个，或者收/发信次数或者收/发信天数大于一定阈值的用户u_{ir}^{j}/u_{is}^{j}与该用户建立单向边，边的方向是消息发送方到接收方，收发消息关系子图可以表示如下：

[0100] $$V_s^i = \{u_i, u_{ir}^j, u_{is}^j\}, \qquad E_s^i = \{(u_i, u_{ir}^j)\} \cup \{(u_{is}^j, u_i)\}$$

[0101] 其中，u_i表示第i个用户节点，u_{ir}^{j}表示收到u_i所发的消息的用户节点，u_{is}^{j}表示发送消息给u_i的用户节点，(u_i, u_{ir}^{j})表示从u_i到u_{ir}^{j}的连边，(u_{is}^{j}, u_i)表示从u_{is}^{j}到u_i的连边。

[0102] (5)查看、点赞、评论和转发关系

[0103] 当社交网络数据的数据类型为查看、点赞、评论和转发关系时，则可以基于发表的每条内容构成一个子图G_s^i。具体的，可以将每一条内容抽象成为一个虚拟节点c，查看内容的用户u_{iv}^{j}与内容建立从内容到用户的单向边(c_i, u_{iv}^{j})，发表/点赞/评论/转发内容的用户u_{iu}^{j}建立内容与用户的双向边，子图可以表示如下：

[0104] $$V_s^i = \{c_i, u_i^j\}, \qquad E_s^i = \{(c_i, u_{iv}^j)\} \cup \{(u_{iu}^j, c_i), (c_i, u_{iu}^j)\}$$

[0105] (6)发送/接收红包关系

[0106] 当社交网络数据的数据类型为发送/接收红包关系时，则可以针对每个用户u_i构成一个子图G_s^i，将与该用户有收/发红包行为的用户u_{ihb}^{j}建立双向边。

[0107] $$V_s^i = \{u_i, u_{ihb}^j\}, \qquad E_s^i = \{(u_i, u_{ihb}^j), (u_{ihb}^j, u_i)\}$$

[0108] (7)用户与地理位置关系

[0109] 当社交网络数据的数据类型为用户与地理位置关系时，则可以针对每个地理位置

构建一个子图G_s^i。将地理位置抽象成为一个虚拟节点geo，将与该地理位置相关联的用户建立双向边，子图可以表示如下：

[0110] $$V_s^i = \left\{geo_i, u_i^j\right\}, \qquad E_s^i = \left\{\left(geo_i, u_i^j\right), \left(u_i^j, geo_i\right)\right\}$$

[0111] 其中，geo_i表征第i个地理位置，u_i^j表示与第i个地理位置相关联的用户节点。用户与地理位置相关联例如可以是用户到过该地理位置，或者该用户在该地理位置签到，亦或者该用户所发表的内容或者信息中携带有该地理位置等等，当然，还可以有其他的相关联的情况，本发明实施例对此不做限制。

[0112] (8)用户属性

[0113] 当社交网络数据的数据类型为用户属性时，则可以针对每种用户属性构建一个子图G_s^i，用户属性例如可以是用户的职业、年龄或者性别等。具体的，还可以将每种属性按照属性值的不同进行划分，进而将每种属性值抽象成为一个虚拟节点p，再将具有该种属性值的用户与该属性值节点建双向边，子图可以表示如下：

[0114] $$V_s^i = \left\{p_i, u_i^j\right\}, \qquad E_s^i = \left\{\left(p_i, u_i^j\right), \left(u_i^j, p_i\right)\right\}$$

[0115] 其中，p_i用于表征编号为i的属性值，u_i^j则为具有编号为i的属性值的第j个用户节点，$\left(p_i, u_i^j\right)$表示从属性值节点p_i到用户节点u_i^j之间的连边，$\left(u_i^j, p_i\right)$表示从用户节点u_i^j到属性值节点p_i用户节点u_i^j之间的连边。

[0116] 如图7所示，为构建群组关系子图的示意图。其中，左侧为获取的群组关系数据，图中的数字节点表示用户节点u_j，g_i表示第i个群组，可以看到，图7中示出了3个群组，即群组g_1～g_3，其中，群组g_1中包括1、2和7这三个群组成员，群组g_2中包括2、4和5这三个群组成员，群组g_3中包括5、6和7这三个群组成员，图7右侧为基于这些群组关系数据构建的群组关系子图集合$G_s^1 \sim G_s^3$。

[0117] 本发明实施例中，通常构造得到的子图G_s^i是一个星状图，节点i即为该星状图的中心节点。其中，同一个用户节点可能会作为非中心节点出现在不同的子图中，例如，用户A加入了群组B和群组C，则可以针对群组B和群组C分别构造子图G_s^b和G_s^c，用户A可同时作为这两个子图的非中心节点。

[0118] 步骤602：获取社交网络关联子图集合中包括的非中心节点的网络表示学习初始向量。

[0119] 本发明实施例中，根据上述构造的社交网络关联子图，从向量存储服务器中读取所需的网络表示学习初始向量，也就是个子图中涉及到的各非中心节点对应的网络表示学习初始向量。

[0120] 步骤603：针对每个子图，获取从每一个非中心节点到中心节点的正向边的第一注意力权重。

[0121] 本发明实施例中，将从非中心节点到中心节点的连边设定为正向边，而从中心节

点到非中心节点的连边设定为反向边，当然，方向也可以反过来，即从非中心节点到中心节点的连边设定为反向边，而从中心节点到非中心节点的连边设定为正向边，本发明实施例对此不做限制，但需要声明的是，本发明实施例中后续均以前者为例进行说明。

[0122] 在每个子图中，中心节点都是与非中心节点存在一定的关联关系的，因此可以综合各非中心节点的特征来对中心节点进行表达，由于考虑到在一个子图中，各非中心节点对于该子图的中心节点的贡献是不同的，例如在一个群组中，更为活跃的群组成员对群组的影响程度更大，因此，在通过各非中心节点的特征来对中心节点进行表达时，可以为各非中心节点赋予不同的权重值，进而使得中心节点的表达更为准确。

[0123] 具体的，本发明实施例中，采用注意力机制，为各非中心节点赋予不同的权重值，即针对每个子图，为每一条从非中心节点到中心节点的连边，计算第一注意力权重。其中，注意力权重可以是单一的注意力权重，也可以是多头注意力权重组。

[0124] 其中，第一注意力权重的计算方式可以包括但不限于如下几种：

[0125] (1)无需进行模型训练的计算方式

[0126] 具体的，可以获取每一个子图中每一个非中心节点的网络表示学习初始向量key与预设参考向量query之间的相似度，进而根据相似度为每一个非中心节点分配第一注意力权重。一般而言，相似度越高，第一注意力权重值越大。

[0127] 其中，预设参考向量query可以为各非中心节点的网络表示学习初始向量的平均向量，即：

[0128] $$\mathrm{query}=\frac{1}{N-1}\sum_{u_i^j\neq u_i} v_i^j$$

[0129] 其中，v_i^j表示子图i中除中心节点之外的顶点，也就是非中心节点。

[0130] 或者，预设参考向量query还可以为各非中心节点的网络表示学习初始向量中长度最长的向量，即：

[0131] $$\mathrm{query}=\underset{v_i^j}{\mathrm{argmax}}\left\|v_i^j\right\|$$

[0132] (2)需要进行模型训练的计算方式

[0133] 本发明实施例中，相较上述无需进行模型训练的计算方式，模型训练的计算方式可以在模型训练的过程中，加入业务监督信息作为目标，使得训练所得的向量调整模型更适用于具体的业务场景或者特定的任务。对于模型训练的过程，将在后续另行介绍，因此这里先不过多介绍。

[0134] 具体的，需要进行模型训练的计算方式即采用基于子图的距离直方图的注意力机制。利用训练完成的向量调整模型，针对每一个子图，计算子图中任意两个非中心节点的网络表示学习初始向量之间的距离，并根据距离的大小构建每个子图对应的距离直方图，得到的每个子图的距离直方图为N*K的矩阵，N为子图中非中心节点的数量，K为直方图所划分的切分点数量，进而通过矩阵根据子图的距离直方图与预设参考向量query(大小为K*1的矩阵)进行相乘，以得到每个子图的第一注意力权重矩阵(大小为N*1的矩阵)，第一注意力权重矩阵中的每一个值为每个子图中的一个非中心节点到中心节点的正向边的第一注意

力权重。

[0135] 其中，预设参考向量query为通过标注了目标值的多组标注数据训练得到的，每组标注数据包括一个非中心节点关联的一个子图以及表征非中心节点与子图之间的关联关系的目标值，预设参考向量query的获取过程可表示为：

[0136] $$\text{query} = \underset{\theta}{\operatorname{argmin}}(L_\theta(v,l)), \text{key} = \left\{H\left(v_i^j, \text{bin}_t\right)\right\}, \text{value} = v_i^j$$

[0137] $$H\left(v_i^j, \text{bin}_t\right) = \sum_{v_i^k \in G_s} \mathbb{I}\{\text{bin}_{t-1} \le dist\left(v_i^j, v_i^k\right) \le \text{bin}_t\}$$

[0138] 其中，l为监督信息，或称监督标签数据，$L_\theta(v,l)$函数表示向量v在监督信息l下，使用θ作为模型参数的损失函数；H为直方图函数；bin_{t-1}和bin_t表示直方图中的的切分点，即距离区间的划分点，例如bin_t表示直方图中的第t个切分点；dist函数表示两个向量的距离；II{}为指示函数，表示当括号内的值为真的时候输出1，否则输出0，用于配合求和符号用于直方图的计数。

[0139] 步骤604：根据每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重获取子图的注意力汇总向量。

[0140] 本发明实施例中，对于一个子图而言，该子图的第一注意力权重计算完成后，则可以根据该子图中各非中心节点对应的第一注意力权重以及非中心节点的网络表示学习初始向量进行融合，得到该子图的注意力汇总向量。其中，注意力汇总向量即为该子图的描述向量，实质上也就是中心节点的描述向量。例如对于一个群组关系而言，中心节点即为该群组，那么注意力汇总向量则是用于表征该群组的整体描述。

[0141] 具体的，若第一注意力权重为单一注意力权重，则可以直接将网络表示学习初始向量于第一注意力权重进行加权求和得到注意力汇总向量，即对于单一注意力权重而言，计算注意力汇总向量的过程即为简单的线性运算，即：

[0142] $$v_i' = \sum_j \alpha_s^{ij} v_i^j$$

[0143] 其中，v_i^j为第i个子图中的第j个非中心节点，α_s^{ij}为第j个非中心节点对应的第一注意力权重，v'_i为i个子图的注意力汇总向量。

[0144] 如图8所示，每个子图中各非中心节点的网络表示学习初始向量与各自对应的注意力权重α_s^{ij}进行加权求和后，则可以得到该子图的注意力汇总向量。

[0145] 具体的，若第一注意力权重为多头注意力权重，即第一注意力权重包括多组第一子注意力权重，则将每一个非中心节点的网络表示学习初始向量分别与每一组第一子注意力权重进行加权求和，得到多组子注意力汇总向量，再将多组子注意力汇总向量进行拼接，以得到注意力汇总向量，具体计算公式如下：

[0146] $$v_i' = W^T \left[\text{concat}\left(\left\{\sum_j \alpha_s^{ijk} v_i^j\right\}_{k=1}^{n}\right)\right]$$

[0147] 其中，$W \in \mathbb{R}^{nd \times d}$，d是向量的维度，$\alpha_s^{ijk}$为n组第一子注意力权重中的第k组，concat为拼接函数，可以看到，在拼接完成后，为使得最终计算得到的汇总向量v'_i的维度与原向量相同，还需要进行矩阵乘法，即每个子注意力汇总向量的长度为d，共n个子注意力汇总向量，则拼接后总长度为nd，则可以将在拼接后的向量的左边乘以nd*d的矩阵W的转置，从而得到最终向量v_i'的长度为d，与原向量的维度的维度相同，其中，矩阵W也可以是通过训练得到的。

[0148] 步骤605：针对每一个非中心节点，根据存在从中心节点到每一个非中心节点的反向边的子图的注意力汇总向量，获取非中心节点的网络表示学习调整向量。

[0149] 本发明实施例中，获取各个子图的汇总向量之后，则可以根据各个非中心节点所关联的子图反向对非中心节点进行调整，从而得到网络表示学习调整向量。为便于网络表示学习调整向量的查询和调用，获取网络表示学习调整向量之后可以将网络表示学习调整向量存储至指定的向量存储服务器中。

[0150] 具体的，反向调整的过程需要用到子图的反向边，按照用户和反向边收集上述的注意力汇总向量，即可得到以原始子图的各非中心节点为中心节点的反向子图，反向子图中的非中心节点为原始子图中的中心节点。反向子图收集完成后通过一个整合函数f对个注意力汇总向量进行计算，进而得到网络表示学习调整向量。其中，与非中心节点存在关联关系的子图是指存在从中心节点到该非中心节点的反向边的子图。

[0151] 例如对于用户A而言，用户A分别为群组B和群组C的群组成员，且在子图中，用户节点A与群组节点B和群组节点C均存在双向边，分别获取群组B和群组C的注意力汇总向量之后，则可以利用群组B和群组C的注意力汇总向量反向得到用户A的向量，所得到的用户A的向量则为用户A的调整向量。

[0152] 其中，整合函数f可以是简单的平均函数，例如对非中心节点所关联的子图的注意力汇总向量进行求均值计算，均值计算的结果即作为非中心节点的网络表示学习调整向量。

[0153] 本发明实施例中，在整合过程中同样可采用注意力机制，也就是计算各子图的权重然后进行加权平均。具体而言，可以获取从反向子图中每一个非中心节点到中心节点的连边的第二注意力权重β_s^{ij}，第二注意力权重用于表征反向子图中每一个非中心节点对中心节点的关联重要程度，进而再根据反向子图的非中心节点对应的注意力汇总向量以及其对应的第二注意力权重，计算得到网络表示学习调整向量。其中，注意力机制与上述步骤603和步骤604的过程类似，也就是说，其同样可以采用无需训练和需要训练两种方式来执行，其中，若是采用需要训练方式，也可以增加监督结果，采用有监督学习的方式训练神经网络作为整合函数f。如图8所示，每个子图的注意力汇总向量与各自对应的注意力权重β_s^{ij}通过整合函数f计算后，则可以得到原始子图中各非中心节点的网络表示学习调整向量。

[0154] 本发明实施例的获取网络表示学习向量的方法，在已有的表示学习结果上进行向量的合成，相比在新数据上重新训练向量计算量大大减小，使得整个计算成本有很大的下降，具体而言，其计算量从O(NLd)下降为O(Nd)，N为节点数量，L为子图的边数量，d为向量维

度。此外，该方法充分利用了社交网络丰富的数据源，将不同类型的数据整合到网络表示学习向量中，提升网络表示学习向量所表达的信息量，为实际业务提供了更好的辅助。除此之外，整个向量获取过程可以封装成为一个算法流程，以便于业务侧的调用，并且使得在实际使用时性能有明显的提升。

[0155] 在实际应用中，若是需要针对特定任务的调整向量，可以通过有监督训练的方式优化向量调整算法中的相关参数，以实现更优的合成结果。通常情况下，可以通过对每个节点的向量进行一定的限定，从而给定一个优化方向进行优化，得到一个最优化的结果，即为节点的表示向量，在优化目标的前提下，在优化的过程中，将网络的结构、节点的信息、边的信息等“嵌入”到各节点向量中，从而使得节点向量表达更为丰富的信息。下面，将针对需要训练的向量调整算法的训练过程进行介绍。如图9所示，为一种可能的向量调整模型的架构示意图，当然，图9所示仅为一种可能的架构，在实际应用时，还可根据实际需求对图9的架构进行合理的调整，例如可以采用不同的注意力权重计算方式，本发明实施例对此不做限制。

[0156] 携带监督信息的标注数据的格式可以为$D=\{\{G_s^i\}^j, L_j\}$，其中G_s^i是指某个特定的子图，而$\{G_s^i\}^j$则表示与用户u_j所关联的所有同类型子图的集合，L_j是该用户所对应的标签，也是最终的学习目标，它可以是数值或者类别。

[0157] 具体的，该向量调整模型主要包括两部分，即左侧的汇总向量计算模型和右侧的调整向量计算模型。本发明实施例中，采用迁移学习的训练方法，也就是说，在实际训练时，汇总向量计算模型和调整向量计算模型分别作为独立的模型进行训练，对左侧的汇总向量计算模型训练完成后，将训练完成得到的汇总向量作为右侧模型的输入量，以及还可以将左侧的神经网络参数迁移至右侧进行训练(即将左侧的神经网络参数作为右侧神经网络的初始值)。其中，图9中圆角矩形内未需要训练的参数，直角矩形内为实际数据，可以为矩阵或者向量，带箭头的实线表示计算步骤，虚线表示一般表示数据或参数的复用。

[0158] 本发明实施例中，图9所示的汇总向量计算模型主要包括基于子图的距离直方图的注意力机制和多层感知器神经网络，其中，基于子图的距离直方图的注意力机制的输入为构建的子图以及各子图中涉及的非中心节点的网络表示学习初始向量，再通过距离函数计算各子图中两两非中心节点之间的距离，形成子图的距离矩阵，并通过直方图函数构造各子图的距离直方图，再通过距离直方图与query根据设定的矩阵算法计算得到各子图的非中心节点的第一注意力权重，再经Softmax归一化，使得各子图的非中心节点的第一注意力权重之和为1，归一化之后的第一注意力权重即为图9中所示的注意力系数α，通过各子图的非中心节点的网络表示学习初始向量以及各非中心节点对应的注意力系数α进行加权求和即可得到该子图的注意力汇总向量。其中，预设参考向量query是需要训练的参数，在初始训练时，可以赋予query一初始值，再通过多层感知器神经网络对获取的汇总向量进行预测得到预测值，进而基于预测值和给定的目标值通过损失函数计算调整梯度，通过梯度的反向传播训练左侧的整个网络。损失函数可以是预定义的类型，也可以是用户定义的外部库。

[0159] 在左侧的汇总向量计算模型训练完成后，为提升训练速度，可以将左侧的汇总向量计算模型中神经网络的结构以及参数迁移至右侧的调整向量计算模型中的神经网络，以

作为调整向量计算模型中的神经网络的初始值进行训练。当然，也可以不迁移，进而可以使得右侧神经网络的结构与左侧神经网络的结构不同。

[0160] 本发明实施例中，右侧的调整向量计算模型实质上即相当于上述的整合函数f。其中，图9所示的调整向量计算模型中也主要由注意力机制和多层感知器神经网络构成，将左侧获取的各子图的汇总向量同时作为注意力机制的key和value，以及建立以各非中心节点为中心节点的反向子图作为输入，通过距离函数计算各反向子图中长度最长的汇总向量作为query，进而基于key与query之间相似度为各value分配第二注意力权重，再经Softmax归一化，使得各反向子图的汇总向量的第二注意力权重之和为1，归一化之后的第二注意力权重即为图9中所示的注意力系数β，通过各反向子图的汇总向量以及对应的注意力系数β进行加权求和即可得到各反向子图的注意力汇总向量，也就是图9所示的中间向量，再通过多层感知器神经网络对中间向量进行信息浓缩，以得到网络表示学习调整向量，再通过线性层对网络表示学习调整向量进行线性变换之后，通过损失函数计算调整梯度，通过梯度的反向传播训练右侧的整个网络。

[0161] 其中，图9所示的调整向量计算模型需要训练线性层参数以及神经网络参数，而预设参考向量query是无需训练的，即采用的是各非中心节点的网络表示学习初始向量中长度最长的向量作为query的方法。

[0162] 本发明实施例中，为便于后续计算调整向量方便取用，训练得到的参数可以保存在指定的路径中。

[0163] 本发明实施例中，为便于业务服务器与向量存储服务器之间的交互，可以为向量存储服务器设置相应的应用程序接口(Application Programming Interface，API)，例如可以包括向量查询API、相似度计算API以及参数训练API等API，当然，还可以包括其他可能的API，本发明实施例对此不做限制。

[0164] (1)向量查询API

[0165] 向量查询API可以用于业务服务器向向量存储服务器请求所需要使用的网络表示学习向量。其中，业务服务器调用向量查询API时，需要携带必要的参数，以便向量存储服务器知晓业务服务器所需的网络表示学习向量，向量查询API可包括但不限于表1中的参数：

[0166]

参数名称	参数类型	参数说明	备注
实体类型	枚举	用于指示汇总向量节点（中心节点）所代表的意义，可为群/组、地理位置、内容、用户属性、用户等	不可为空，无缺省值
时间点	64位整形	参数位置填入时间点，时间点可采用世界标准世界（UTC），用于指示向量存储服务器返回在此时间点前的最新数据	不可为空，无缺省值
关联数据	枚举	用于指示需要在汇总步骤中进行关联的数据，即子图的非中心节点的含义，通常为用户，也可以是群	可为空，缺省值是用户
向量ID	64位整形数组	用于指示需要返回的向量ID集合	不可为空，无缺省值
使用模型	字符串	用于指示训练得到的模型参数对应的路径，以指示调用其对应训练好的模型进行计算	可为空，缺省值是空

[0167] 表1

[0168] 本发明实施例中，业务服务器调用向量查询API，可以指定实体类型、关联数据类

型以及计算向量所使用的模型等信息，向量存储服务器则会向业务服务器返回相应的向量，例如向量存储服务器返回的内容可以是<向量ID,向量值>的键值对集合。其中，实体类型用于指示业务服务器所需的网络表示学习向量对应的节点类型，例如可以是用户节点、内容节点或者群组节点等，其参数类型为枚举类型，业务服务器调用向量查询API时，实体类型所在位置参数不可为空，且该实体类型所在位置参数没有对应的缺省值，其他参数则可以此类推。

[0169] (2)相似度计算API

[0170] 相似度计算API可以用于业务服务器向向量存储服务器请求计算两个或指定两组网络表示学习向量之间的相似度。相似度计算API可以在向量查询API基础上，将向量查询API中的参数“向量ID”替换为表2中的参数：

[0171]

参数名称	参数类型	参数说明	备注
向量ID1	64位整形数组	用于指示第一个向量的ID，或者第一组向量对应的ID	不可为空，无缺省值
向量ID2	64位整形数组	用于指示第二个向量的ID，或者第二组向量对应的ID	不可为空，无缺省值
掩码矩阵	稀疏布尔（bool）矩阵	用于标记需要计算的向量对，大小为第一组向量数量乘以第二组向量数量，为真的点表示需要计算距离。如果该参数为空，则返回两集合笛卡尔积的距离矩阵。	可为空，缺省值是空
度量类型	枚举	用于指示向量相似度度量的类型，可为内积、余弦、欧氏、正则化欧氏等。	可为空，缺省值可以是内积相似度

[0172] 表2

[0173] 本发明实施例中，向量存储服务器返回的内容是相似度稀疏矩阵，表示第一组向量到第二组向量的相似度。

[0174] (3)参数训练API

[0175] 参数训练API可以用于业务服务器向向量存储服务器提供业务监督信息作为目标值，训练上述向量调整模型参数所用。参数训练API可以在向量查询API基础上，将向量查询API中的参数“向量ID”和“使用模型”替换为表3中的参数：

[0176]

参数名称	参数类型	参数说明	备注
训练集	字符串	用于指示训练数据和标签的路径，通常为分布式存储系统中的路径	不可为空
测试集	字符串	用于指示测试数据和标签（目标值）的路径	不可为空
验证集	字符串	用于指示验证数据和标签（目标值）的路径	可为空
超参数	键值对	该参数为由键值对构成的超参数集合，一般用于指示神经网络的层数、激活函数、学习率、正则化系数、正则化类型	可为空,为空时采用预定的缺省值
损失函数	字符串或函数库地址	该参数为训练时使用的损失函数。对于字符串可为 mse(均方误差)、cross_ent(交叉熵)、hinge(活页函数)等，用户也可以自定义损失函数，通过函数库地址链接进来	可为空,默认采用均方误差 mse
反向注意力	bool	用于指示是否使用反向注意力机制，即在计算调整向量时加入注意力机制	可为空,为空时采用默认值 False
参数保存路径	字符串	用于指示训练得到的参数保存的路径，通常为分布式存储系统中的路径	不可为空，无缺省值

[0177] 表3

[0178] 本发明实施例中,对于API的设计,还可以加入提供上传/下载路径的参数,使得模型、训练数据、验证数据、测试数据以及超参数都能够从外部地址读取得到,这样可以方便数据互通的能力,而不必将所有数据存储在指定的分布式系统中。当然,在具体实现时,需要添加将这些数据传输到分布式系统各个节点的功能模块,以免造成单点读取瓶颈。

[0179] 综上所述,本发明实施例利用现有的网络表示学习向量作为前提数据,例如基于Node2Vec算法得到的网络表示学习向量,在此基础上,再通过附加新的网络结构(例如社群和讨论组等)或者新的信息(例如消息互动等)对之前的学习得到的向量进行调整,使得新的信息能够融合到其中。同时,还可以加入监督信息对调整过程提供目标,可以实现对有监督学习的近似。此外,本发明实施例采用注意力机制,其参数规模相比之前的表示学习方法要少很多,进而降低计算开销,节约成本。

[0180] 由上述描述可知,本发明实施例的整个过程主要包括两个步骤,因此可以随时调整需要融入的信息,使得算法的灵活性和易用性有更大的提升,而不必局限于复杂图构造和参数调节等技巧中,最后得到的结果信息表达相比原始的向量更加丰富,对于业务使用更为简单有效。

[0181] 请参见图10,基于同一发明构思,本发明实施例还提供了一种获取网络表示学习向量的装置100,包括:

[0182] 图构造单元1001,用于根据社交网络数据构造社交网络关联子图集合,社交网络关联子图集合中的每一个子图包括一个中心节点以及至少一个非中心节点,用于表征各非中心节点与中心节点之间的关联关系;

[0183] 初始向量获取单元1002,用于获取社交网络关联子图集合中包括的非中心节点的网络表示学习初始向量;

[0184] 向量汇总单元1003,用于针对每个子图,获取从每一个非中心节点到中心节点的正向边的第一注意力权重,并根据每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重获取子图的注意力汇总向量,第一注意力权重用于表征

每一个非中心节点对中心节点的关联重要程度，注意力汇总向量用于表征每一个子图的中心节点的特征描述；

[0185] 向量调整单元1004，用于针对每一个非中心节点，根据存在从中心节点到该非中心节点的反向边的子图的注意力汇总向量，获取非中心节点的网络表示学习调整向量。

[0186] 可选的，图构造单元1001，具体用于：

[0187] 根据获取的社交网络数据的类型，获取各个社交网络关联子图的中心节点；

[0188] 基于社交网络数据中非中心节点与各个中心节点之间的关联关系，构造非中心节点与各个中心节点之间的连边，以形成社交网络关联子图。

[0189] 可选的，向量汇总单元1003，具体用于：

[0190] 获取每一个子图中任意两个非中心节点的网络表示学习初始向量之间的距离，并根据距离的大小构建每个子图对应的距离直方图；

[0191] 根据距离直方图与预设参考向量得到每个子图的第一注意力权重矩阵，第一注意力权重矩阵中的每一个值为每个子图中的一个非中心节点到中心节点的正向边的第一注意力权重；其中，预设参考向量为通过标注了目标值的多组标注数据训练得到的，每组标注数据包括一个非中心节点关联的一个子图以及表征非中心节点与子图之间的关联关系的目标值。

[0192] 可选的，向量汇总单元1003，具体用于：

[0193] 获取每一个子图中每一个非中心节点的网络表示学习初始向量与预设参考向量之间的相似度；其中，预设参考向量为各非中心节点的网络表示学习初始向量的平均向量，或者，预设参考向量为各非中心节点的网络表示学习初始向量中长度最长的向量；

[0194] 基于相似度为每一个非中心节点分配第一注意力权重。

[0195] 可选的，向量汇总单元1003，具体用于：

[0196] 将每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重进行加权求和，以得到注意力汇总向量。

[0197] 可选的，向量汇总单元1003，具体用于：

[0198] 将每一个非中心节点的网络表示学习初始向量分别与每一组第一子注意力权重进行加权求和，得到多组子注意力汇总向量；

[0199] 将多组子注意力汇总向量进行拼接，以得到注意力汇总向量。

[0200] 可选的，向量调整单元1004，具体用于：

[0201] 获取从每一个中心节点到非中心节点的反向边的第二注意力权重，第二注意力权重用于表征每一个中心节点对非中心节点的关联重要程度；

[0202] 根据每一个子图的注意力汇总向量以及每一个子图的中心节点对应的第二注意力权重，得到网络表示学习调整向量。

[0203] 可选的，装置还包括向量存放单元1005，用于：

[0204] 将网络表示学习调整向量存储至指定的向量存储服务器中，以使得在业务服务器请求从向量存储服务器获取网络表示学习调整向量时，向量存储服务器能够将网络表示学习调整向量提供给业务服务器。

[0205] 该装置可以用于执行图3～图9所示的实施例中所示的方法，因此，对于该装置的各功能模块所能够实现的功能等可参考图3～图9所示的实施例的描述，不多赘述。其中，向

量存放单元1005虽然在图10中一并示出，但需要知道的是，向量存放单元1005并不是必选的功能单元，因此在图10中以虚线示出。

[0206] 请参见图11，基于同一技术构思，本发明实施例还提供了一种计算机设备110，可以包括存储器1101和处理器1102。

[0207] 所述存储器1101，用于存储处理器1102执行的计算机程序。存储器1101可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序等；存储数据区可存储根据计算机设备的使用所创建的数据等。处理器1102，可以是一个中央处理单元(central processing unit，CPU)，或者为数字处理单元等等。本发明实施例中不限定上述存储器1101和处理器1102之间的具体连接介质。本发明实施例在图11中以存储器1101和处理器1102之间通过总线1103连接，总线1103在图11中以粗线表示，其它部件之间的连接方式，仅是进行示意性说明，并不引以为限。所述总线1103可以分为地址总线、数据总线、控制总线等。为便于表示，图11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。

[0208] 存储器1101可以是易失性存储器(volatile memory)，例如随机存取存储器(random-access memory，RAM)；存储器1101也可以是非易失性存储器(non-volatile memory)，例如只读存储器，快闪存储器(flash memory)，硬盘(hard disk drive，HDD)或固态硬盘(solid-state drive，SSD)、或者存储器1101是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1101可以是上述存储器的组合。

[0209] 处理器1102，用于调用所述存储器1101中存储的计算机程序时执行如图3～图9中所示的实施例中设备所执行的方法。

[0210] 在一些可能的实施方式中，本发明提供的方法的各个方面还可以实现为一种程序产品的形式，其包括程序代码，当所述程序产品在计算机设备上运行时，所述程序代码用于使所述计算机设备执行本说明书上述描述的根据本发明各种示例性实施方式的方法中的步骤，例如，所述计算机设备可以执行如图3～图9中所示的实施例中设备所执行的方法。

[0211] 所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

[0212] 尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

[0213] 显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内，则本发明也意图包含这些改动和变型在内。

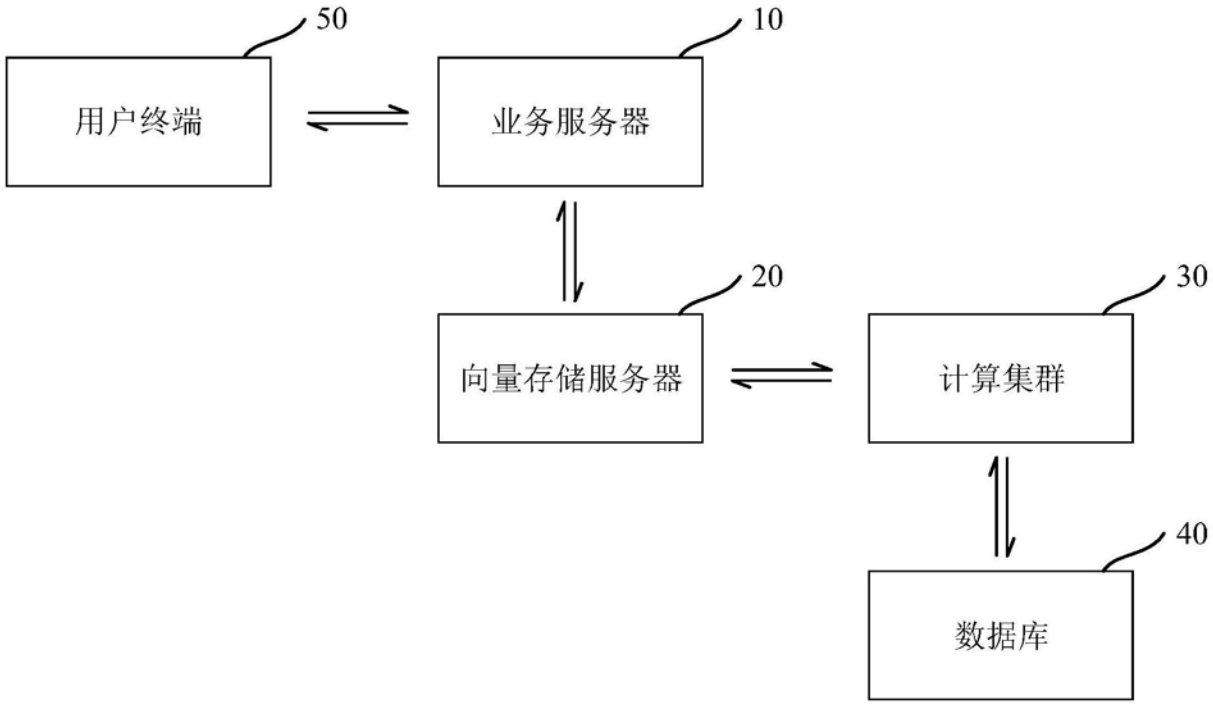

50
用户终端
10
业务服务器
20
向量存储服务器
30
计算集群
40
数据库

图1

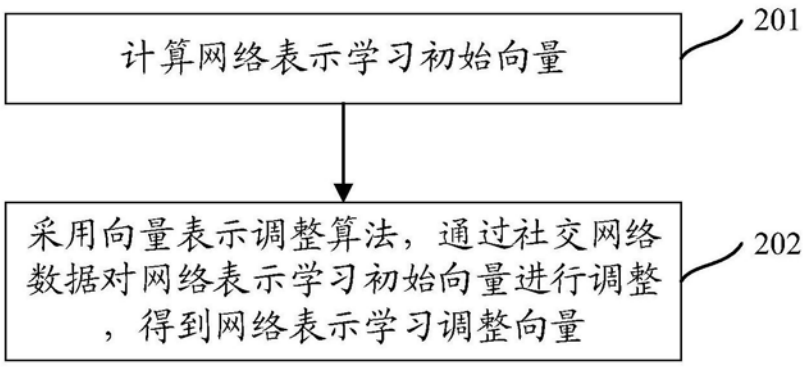

计算网络表示学习初始向量
201
采用向量表示调整算法，通过社交网络数据对网络表示学习初始向量进行调整，得到网络表示学习调整向量
202

图2

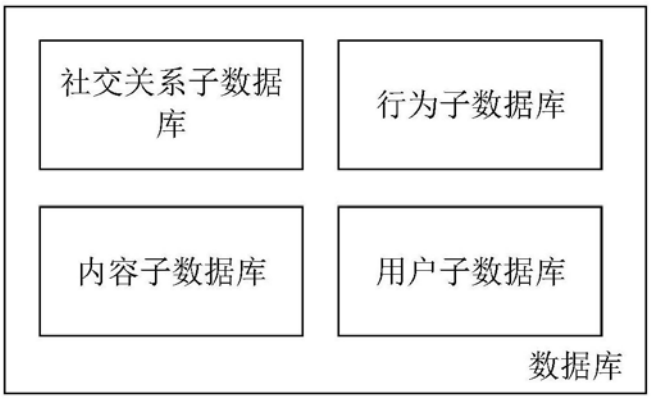


图3

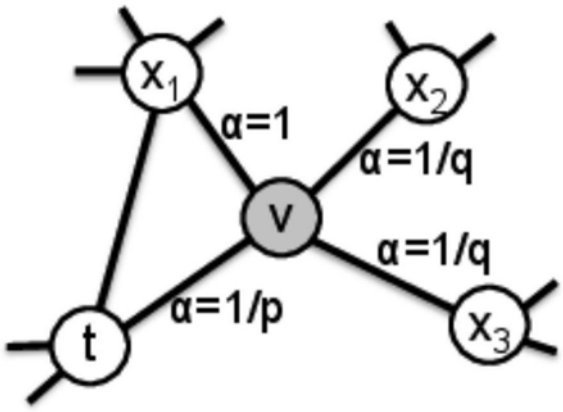


图4

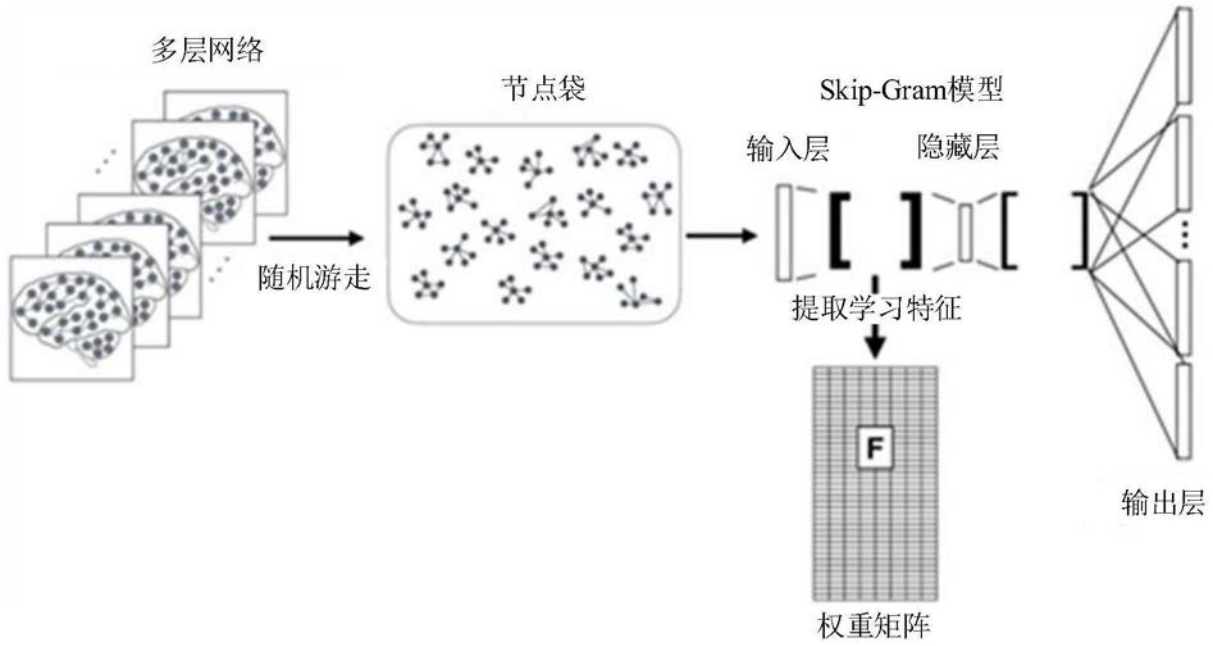


图5

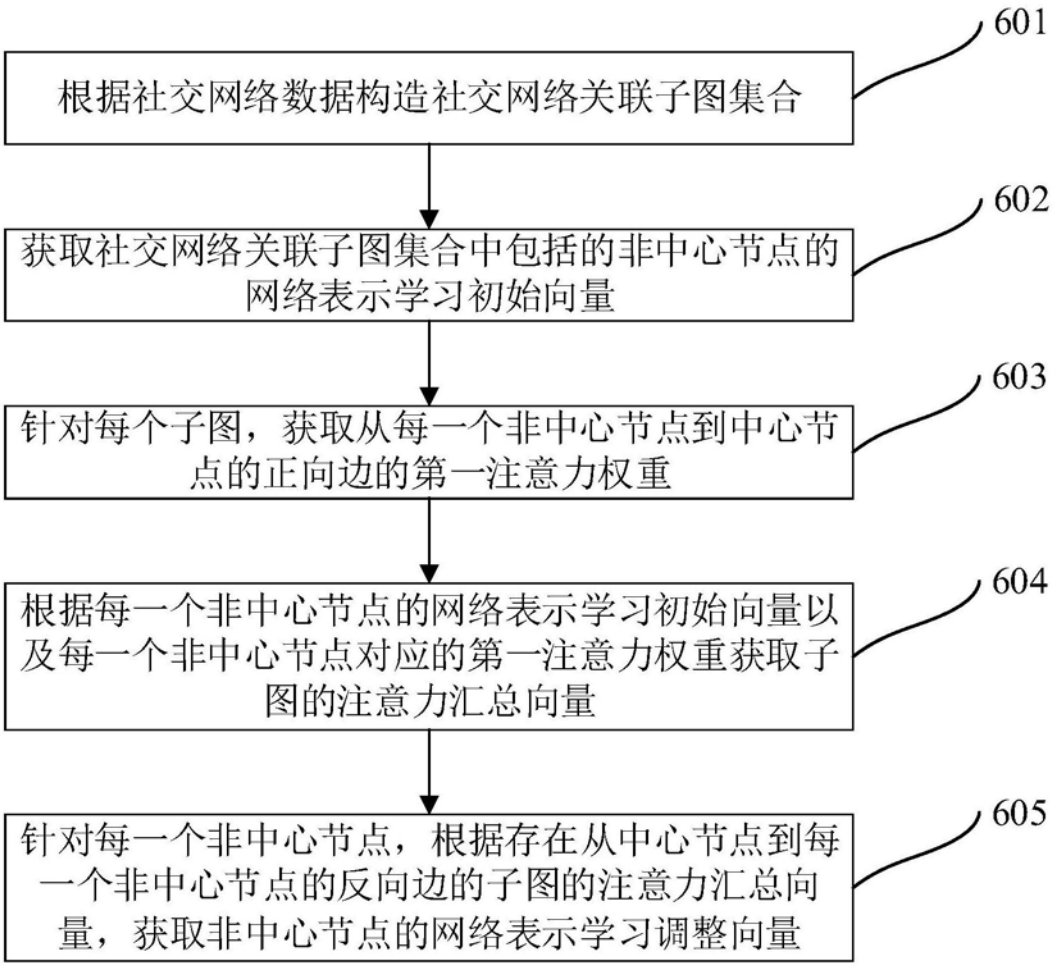

601
根据社交网络数据构造社交网络关联子图集合
602
获取社交网络关联子图集合中包括的非中心节点的网络表示学习初始向量
603
针对每个子图，获取从每一个非中心节点到中心节点的正向边的第一注意力权重
604
根据每一个非中心节点的网络表示学习初始向量以及每一个非中心节点对应的第一注意力权重获取子图的注意力汇总向量
605
针对每一个非中心节点，根据存在从中心节点到每一个非中心节点的反向边的子图的注意力汇总向量，获取非中心节点的网络表示学习调整向量

图6

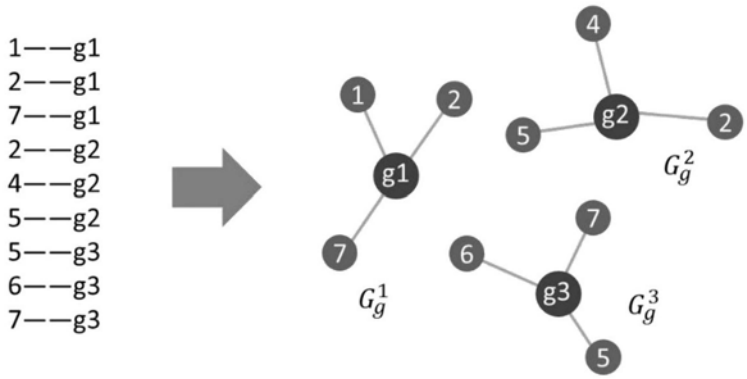

1——g1
2——g1
7——g1
2——g2
4——g2
5——g2
5——g3
6——g3
7——g3
g1
g2
g3
G_g^1
G_g^2
G_g^3

图7

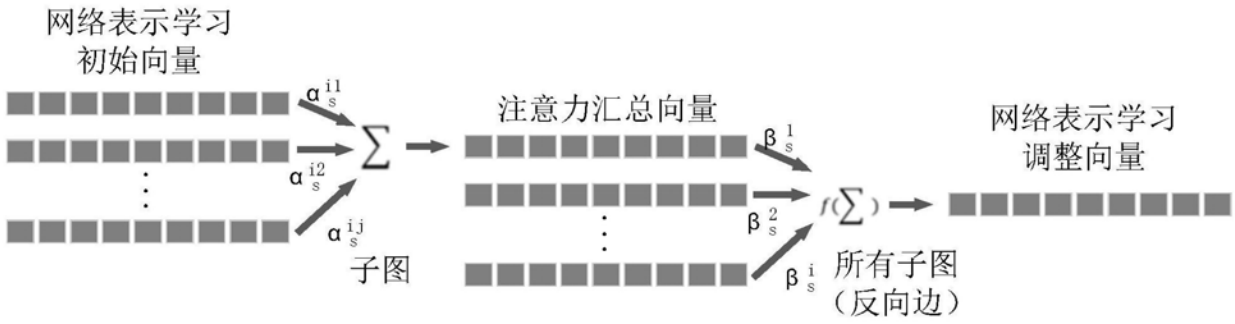

网络表示学习
初始向量
子图
注意力汇总向量
所有子图
（反向边）
网络表示学习
调整向量

图8

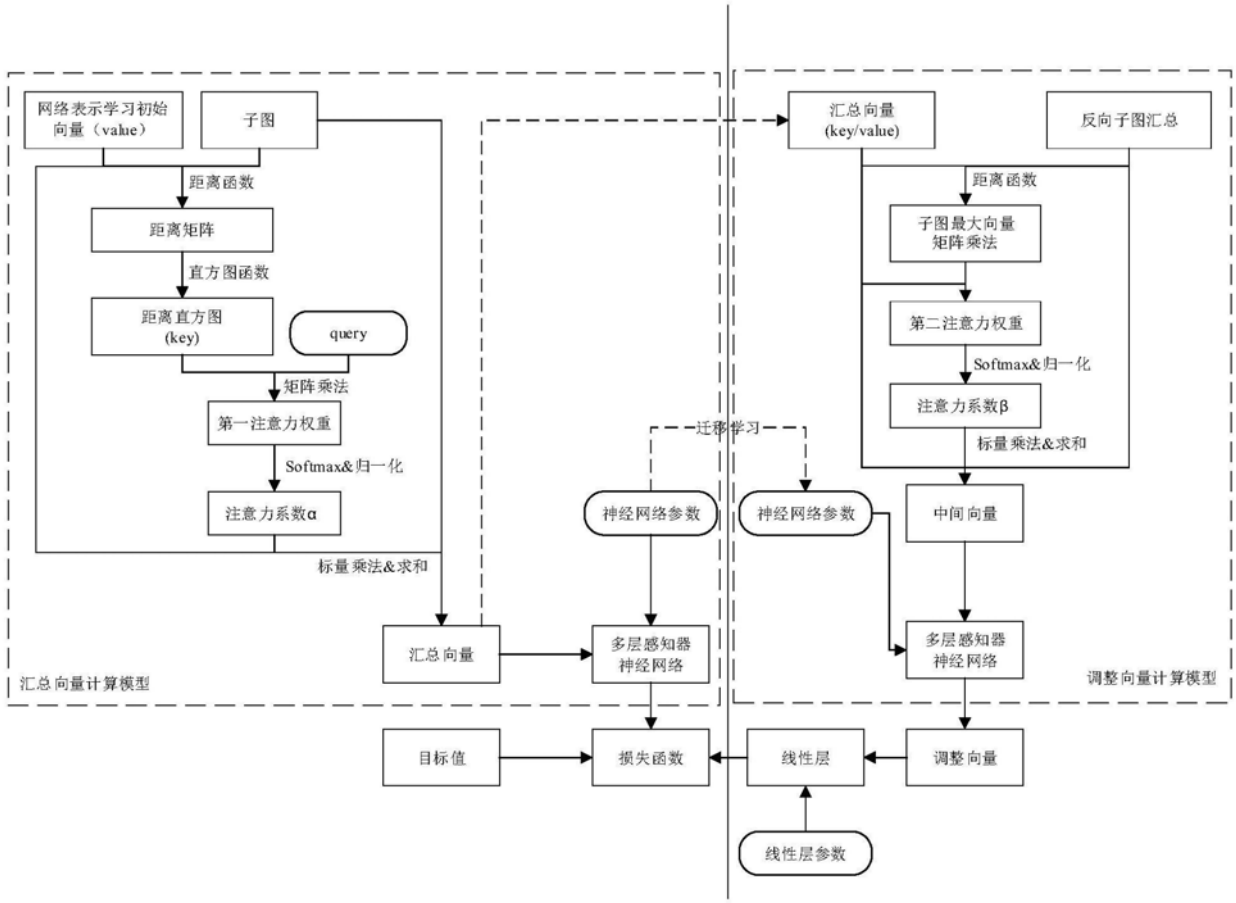

网络表示学习初始向量（value）
子图
距离函数
距离矩阵
直方图函数
距离直方图(key)
query
矩阵乘法
第一注意力权重
Softmax&归一化
注意力系数α
标量乘法&求和
汇总向量
多层感知器神经网络
汇总向量计算模型
迁移学习
神经网络参数
神经网络参数
汇总向量(key/value)
反向子图汇总
距离函数
子图最大向量矩阵乘法
第二注意力权重
Softmax&归一化
注意力系数β
标量乘法&求和
中间向量
多层感知器神经网络
调整向量计算模型
目标值
损失函数
线性层
调整向量
线性层参数

图9

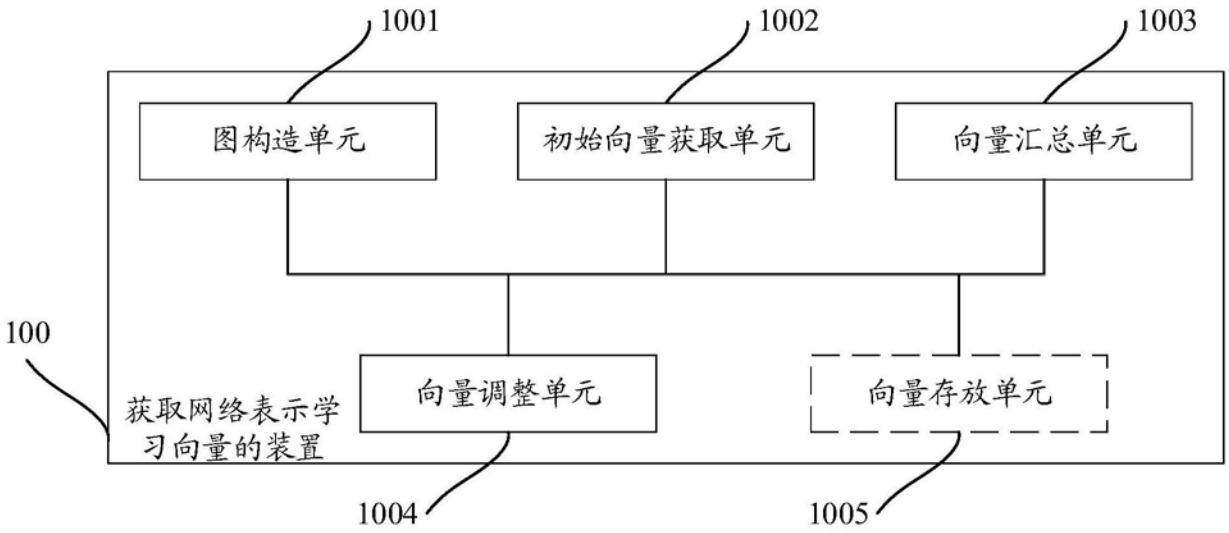

1001
1002
1003
图构造单元
初始向量获取单元
向量汇总单元
100
获取网络表示学习向量的装置
向量调整单元
向量存放单元
1004
1005

图10

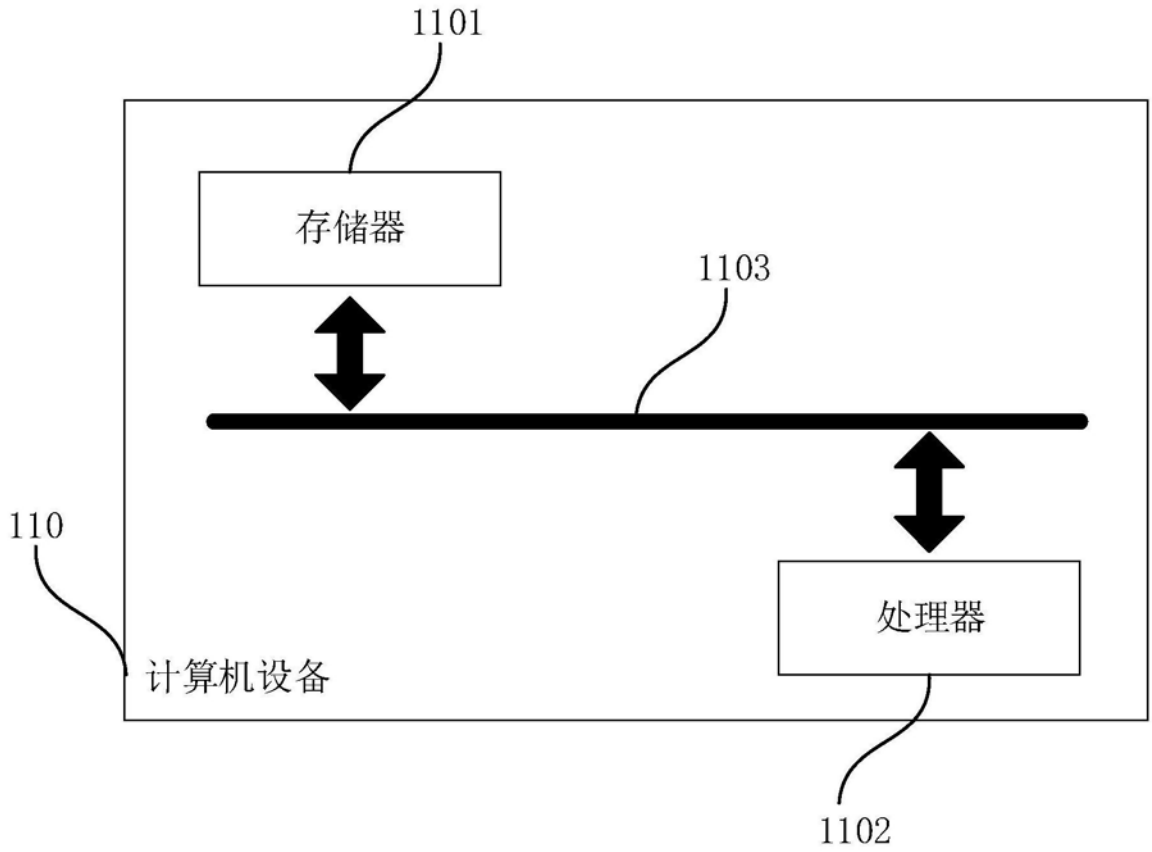

1101
存储器
1103
110
计算机设备
处理器
1102

图11