(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 108520470 A
(43)申请公布日 2018.09.11

(21)申请号 201710113238.9

(22)申请日 2017.02.28

(71)申请人 百度在线网络技术(北京)有限公司
地址 100085 北京市海淀区上地十街10号百度大厦三层

(72)发明人 马超 朱琛 祝恒书 熊辉

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
代理人 王达佐 马晓亚

(51)Int.Cl.
G06Q 50/00(2012.01)

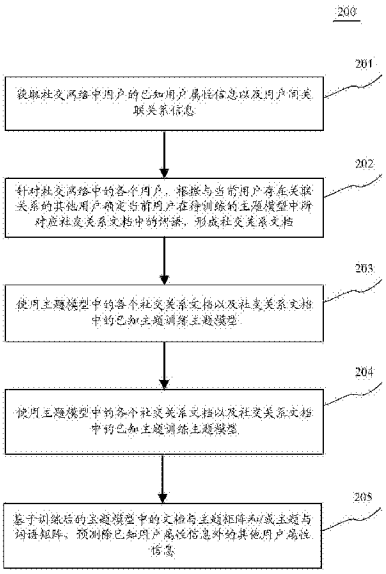
权利要求书3页 说明书11页 附图3页

(54)发明名称

用于生成用户属性信息的方法和装置

(57)摘要

本申请公开了用于生成用户属性信息的方法和装置。该方法的一具体实施方式包括:获取社交网络中用户的已知用户属性信息以及用户间关联关系信息;针对社交网络中的各个用户,根据与当前用户存在关联关系的其他用户确定当前用户在待训练的主题模型中所对应社交关系文档中的词语;根据已知用户属性信息,设置已知用户属性信息所涉及的用户对应的社交关系文档中的已知主题;训练主题模型;基于训练后的主题模型中的文档与主题矩阵和/或主题与词语矩阵,预测除已知用户属性信息外的其他用户属性信息。该实施方式实现了在仅有部分用户属性信息时对其他用户属性信息的预测。



CN 108520470 A

1.一种用于生成用户属性信息的方法，其特征在于，所述方法包括：

获取社交网络中用户的已知用户属性信息以及用户间关联关系信息，所述用户间关联关系信息用于表征不同用户之间是否存在关联关系；

针对所述社交网络中的各个用户，根据与当前用户存在关联关系的其他用户确定所述当前用户在待训练的主题模型中所对应社交关系文档中的词语，形成社交关系文档；

根据所述已知用户属性信息，设置所述已知用户属性信息所涉及的用户对应的社交关系文档中的已知主题；

使用所述主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练所述主题模型；

基于训练后的主题模型中的文档与主题矩阵和/或主题与词语矩阵，预测除所述已知用户属性信息外的其他用户属性信息。

2.根据权利要求1所述的方法，其特征在于，所述用户间关联关系信息还用于表征用户之间存在的关联关系的关联程度。

3.根据权利要求1或2所述的方法，其特征在于，所述获取社交网络中用户的已知用户属性信息以及用户间关联关系信息，包括：

获取所述社交网络中用户的操作行为信息；

根据所获取的操作行为信息，生成所述用户间关联关系信息。

4.根据权利要求2所述的方法，其特征在于，所述方法还包括：

构造以所述社交网络中的各个用户为节点、关联程度作为边的社交网络图；

根据所述已知用户属性信息以及所预测的其他用户属性信息，对所述已知用户属性信息以及所述其他用户属性信息所涉及的用户在所述社交网络图的节点设置已知标签；

对所述社交网络图执行标签传播算法；

根据执行标签传播算法后的社交网络图的节点中概率最大的标签，生成节点所对应的用户的用户属性信息。

5.根据权利要求3所述的方法，其特征在于，所述根据所述已知用户属性信息以及所预测的其他用户属性信息，对所述已知用户属性信息以及所述其他用户属性信息所涉及的用户在所述社交网络图的节点设置已知标签，包括：

根据所述已知用户属性信息以及设定规则，确定所述已知用户属性信息所涉及的用户的初始标签并将初始标签分配给所述社交网络图对应的节点；

根据所预测的其他用户属性以及所述设定规则，确定所述其他用户属性信息所涉及的用户的初始标签并进行衰减后分配给所述社交网络图对应的节点。

6.根据权利要求5所述的方法，其特征在于，所述对所述社交网络图执行标签传播算法，包括：

迭代执行标签传播步骤，直至各个节点的标签收敛；

其中，每次标签传播步骤包括：将每个节点的标签按照节点之间的边所对应的概率传递至相邻节点；将所述已知用户属性信息涉及的用户对应的节点重置为节点的初始标签，将所述其他用户属性信息涉及的用户对应的节点重置为对节点的初始标签进行衰减所得到的标签。

7.一种用于生成用户属性信息的装置，其特征在于，所述装置包括：

获取单元，用于获取社交网络中用户的已知用户属性信息以及用户间关联关系信息，所述用户间关联关系信息用于表征不同用户之间是否存在关联关系；

确定单元，用于针对所述社交网络中的各个用户，根据与当前用户存在关联关系的其他用户确定所述当前用户在待训练的主题模型中所对应社交关系文档中的词语，形成社交关系文档；

主题设置单元，用于根据所述已知用户属性信息，设置所述已知用户属性信息所涉及的用户对应的社交关系文档中的已知主题；

训练单元，用于使用所述主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练所述主题模型；

预测单元，用于基于训练后的主题模型中的文档与主题矩阵和/或主题与词语矩阵，预测除所述已知用户属性信息外的其他用户属性信息。

8.根据权利要求7所述的装置，其特征在于，所述用户间关联关系信息还用于表征用户之间存在的关联关系的关联程度。

9.根据权利要求7或8所述的装置，其特征在于，所述获取单元进一步用于：

获取所述社交网络中用户的操作行为信息；

根据所获取的操作行为信息，生成所述用户间关联关系信息。

10.根据权利要求8所述的装置，其特征在于，所述装置还包括：

构造单元，用于构造以所述社交网络中的各个用户为节点、关联程度作为边的社交网络图；

标签设置单元，用于根据所述已知用户属性信息以及所预测的其他用户属性信息，对所述已知用户属性信息以及所述其他用户属性信息所涉及的用户在所述社交网络图的节点设置已知标签；

执行单元，用于对所述社交网络图执行标签传播算法；

生成单元，用于根据执行标签传播算法后的社交网络图的节点中概率最大的标签，生成节点所对应的用户的用户属性信息。

11.根据权利要求10所述的装置，其特征在于，所述标签设置单元进一步用于：

根据所述已知用户属性信息以及设定规则，确定所述已知用户属性信息所涉及的用户的初始标签并将初始标签分配给所述社交网络图对应的节点；

根据所预测的其他用户属性以及所述设定规则，确定所述其他用户属性信息所涉及的用户的初始标签并进行衰减后分配给所述社交网络图对应的节点。

12.根据权利要求11所述的装置，其特征在于，所述执行单元进一步用于：

迭代执行标签传播步骤，直至各个节点的标签收敛；

其中，每次标签传播步骤包括：将每个节点的标签按照节点之间的边所对应的概率传递至相邻节点；将所述已知用户属性信息涉及的用户对应的节点重置为节点的初始标签，将所述其他用户属性信息涉及的用户对应的节点重置为对节点的初始标签进行衰减所得到的标签。

13.一种设备，包括：

一个或多个处理器；

存储装置，用于存储一个或多个程序，

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求1-6中任一所述的方法。

14.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现如权利要求1-6中任一所述的方法。

用于生成用户属性信息的方法和装置

技术领域

[0001] 本申请涉及计算机技术领域，具体涉及数据挖掘技术领域，尤其涉及用于生成用户属性信息的方法和装置。

背景技术

[0002] 用户画像是指最能代表某个用户的一系列特征，包括但不限于年龄、性别、居住地、工作地点等。用户画像数据有利于执行对用户的个性化信息推送、个性化搜索等个性化服务。然而，用户在互联网中填写的数据常常并不完整，因此有必要通过数据挖掘技术从互联网数据中挖掘出较为完整的用户画像数据。

[0003] 现有技术在从互联网领域中挖掘用户画像数据时，通常采用传统的有监督的分类方法以及无监督的聚类方法。然而，前一种方法仅仅基于已知信息的用户做模型训练，但由于已知的信息首先很少、难以获得且真假难辨，会导致分析结果会产生有很大的偏差。同时分类算法的准确度极大的依赖于模型的特征选取，而特征选取通常依靠人工设计，会导致模型的适用范围和通用性比较局限。后一种方法则完全不考虑用户自己的信息，仅仅基于网络的无监督的方法的性能主要取决于启发式的网络划分结果的好坏，预测的结果亦不够准确。

发明内容

[0004] 本申请的目的在于提出一种改进的用于生成用户属性信息的方法和装置，来解决以上背景技术部分提到的技术问题。

[0005] 第一方面，本申请实施例提供了一种用于生成用户属性信息的方法，该方法包括：获取社交网络中用户的已知用户属性信息以及用户间关联关系信息，用户间关联关系信息用于表征不同用户之间是否存在关联关系；针对社交网络中的各个用户，根据与当前用户存在关联关系的其他用户确定当前用户在待训练的主题模型中所对应社交关系文档中的词语，形成社交关系文档；根据已知用户属性信息，设置已知用户属性信息所涉及的用户对应的社交关系文档中的已知主题；使用主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练主题模型；基于训练后的主题模型中的文档与主题矩阵和/或主题与词语矩阵，预测除已知用户属性信息外的其他用户属性信息。

[0006] 在一些实施例中，用户间关联关系信息还用于表征用户之间存在的关联关系的关联程度。

[0007] 在一些实施例中，上述获取社交网络中用户的已知用户属性信息以及用户间关联关系信息，包括：获取社交网络中用户的操作行为信息；根据所获取的操作行为信息，生成用户间关联关系信息。

[0008] 在一些实施例中，上述方法还包括：构造以社交网络中的各个用户为节点、关联程度作为边的社交网络图；根据已知用户属性信息以及所预测的其他用户属性信息，对已知用户属性信息以及其他用户属性信息所涉及的用户在社交网络图的节点设置已知标签；对

社交网络图执行标签传播算法；根据执行标签传播算法后的社交网络图的节点中概率最大的标签，生成节点所对应的用户的用户属性信息。

[0009] 在一些实施例中，上述根据已知用户属性信息以及所预测的其他用户属性信息，对已知用户属性信息以及其他用户属性信息所涉及的用户在社交网络图的节点设置已知标签，包括：根据已知用户属性信息以及设定规则，确定已知用户属性信息所涉及的用户的初始标签并将初始标签分配给社交网络图对应的节点；根据所预测的其他用户属性以及设定规则，确定其他用户属性信息所涉及的用户的初始标签并进行衰减后分配给社交网络图对应的节点。

[0010] 在一些实施例中，上述对社交网络图执行标签传播算法，包括：迭代执行标签传播步骤，直至各个节点的标签收敛；其中，每次标签传播步骤包括：将每个节点的标签按照节点之间的边所对应的概率传递至相邻节点；将已知用户属性信息涉及的用户对应的节点重置为节点的初始标签，将其他用户属性信息涉及的用户对应的节点重置为对节点的初始标签进行衰减所得到的标签。

[0011] 第二方面，本申请实施例提供了一种用于生成用户属性信息的装置，该装置包括：获取单元，用于获取社交网络中用户的已知用户属性信息以及用户间关联关系信息，用户间关联关系信息用于表征不同用户之间是否存在关联关系；确定单元，用于针对社交网络中的各个用户，根据与当前用户存在关联关系的其他用户确定当前用户在待训练的主题模型中所对应社交关系文档中的词语，形成社交关系文档；主题设置单元，用于根据已知用户属性信息，设置已知用户属性信息所涉及的用户对应的社交关系文档中的已知主题；训练单元，用于使用主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练主题模型；预测单元，用于基于训练后的主题模型中的文档与主题矩阵和/或主题与词语矩阵，预测除已知用户属性信息外的其他用户属性信息。

[0012] 在一些实施例中，上述用户间关联关系信息还用于表征用户之间存在的关联关系的关联程度。

[0013] 在一些实施例中，获取单元进一步用于：获取社交网络中用户的操作行为信息；根据所获取的操作行为信息，生成用户间关联关系信息。

[0014] 在一些实施例中，上述装置还包括：构造单元，用于构造以社交网络中的各个用户为节点、关联程度作为边的社交网络图；标签设置单元，用于根据已知用户属性信息以及所预测的其他用户属性信息，对已知用户属性信息以及其他用户属性信息所涉及的用户在社交网络图的节点设置已知标签；执行单元，用于对社交网络图执行标签传播算法；生成单元，用于根据执行标签传播算法后的社交网络图的节点中概率最大的标签，生成节点所对应的用户的用户属性信息。

[0015] 在一些实施例中，标签设置单元进一步用于：根据已知用户属性信息以及设定规则，确定已知用户属性信息所涉及的用户的初始标签并将初始标签分配给社交网络图对应的节点；根据所预测的其他用户属性以及设定规则，确定其他用户属性信息所涉及的用户的初始标签并进行衰减后分配给社交网络图对应的节点。

[0016] 在一些实施例中，执行单元进一步用于：迭代执行标签传播步骤，直至各个节点的标签收敛；其中，每次标签传播步骤包括：将每个节点的标签按照节点之间的边所对应的概率传递至相邻节点；将已知用户属性信息涉及的用户对应的节点重置为节点的初始标签，

将其他用户属性信息涉及的用户对应的节点重置为对节点的初始标签进行衰减所得到的标签。

[0017] 第三方面，本申请提供一种设备，包括：一个或多个处理器；存储装置，用于存储一个或多个程序，当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如第一方面中任一项所描述的方法。

[0018] 第四方面，本申请提供一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现如第一方面中任一项所描述的方法。

[0019] 本申请实施例提供的用于生成用户属性信息的方法和装置，将部分用户属性信息以及社交网络中用户的关联关系作为已知条件，并基于主题模型预测出其他的用户属性信息，实现了在仅有部分用户属性信息时对其他用户属性信息的预测。

附图说明

[0020] 通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更明显：

[0021] 图1是本申请可以应用于其中的示例性系统架构图；

[0022] 图2是根据本申请的用于生成用户属性信息的方法的一个实施例的流程图；

[0023] 图3是根据本申请的用于生成用户属性信息的方法的又一个实施例的流程图；

[0024] 图4是根据本申请的用于生成用户属性信息的装置的一个实施例的结构示意图；

[0025] 图5是适于用来实现本申请实施例的设备的计算机系统的结构示意图。

具体实施方式

[0026] 下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

[0027] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0028] 图1示出了可以应用本申请的用于生成用户属性信息的方法或用于生成用户属性信息的装置的实施例的示例性系统架构100。

[0029] 如图1所示，系统架构100可以包括终端设备101、102、103，网络104、106和服务器105、107、108。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质，网络106则用以在服务器105和服务器107、108之间提供通信链路的介质。网络104、网络106可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

[0030] 用户可以使用终端设备101、102、103通过网络104与服务器105交互，以接收或发送消息等。终端设备101、102、103上可以安装有各种通讯客户端应用，例如网页浏览器应用、即时通信工具、邮箱客户端、社交平台软件等。

[0031] 终端设备101、102、103可以是各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、MP3播放器(Moving Picture Experts Group Audio Layer III，动态影像专家压缩标准音频层面3)、MP4(Moving Picture Experts Group Audio Layer IV，动态影像专家压缩标准音频层面4)播放器、膝上型便携计算机和台式计算机等等。

[0032] 服务器107、服务器108可以是各种提供数据支持的服务器，例如社交网站服务器，其提供的数据可以供服务器105进行进一步分析。

[0033] 服务器105可以是执行数据分析的服务器，例如对从终端设备101、102、103和/或服务器107、服务器108获取的原始数据进行数据分析，以生成用户画像数据。

[0034] 需要说明的是，本申请实施例所提供的用于生成用户属性信息的方法一般由服务器105执行，有些情况下也可以由终端设备101、102、103执行；相应地，用于生成用户属性信息的装置一般设置于服务器105中，有些情况下也可以设置在终端设备101、102、103中。还需要说明的是，某些情况下系统架构100也可以不包括终端设备101、102、103或服务器107、108，某些情况下服务器107、108可以与服务器105是同一设备，还有些情况下

[0035] 应该理解，图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

[0036] 继续参考图2，示出了根据本申请的用于生成用户属性信息的方法的一个实施例的流程200。该用于生成用户属性信息的方法，包括以下步骤：

[0037] 步骤201，获取社交网络中用户的已知用户属性信息以及用户间关联关系信息。

[0038] 在本实施例中，用于生成用户属性信息的方法运行于其上的电子设备(例如图1所示的服务器105)可以本地或远程获取社交网络中用户的已知用户属性信息以及用户间关联关系信息。社交网络可以包括但不限于微博、论坛、即时通讯等。

[0039] 其中，用户属性信息是用于表征用户特征的信息，已知用户属性信息是可以通过一定方法预先知晓的用户属性信息，通常已知用户属性信息仅为完整的用户属性信息中的部分信息。例如，当完整的用户属性信息需要包括年龄、性别、居住地、工作地点，已知用户信息所涉及的某一个用户可能只包括年龄、性别等信息，而缺乏居住地、工作地点等方面的信息。

[0040] 实践中，当已知用户属性信息或用户间关联关系信息预先存储在电子设备本地时，可以直接从本地获取。当这些信息未预先存储在本地时，电子设备则可以通过有线连接方式或者无线连接方式从远程的终端设备或服务器获取这些信息。例如，电子设备可以从社交网络服务器中获取用户间关联关系信息。用户间关联关系信息可以包括但不限于用户之间关注的信息、用户之间互相转发所发布内容的信息、用户之间存在好友关系的信息、用户之间处于同一个社群的信息。对方信息用户之间再例如，电子设备还可以从社交网站服务器中获取用户已填写的部分用户属性信息作为已知属性信息，也可以从社交网站中的用户所使用的终端获取已知属性信息。实践中，电子设备可以直接获取到已知用户属性信息或用户间关联关系信息，也可以先获取其他信息再进行分析后得到已知用户属性信息或用户间关联关系信息。

[0041] 步骤202，针对社交网络中的各个用户，根据与当前用户存在关联关系的其他用户确定当前用户在待训练的主题模型中所对应社交关系文档中的词语，形成社交关系文档。

[0042] 在本实施例中，基于步骤201中得到的用户间关联关系信息，上述电子设备(例如图1所示的服务器)可以为社交网络中的各个用户生成社交关系文档。需要说明的是，步骤202针对的用户，可以不是所有用户，而可以是仅对其中的一部分用户进行分析。实践中，可以对社交网络中的用户信息进行预处理，将某些异常的用户剔除出分析对象。例如，可以根据一定的标准计算用户的特别度，从而可以将特别度较高或较低的用户剔除出去。这些去

除用户的方法仅为示例，实践中不限于所列出的方法。

[0043] 在生成社交关系文档时，每个社交关系文档中的词语是根据与对应的用户有关联关系的其他用户确定的。实践中，词语可以使用与当前用户具有关联关系的用户的用户标识。用户标识包括但不限于真实姓名、网名、身份标识符或其他代号。该社交关系文档可以作为待训练的主题模型中的输入数据。例如，当关联关系采用好友关系时，某个用户对应的社交关系文档中可以用w_i表示用户的第i个好友，则好友的数目即是文档中的单词总数，最终所有用户的社交关系文档形成的社交关系文档集合中文档的数目便为社交网络中用户的数目。在主题模型中，假定两个用户之所以会产生关联关系，是因为他们在某些用户属性上的相似性，因此可以在主题模型中的主题去表征这些会导致关联关系的用户属性。例如，可以用用户的社交关系文档中的第i个主题z_i表征导致当前用户和第i个有关联关系的其他用户产生关联关系的属性。

[0044] 步骤203，根据已知用户属性信息，设置已知用户属性信息所涉及的用户在主题模型所对应社交关系文档中的已知主题。

[0045] 在本实施例中，基于步骤201所获取的已知用户属性信息，电子设备可以确定已知用户属性信息所涉及的用户，并根据已知用户属性信息设置这些用户所对应的社交关系文档的已知主题。例如，一个用户已经有属性“纽约”，则该用户与其有关联关系的用户之间产生关联关系的主题就不会被设置为表征西雅图的主题。实践中，可以通过开关变量$\Lambda^{(d)}$和$\Lambda'^{(k)}$分别表征主题模型的文档与主题矩阵和主题与词语矩阵。开关变量可以使用向量表示。例如，如果$\Lambda^{(d)}$的某个维度为0，表示这个用户不可能被分配到对应的属性，同样得，如果$\Lambda'^{(k)}$的某个维度为0表示对应的属性不可能分配到某个用户。

[0046] 步骤204，使用主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练主题模型。

[0047] 在本实施例中，由于通过步骤203和步骤204分别生成了主题模型的社交关系文档以及上述社交关系文档，电子设备可以对该主题模型进行训练。主题模型的训练过程可以对不同属性之间的潜在关系进行分析，从而能从不同的维度解析用户之间关联关系的构成原因。例如，如果社交网络中A和B是好友关系，那么他们之所以建立关系可能是因为他们有相似的兴趣爱好属性，或者在同一个地方，又或者有相同的教育背景。可选的，在训练时，所采用的算法可以包括但不限于最大期望算法(EM，Expectation Maximization Algorithm)、吉布斯(Gibbs)采样算法中的任一项。

[0048] 步骤205，基于训练后的主题模型中的文档与主题矩阵和/或主题与词语矩阵，预测除已知用户属性信息外的其他用户属性信息。

[0049] 在本实施例中，基于步骤204中对主题模型的训练，电子设备可以得到训练后的主题模型中的文档以及主题矩阵和/或主题与词语矩阵。前者可以表征特定用户在各种用户属性中的分布，后者则可以表征特定用户属性在各个用户中的分布。因此，基于主题模型中的文档以及主题矩阵和主题与词语矩阵中的至少一项，即可预测出除上述已知用户属性信息外的其他用户属性信息。需要说明的是，所预测的其他用户属性信息可能覆盖到所有未知用户属性，也可能仅覆盖到一部分的未知用户属性。

[0050] 在本实施例的一些可选的实现方式中，上述用户间关联关系信息还用于表征用户之间存在的关联关系的关联程度。在该实现方式中，用户间关联关系信息还用于表征用户

之间存在的关联关系的关联程度可以同时表征不同用户之间是否存在关联关系以及当存在关联关系时关联关系的关联程度，其中，关联程度是对关联关系强弱的定量表示。

[0051] 在本实施例的一些可选实现方式中，步骤201可以具体包括：获取社交网络中不同用户的操作行为信息；根据操作行为信息之间的关联性，生成用户间关联关系信息。在该实现方式中，电子设备可以首先获取社交网络中不同用户的操作行为信息。例如，该操作行为信息对应的行为可以是发起添加好友请求的行为、同意或拒绝好友添加请求的行为、发布媒体信息的行为、转发媒体信息的行为、收藏媒体信息的行为、关注某个用户的行为等。在根据所获取的操作行为信息生成用户间关联关系信息时，可以根据单个用户的行为操作信息确定用户间关联关系信息。例如，当用户甲关注用户乙时，可以认为用户甲与用户乙之前存在关联关系，还可以在需要时基于该行为的类型或其他行为属性确定关联关系的关联程度。此外，还可以根据两个用户的行为操作信息之间的关联性确定用户间关联关系信息。例如，当用户丙发布一条媒体信息，而用户丁转发同一条媒体信息，可以认定用户丙和用户丁之间存在关联关系。关联关系的确定规则可以按照实际需求进行设置，例如可以在用户丁转发用户丁的行为操作的次数达到阈值才认定两者之间存在关联关系。再例如，还可以根据用户丁转发用户丁的行为操作的次数确定所存在关联关系的关联程度。

[0052] 本申请的上述实施例提供的方法可以将部分用户属性信息以及社交网络中用户的关联关系作为已知条件，并基于主题模型预测出其他的用户属性信息，实现了在仅有部分用户属性信息时对其他用户属性信息的预测。

[0053] 进一步参考图3，其示出了用于生成用户属性信息的方法的又一个实施例的流程300。该用于生成用户属性信息的方法的流程300，包括以下步骤：

[0054] 步骤301，获取社交网络中用户的已知用户属性信息以及用户间关联关系信息。

[0055] 在本实施例中，用户间关联关系信息除了用于表征不同用户之间是否存在关联关系外，还用于表征用户之间存在的关联关系的关联程度。

[0056] 步骤302，针对社交网络中的各个用户，根据与当前用户存在关联关系的其他用户确定当前用户在待训练的主题模型中所对应社交关系文档中的词语，形成社交关系文档。

[0057] 在本实施例中，步骤302的具体处理可以参考图2对应实施例中的步骤202，这里不再赘述。

[0058] 步骤303，根据已知用户属性信息，设置已知用户属性信息所涉及的用户对应的社交关系文档中的已知主题。

[0059] 在本实施例中，步骤303的具体处理可以参考图2对应实施例中过的步骤203，这里不再赘述。

[0060] 步骤304，使用主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练主题模型。

[0061] 在本实施例中，步骤304的具体处理可以参考图2对应实施例中的步骤204，这里不再赘述。

[0062] 步骤305，基于训练后的主题模型中的文档与主题矩阵和/或主题与词语矩阵，预测除已知用户属性信息外的其他用户属性信息。

[0063] 在本实施例中，步骤305的具体处理可以参考图2对应实施例中的步骤205，这里不再赘述。

[0064] 步骤306，构造以社交网络中的各个用户为节点、关联程度作为边的社交网络图。

[0065] 在本实施例中，电子设备可以构造社交网络图。其中，社交网络图中的各个节点与社交网络中的各个用户对应，社交网络图中节点之间的边则与用户间关联关系的关联程度对应。当用户之间无关联关系时，节点之间可以不存在连接的边；当用户之间存在关联关系时，则可以根据关联程度的值确定节点之间的边的长度。

[0066] 步骤307，根据已知用户属性信息以及所预测的其他用户属性信息，对已知用户属性信息以及其他用户属性信息所涉及的用户在社交网络图的节点设置已知标签。

[0067] 在本实施例中，基于步骤201所获取的已知用户属性信息以及基于步骤305所预测的其他用户属性信息，电子设备可以使用这些用户属性信息对相应用户在社交网络图的节点设置已知标签。例如，在根据已知用户信息或预测的其他用户属性信息确定用户存在某种用户属性时，可以将所涉及的节点对应于该用户属性的标签设置为1，否则设置为0。所列出的设置方案仅为示例，实践中不限于该方案。需要说明的是，设置已知标签所使用的已知用户属性信息以及所预测的其他用户属性信息可以是前述步骤中已知用户属性信息以及所预测的其他用户属性信息中的部分信息。例如，步骤305中所预测出的用户属性信息，可以根据一定标准选择出可信度比较强用于步骤307的操作，由于缺乏数据导致一些可信度比较差的用户属性信息不用于设置社交关系图中的已知标签，这些用户属性信息可以在后续步骤中生成。

[0068] 步骤308，对社交网络图执行标签传播算法。

[0069] 在本实施例中，由于基于步骤307和步骤307生成了社交网络图且已设置了社交网络的部分标签，电子设备可以对社交网络图执行标签传播算法，从而可以根据社交网络图中的边在节点之间传递标签。

[0070] 在前述步骤使用主题模型生成用户属性信息时，当两个用户之间存在的共同关联关系的用户较少时，主题模型很难准确的描述两个用户产生关联关系的形成原因。便签传播模型可以对前述生成的结果进行精调，因为标签传播算法仅仅考虑网络的结构信息，利用用户之间关联关系的关联程度作为相似度度量方法可以避免前述问题。例如，用欧式距离度量的时候，两个用户之间的转移概率并不会因为共同好友的数量有所改变。

[0071] 标签传播算法是一个迭代算法，每一次迭代的第一步是根据转移矩阵将已知标签的节点的标签转移到未知标签的节点，再将已知标签的节点的标签重置回传播前的值。重复每次迭代的步骤，直到标签收敛。

[0072] 步骤309，根据执行标签传播算法后的社交网络图的节点中概率最大的标签，生成节点所对应的用户的用户属性信息。

[0073] 在本实施例中，基于步骤308对社交网络图执行标签传播算法所得到的社交网络图，电子设备可以根据针对该社交网络图的节点，根据节点中概率最大的标签生成节点所对应的用户的用户属性信息。

[0074] 在本实施例的一些可选实现方式中，步骤307可以包括：根据已知用户属性信息以及设定规则，确定已知用户属性信息所涉及的用户的初始标签并将初始标签分配给社交网络图对应的节点；根据所预测的其他用户属性以及设定规则，确定其他用户属性信息所涉及的用户的初始标签并进行衰减后分配给社交网络图对应的节点。在该实现方式中，在对所预测的其他用户属性所涉及的用户对应的节点分配标签时，除了与对已知用户属性所涉

及的用户对应的节点分配标签执行相同的计算初始标签外，还需要对初始标签进行衰减。例如，根据已知用户属性所得到的初始标签形成的矩阵为Y，根据预测的其他用户属性所得到的初始标签形成的矩阵为∈时，其中矩阵Y和∈中标签为1代表用户具有某种属性。此时，还需要对∈乘以置信度1-σ(σ是大于0且小于1的数)以对初始标签进行一定程度的衰减。σ可以预先选择一定的合理值。可选，σ可以是在测试集中测试得到后并在验证集验证通过而确定。在该实现方式中，考虑到预测出的用户属性的准确程度不及直接获取到的用户属性，因此对预测的其他用户属性所对应的初始标签额外增加了衰减步骤，减低其在标签传播算法中的重要程度，从而使得通过执行标签算法的社交关系图能生成更准确的用户属性信息。

[0075] 在本实施例中，上述步骤309可以具体包括：迭代执行标签传播步骤，直至各个节点的标签收敛；其中，每次标签传播步骤包括：将每个节点的标签按照节点之间的边所对应的概率传递至相邻节点；将已知用户属性信息涉及的用户对应的节点重置为节点的初始标签，将其他用户属性信息涉及的用户对应的节点重置为对节点的初始标签进行衰减所得到的标签。在该实现方式中，在对已知标签的节点进行标签重置时，将预测出的用户属性对应的节点重置回前述衰减后的标签，从而保证计算结果的准确性。实践中，在使用该方法进行重置时，可以通过将每次执行标签传播操作后各标签的Y重置为σ×Y+(1-σ)∈。其中，1-σ为选取的置信度。

[0076] 从图3中可以看出，与图2对应的实施例相比，本实施例中的用于生成用户属性信息的方法的流程300突出了利用已知用户属性与预测出的用户属性确定社交关系图中节点的已知标签，并执行标签算法得到最终的用户属性，可见该实施例能够利用标签算法对用户属性进行进一步精调，从而能够更完整与更准确地生成用户属性信息。该实施例的用途包括但不限于为人才网站提供用户画像。

[0077] 进一步参考图4，作为对上述各图所示方法的实现，本申请提供了一种用于生成用户属性信息的装置的一个实施例，该装置实施例与图2所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

[0078] 如图4所示，本实施例的用于生成用户属性信息的装置400包括：获取单元401、确定单元402、主题设置单元403、训练单元404和预测单元405。其中，获取单元401用于获取社交网络中用户的已知用户属性信息以及用户间关联关系信息，用户间关联关系信息用于表征不同用户之间是否存在关联关系；确定单元402用于针对社交网络中的各个用户，根据与当前用户存在关联关系的其他用户确定当前用户在待训练的主题模型中所对应社交关系文档中的词语，形成社交关系文档；主题设置单元403用于根据已知用户属性信息，设置已知用户属性信息所涉及的用户对应的社交关系文档中的已知主题；训练单元404用于使用主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练主题模型；而预测单元405用于基于训练后的主题模型中的文档与主题矩阵和/或主题与词语矩阵，预测除已知用户属性信息外的其他用户属性信息。

[0079] 在本实施例中，获取单元401、确定单元402、主题设置单元403、训练单元404和预测单元405可以分别参考图2对应实施例中的步骤201、步骤202、步骤203、步骤204以及步骤205，这里不再赘述。

[0080] 在本实施例的一些可选实现方式中，上述用户间关联关系信息还用于表征用户之

间存在的关联关系的关联程度。该实现方式的具体处理可以参考图2对应实施例中相应的实现方式，这里不再赘述。

[0081] 在本实施例的一些可选实现方式中，获取单元401进一步用于：获取社交网络中用户的操作行为信息；根据所获取的操作行为信息，生成用户间关联关系信息。该实现方式的具体处理可以参考图2对应实施例中相应的实现方式，这里不再赘述。

[0082] 在本实施例的一些可选实现方式中，装置400还包括：构造单元(未示出)，用于构造以社交网络中的各个用户为节点、关联程度作为边的社交网络图；标签设置单元(未示出)，用于根据已知用户属性信息以及所预测的其他用户属性信息，对已知用户属性信息以及其他用户属性信息所涉及的用户在社交网络图的节点设置已知标签；执行单元(未示出)，用于对社交网络图执行标签传播算法；生成单元(未示出)，用于根据执行标签传播算法后的社交网络图的节点中概率最大的标签，生成节点所对应的用户的用户属性信息。该实现方式的具体处理可以参考图3对应实施例中相应的步骤，这里不再赘述。

[0083] 在本实施例的一些可选实现方式中，标签设置单元可以进一步用于：根据已知用户属性信息以及设定规则，确定已知用户属性信息所涉及的用户的初始标签并将初始标签分配给社交网络图对应的节点；根据所预测的其他用户属性以及设定规则，确定其他用户属性信息所涉及的用户的初始标签并进行衰减后分配给社交网络图对应的节点。该实现方式的具体处理可以参考图3对应实施例中相应的步骤，这里不再赘述。

[0084] 在本实施例的一些可选实现方式中，执行单元进一步用于：迭代执行标签传播步骤，直至各个节点的标签收敛；其中，每次标签传播步骤包括：将每个节点的标签按照节点之间的边所对应的概率传递至相邻节点；将已知用户属性信息涉及的用户对应的节点重置为节点的初始标签，将其他用户属性信息涉及的用户对应的节点重置为对节点的初始标签进行衰减所得到的标签。该实现方式的具体处理可以参考图3对应实施例中相应的步骤，这里不再赘述。

[0085] 本申请还提供了一种设备，该设备包括：一个或多个处理器；存储装置，用于存储一个或多个程序，当一个或多个程序被一个或多个处理器执行，使得一个或多个处理器实现图2或3对应的实施例或实施例中任一可选实现方式所描述的方法。其中，该设备可以是终端设备或服务器。图5示出了适于用来实现本申请实施例的设备的计算机系统500的结构示意图。图5示出的设备仅仅是一个示例，不应对本申请实施例的功能和使用范围带来任何限制。

[0086] 如图5所示，计算机系统500包括中央处理单元(CPU)501，其可以根据存储在只读存储器(ROM)502中的程序或者从存储部分508加载到随机访问存储器(RAM)503中的程序而执行各种适当的动作和处理。在RAM 503中，还存储有系统500操作所需的各种程序和数据。CPU 501、ROM 502以及RAM 503通过总线504彼此相连。输入/输出(I/O)接口505也连接至总线504。

[0087] 以下部件连接至I/O接口505：包括键盘、鼠标等的输入部分506；包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分507；包括硬盘等的存储部分508；以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分509。通信部分509经由诸如因特网的网络执行通信处理。驱动器510也根据需要连接至I/O接口505。可拆卸介质511，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器510上，以便于从其上读出

的计算机程序根据需要被安装入存储部分508。

[0088] 特别地，根据本公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分509从网络上被下载和安装，和/或从可拆卸介质511被安装。在该计算机程序被中央处理单元(CPU)501执行时，执行本申请的方法中限定的上述功能。需要说明的是，本申请所描述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机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。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本申请中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本申请中，计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：无线、电线、光缆、RF等等，或者上述的任意合适的组合。

[0089] 附图中的流程图和框图，图示了按照本申请各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0090] 描述于本申请实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元也可以设置在处理器中，例如，可以描述为：一种处理器包括获取单元、确定单元、主题设置单元、训练单元和预测单元。其中，这些单元的名称在某种情况下并不构成对该单元本身的限定，例如，获取单元还可以被描述为“获取社交网络中用户的已知用户属性信息以及用户间关联关系信息的单元”。

[0091] 作为另一方面，本申请还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的设备中所包含的；也可以是单独存在，而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该设备执行时，使得该设备：获取社交网络中用户的已知用户属性信息以及用户间关联关系信息，所述用户间关

联关系信息用于表征不同用户之间是否存在关联关系;针对所述社交网络中的各个用户,根据与当前用户存在关联关系的其他用户确定所述当前用户在待训练的主题模型中所对应社交关系文档中的词语,形成社交关系文档;根据所述已知用户属性信息,设置所述已知用户属性信息所涉及的用户对应的社交关系文档中的已知主题;使用所述主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练所述主题模型;基于训练后的主题模型中的文档与主题矩阵和/或主题与词语矩阵,预测除所述已知用户属性信息外的其他用户属性信息。

[0092] 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解,本申请中所涉及的发明范围,并不限于上述技术特征的特定组合而成的技术方案,同时也应涵盖在不脱离上述发明构思的情况下,由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

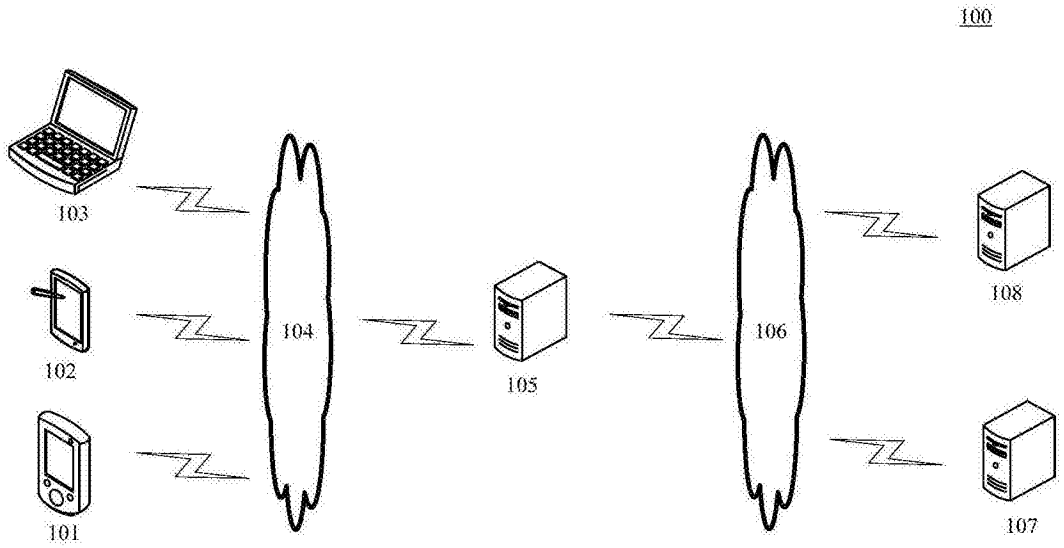

100
103
102
101
104
105
106
108
107

图1

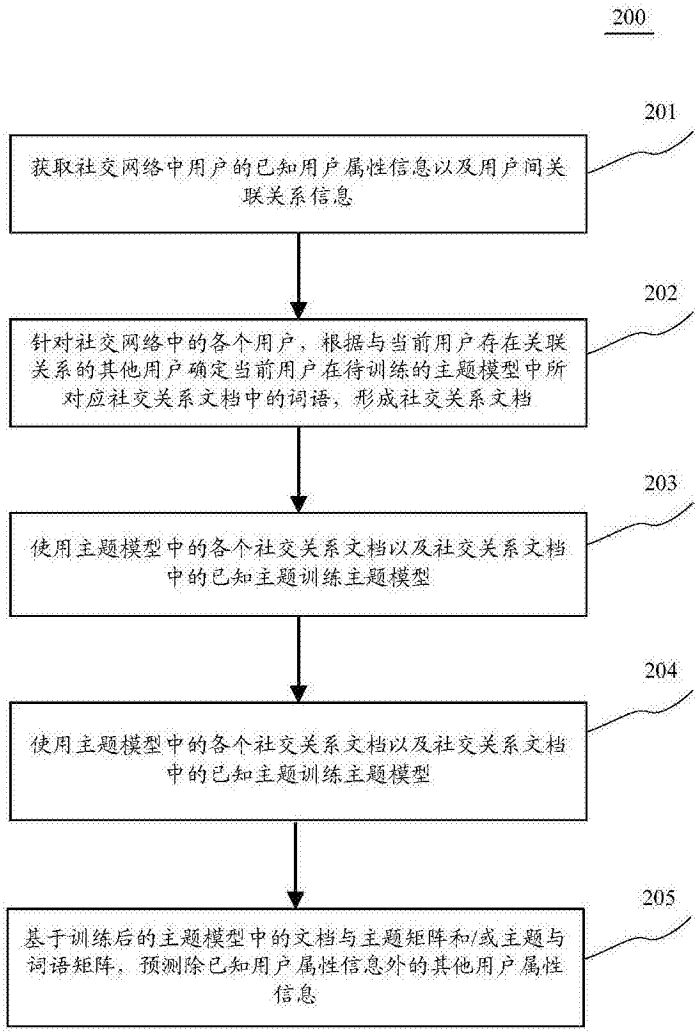

200
201
获取社交网络中用户的已知用户属性信息以及用户间关联关系信息
202
针对社交网络中的各个用户，根据与当前用户存在关联关系的其他用户确定当前用户在待训练的主题模型中所对应社交关系文档中的词语，形成社交关系文档
203
使用主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练主题模型
204
使用主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练主题模型
205
基于训练后的主题模型中的文档与主题矩阵和/或主题与词语矩阵，预测除已知用户属性信息外的其他用户属性信息

图2

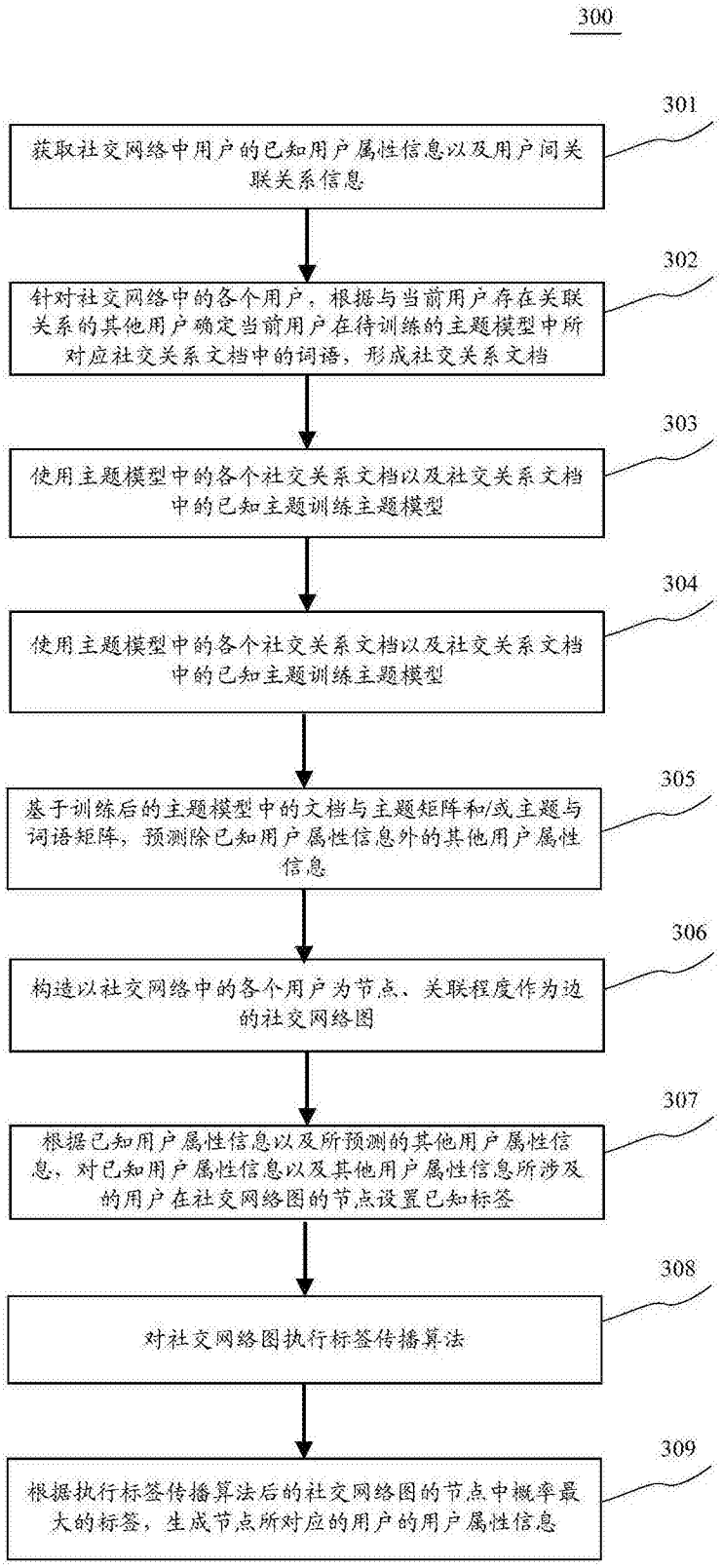

300
301
获取社交网络中用户的已知用户属性信息以及用户间关联关系信息
302
针对社交网络中的各个用户，根据与当前用户存在关联关系的其他用户确定当前用户在待训练的主题模型中所对应社交关系文档中的词语，形成社交关系文档
303
使用主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练主题模型
304
使用主题模型中的各个社交关系文档以及社交关系文档中的已知主题训练主题模型
305
基于训练后的主题模型中的文档与主题矩阵和/或主题与词语矩阵，预测除已知用户属性信息外的其他用户属性信息
306
构造以社交网络中的各个用户为节点、关联程度作为边的社交网络图
307
根据已知用户属性信息以及所预测的其他用户属性信息、对已知用户属性信息以及其他用户属性信息所涉及的用户在社交网络图的节点设置已知标签
308
对社交网络图执行标签传播算法
309
根据执行标签传播算法后的社交网络图的节点中概率最大的标签，生成节点所对应的用户的用户属性信息

图3

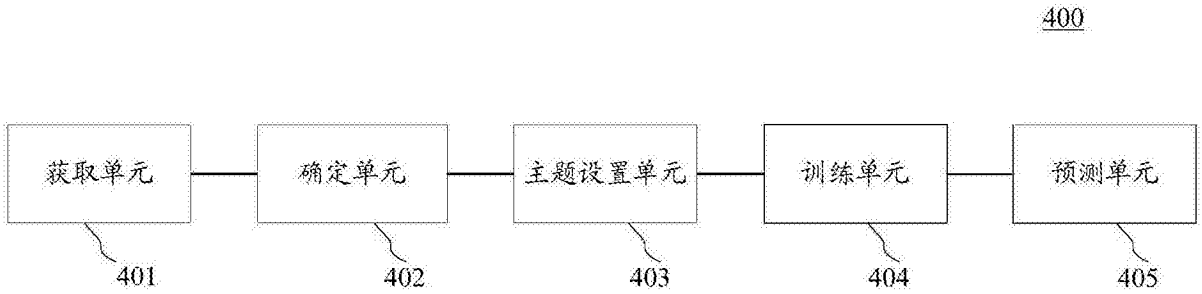

400
获取单元
确定单元
主题设置单元
训练单元
预测单元
401
402
403
404
405

图4

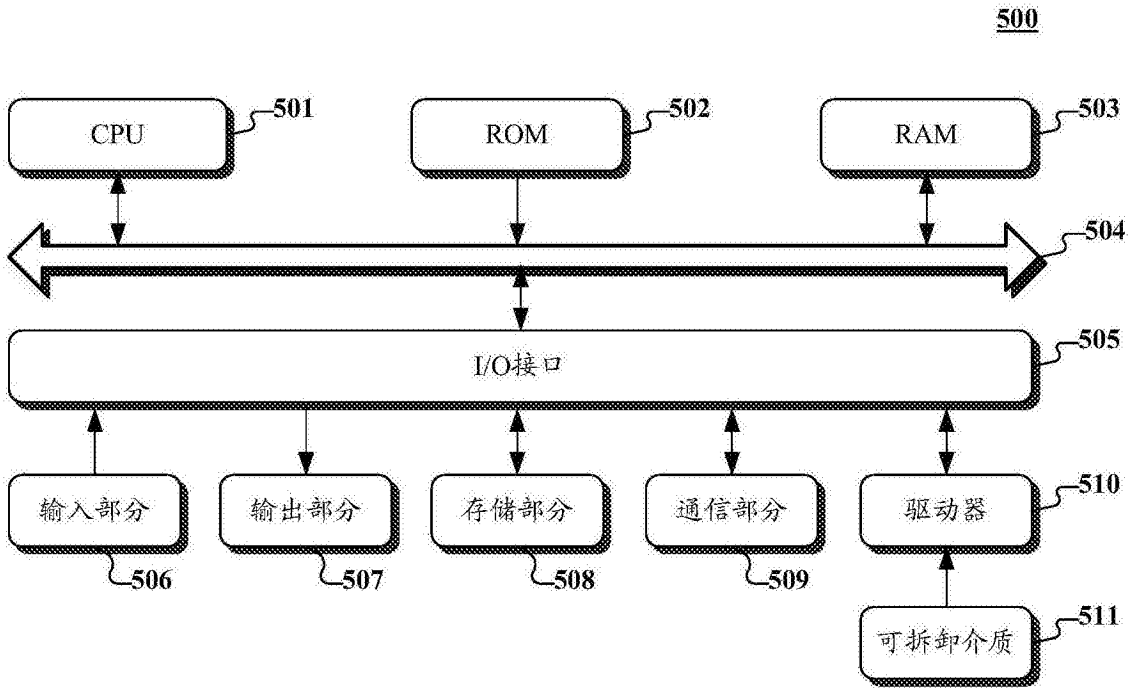

500
CPU
501
ROM
502
RAM
503
504
I/O接口
505
输入部分
506
输出部分
507
存储部分
508
通信部分
509
驱动器
510
可拆卸介质
511

图5