(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 111611388 A
(43)申请公布日 2020.09.01

(21)申请号 202010475295.3

(22)申请日 2020.05.29

(71)申请人 北京学之途网络科技有限公司
地址 100000 北京市海淀区北三环西路25号27号楼二层2020室

(72)发明人 冯允

(74)专利代理机构 北京超成律师事务所 11646
代理人 孔默

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/9536(2019.01)
G06F 40/289(2020.01)
G06N 3/04(2006.01)

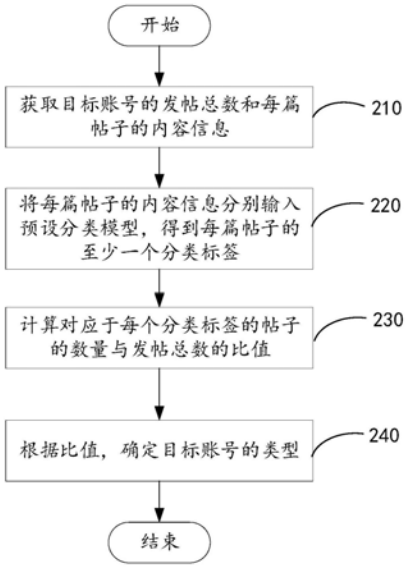
权利要求书2页 说明书5页 附图3页

(54)发明名称

账号分类方法、装置和设备

(57)摘要

本申请实施例提供一种账号分类方法、装置和设备，所述账号分类方法包括：获取目标账号的发帖总数和每篇帖子的内容信息；将所述每篇帖子的内容信息分别输入预设分类模型，得到所述每篇帖子的至少一个分类标签；计算对应于每个所述分类标签的帖子的数量与所述发帖总数的比值；根据所述比值，确定所述目标账号的类型。本申请实现了提高对账号进行分类的效率和准确性。



CN 111611388 A

1.一种账号分类方法,其特征在于,包括:

获取目标账号的发帖总数和每篇帖子的内容信息;

将所述每篇帖子的内容信息分别输入预设分类模型,得到所述每篇帖子的至少一个分类标签;

计算对应于每个所述分类标签的帖子的数量与所述发帖总数的比值;

根据所述比值,确定所述目标账号的类型。

2.根据权利要求1所述的方法,其特征在于,构建所述预设分类模型的步骤包括:

获取多个样本的内容信息和所述多个样本的分类标签;

对所述多个样本的内容信息进行特征提取,得到所述多个样本的特征数据;

根据所述多个样本的分类标签和所述特征数据,构建所述预设分类模型。

3.根据权利要求2所述的方法,其特征在于,所述对所述多个样本的内容信息进行特征提取,得到所述多个样本的特征数据,包括:

对所述多个样本的内容信息进行预处理,得到预处理数据;

对所述预处理数据进行向量化处理,得到所述多个样本的词向量数据;

对所述词向量数据进行卷积池化处理,得到所述多个样本的特征数据。

4.根据权利要求3所述的方法,其特征在于,所述对所述多个样本的内容信息进行预处理,得到预处理数据,包括:

对所述多个样本的内容信息进行分词处理;

去除分词处理后的所述多个样本的内容信息中的停用词。

5.根据权利要求1所述的方法,其特征在于,所述根据所述比值,确定所述目标账号的类型,包括:

判断所述比值是否大于预设阈值;

当所述比值大于所述预设阈值时,将对应于所述比值的所述分类标签确定为所述目标账号的类型。

6.一种账号分类装置,其特征在于,包括:

获取模块,用于获取目标账号的发帖总数和每篇帖子的内容信息;

分类模块,用于将所述每篇帖子的内容信息分别输入预设分类模型,得到所述每篇帖子的至少一个分类标签;

计算模块,用于计算对应于每个所述分类标签的帖子的数量与所述发帖总数的比值;

确定模块,用于根据所述比值,确定所述目标账号的类型。

7.根据权利要求6所述的装置,其特征在于,还包括构建模块,用于:

获取多个样本的内容信息和所述多个样本的分类标签;

对所述多个样本的内容信息进行特征提取,得到所述多个样本的特征数据;

根据所述多个样本的分类标签和所述特征数据,构建所述预设分类模型。

8.根据权利要求7所述的装置,其特征在于,所述构建模块用于:

对所述多个样本的内容信息进行预处理,得到预处理数据;

对所述预处理数据进行向量化处理,得到所述多个样本的词向量数据;

对所述词向量数据进行卷积池化处理,得到所述多个样本的特征数据。

9.根据权利要求6所述的装置,其特征在于,所述确定模块用于:

判断所述比值是否大于预设阈值；

当所述比值大于所述预设阈值时，将对应于所述比值的所述分类标签确定为所述目标账号的类型。

10.一种电子设备，其特征在于，包括：

存储器，用以存储计算机程序；

处理器，用以执行如权利要求1至5中任一项所述的方法。

账号分类方法、装置和设备

技术领域

[0001] 本申请涉及计算机技术领域，具体而言，涉及一种账号分类方法、装置和设备。

背景技术

[0002] 随着互联网时代的到来，社交网络日益成为人们生活中重要的组成部分，人们可以使用社交账号在社交网络上发表意见，分享自己的想法，在这个过程中就会出现一些知名度高，号召力强的用户，这些用户被称为KOL(Key Opinion Leader,关键意见领袖)，KOL的社交账号大多拥有较多的粉丝，因而，KOL在社交账号上发表的信息受到的关注度较高，具有一定的影响力。

[0003] 在挑选KOL时，经常需要对KOL进行分类，现有技术中，一般是根据用户自己填写的简介描述或认证信息进行分类，准确性较低，而且在用户没有填写与账号类别相关的信息时，无法对社交账号进行分类。

发明内容

[0004] 本申请实施例的目的在于提供一种账号分类方法、装置和设备，用以实现提高对账号进行分类的效率和准确性。

[0005] 本申请实施例第一方面提供了一种账号分类方法，包括：获取目标账号的发帖总数和每篇帖子的内容信息；将所述每篇帖子的内容信息分别输入预设分类模型，得到所述每篇帖子的至少一个分类标签；计算对应于每个所述分类标签的帖子的数量与所述发帖总数的比值；根据所述比值，确定所述目标账号的类型。

[0006] 于一实施例中，构建所述预设分类模型的步骤包括：获取多个样本的内容信息和所述多个样本的分类标签；对所述多个样本的内容信息进行特征提取，得到所述多个样本的特征数据；根据所述多个样本的分类标签和所述特征数据，构建所述预设分类模型。

[0007] 于一实施例中，所述对所述多个样本的内容信息进行特征提取，得到所述多个样本的特征数据，包括：对所述多个样本的内容信息进行预处理，得到预处理数据；对所述预处理数据进行向量化处理，得到所述多个样本的词向量数据；对所述词向量数据进行卷积池化处理，得到所述多个样本的特征数据。

[0008] 于一实施例中，所述对所述多个样本的内容信息进行预处理，得到预处理数据，包括：对所述多个样本的内容信息进行分词处理；去除分词处理后的所述多个样本的内容信息中的停用词。

[0009] 于一实施例中，所述根据所述比值，确定所述目标账号的类型，包括：判断所述比值是否大于预设阈值；当所述比值大于所述预设阈值时，将对应于所述比值的所述分类标签确定为所述目标账号的类型。

[0010] 本申请实施例第二方面提供了一种账号分类装置，包括：获取模块，用于获取目标账号的发帖总数和每篇帖子的内容信息；分类模块，用于将所述每篇帖子的内容信息分别输入预设分类模型，得到所述每篇帖子的至少一个分类标签；计算模块，用于计算对应于每

个所述分类标签的帖子的数量与所述发帖总数的比值;确定模块,用于根据所述比值,确定所述目标账号的类型。

[0011] 于一实施例中,还包括构建模块,用于:获取多个样本的内容信息和所述多个样本的分类标签;对所述多个样本的内容信息进行特征提取,得到所述多个样本的特征数据;根据所述多个样本的分类标签和所述特征数据,构建所述预设分类模型。

[0012] 于一实施例中,所述构建模块具体用于:对所述多个样本的内容信息进行预处理,得到预处理数据;对所述预处理数据进行向量化处理,得到所述多个样本的词向量数据;对所述词向量数据进行卷积池化处理,得到所述多个样本的特征数据。

[0013] 于一实施例中,所述构建模块具体用于:对所述多个样本的内容信息进行分词处理;去除分词处理后的所述多个样本的内容信息中的停用词。

[0014] 于一实施例中,所述确定模块用于:判断所述比值是否大于预设阈值;当所述比值大于所述预设阈值时,将对应于所述比值的所述分类标签确定为所述目标账号的类型。

[0015] 本申请实施例第三方面提供了一种电子设备,包括:存储器,用以存储计算机程序;处理器,用以执行本申请实施例第一方面及其任一实施例的方法。

[0016] 本申请实施例第四方面提供了一种非暂态电子设备可读存储介质,包括:程序,当其藉由电子设备运行时,使得所述电子设备执行本申请实施例第一方面及其任一实施例的方法。

[0017] 在本申请实施例中,获取待分类账号的历史发帖内容,通过预设分类模型对待分类账号的历史发帖内容进行分类,并根据各个类型的发帖量在发帖总数中的比例,最终确定待分类账户的类型,有效提高了对账号进行分类的效率和准确性。

附图说明

[0018] 为了更清楚地说明本申请实施例的技术方案,下面将对本申请实施例中所需要使用的附图作简单地介绍,应当理解,以下附图仅示出了本申请的某些实施例,因此不应被看作是对范围的限定,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,还可以根据这些附图获得其他相关的附图。

[0019] 图1为本申请一实施例的电子设备的结构示意图;

[0020] 图2为本申请一实施例的账号分类方法的流程示意图;

[0021] 图3为本申请一实施例中构建预设分类模型的流程示意图;

[0022] 图4为本申请一实施例的账号分类装置的结构示意图;

[0023] 图5为本申请另一实施例的账号分类装置的结构示意图。

[0024] 附图标记:

[0025] 100-电子设备,110-总线,120-处理器,130-存储器,400-账号分类装置,410-获取模块,420-分类模块,430-计算模块,440-确定模块,450-构建模块。

具体实施方式

[0026] 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描述。

[0027] 在本申请的描述中,术语“第一”、“第二”等仅用于区分描述,并不表示排列序号,也不能理解为指示或暗示相对重要性。

[0028] 在本申请的描述中，除非另有明确的规定和限定，术语“安装”、“设置”、“设有”、“连接”、“配置为”应做广义理解。例如，可以是固定连接，也可以是可拆卸连接，或整体式构造；可以是机械连接，也可以是电连接；可以是直接相连，也可以是通过中间媒介间接相连，又或者是两个装置、元件或组成部分之间内部的连通。对于本领域的普通技术人员而言，可以根据具体情况理解上述术语在本申请中的具体含义。

[0029] 请参看图1，其为本申请一实施例的电子设备100的结构示意图，包括至少一个处理器120和存储器130，图1中以一个处理器为例。处理器120和存储器130通过总线110连接，存储器130存储有可被至少一个处理器120执行的指令，指令被至少一个处理器120执行，以使至少一个处理器120执行如下述实施例中的账号分类方法。

[0030] 如图2所示，其为本申请一实施例的账号分类方法的流程示意图，该方法可由图1所示的电子设备100来执行，以实现提高对账号进行分类的效率和准确性。该方法包括如下步骤：

[0031] 步骤210：获取目标账号的发帖总数和每篇帖子的内容信息。

[0032] 在上述步骤中，目标账号可以是在各类社交平台上注册的账号，包括但不限于：微博账号、公众号、贴吧账号等。可以采用API(Application Programming Interface，应用程序接口)、网络爬虫等方法来获取目标账号的发帖总数和每篇帖子的内容信息，该内容信息可以包括文本信息、图片、音频、视频等。

[0033] 步骤220：将每篇帖子的内容信息分别输入预设分类模型，得到每篇帖子的至少一个分类标签。

[0034] 在上述步骤中，可以通过机器学习的方式预先构建预设分类模型，然后将每篇帖子的内容信息分别输入预设分类模型，预设分类模型根据每篇帖子的内容信息的特征，输出对应的分类标签，其中，每篇帖子可以有一个或多个分类标签，该分类标签包括但不限于：旅游、美食、摄影、音乐、运动等。

[0035] 步骤230：计算对应于每个分类标签的帖子的数量与发帖总数的比值。

[0036] 在上述步骤中，假设目标账号的发帖总数为X，对应于第一分类标签的帖子的数量为x_1，对应于第二分类标签的帖子的数量为x_2，对应于第三分类标签的帖子的数量为x_3，对应于第四分类标签的帖子的数量为x_4，则计算对应于每个分类标签的帖子的数量与发帖总数的比值，第一分类标签的比值为(x_1/X)，第二分类标签的比值为(x_2/X)，第三分类标签的比值为(x_3/X)，第四分类标签的比值为(x_4/X)，以此类推。

[0037] 步骤240：根据比值，确定目标账号的类型。

[0038] 在上述步骤中，可以判断分类标签的比值是否大于预设阈值，当比值大于预设阈值时，将对应于该比值的分类标签确定为目标账号的类型。目标账号的类型可以是一个或多个，所有比值大于预设阈值的分类标签都可以确定为目标账号的类型。该预设阈值可以在0至1的范围内根据实际情况设定，于一实施例中，预设阈值为0.5，将比值大于0.5的分类标签确定为目标账号的类型。

[0039] 于一实施例中，可以根据不同分类标签的浏览量、阅读量或转发量等，针对每个分类标签设定不同的预设阈值，判断分类标签的比值是否大于对应于该分类标签的预设阈值，当分类标签的比值大于对应于该分类标签的预设阈值时，将该分类标签确定为目标账号的类型。

[0040] 于一实施例中，可以对分类标签的比值按大小进行排序，将比值最大的分类标签确定为目标账号的类型。

[0041] 如图3所示，其为本申请一实施例中构建预设分类模型的流程示意图，该方法可由图1所示的电子设备100来执行，该方法包括如下步骤：

[0042] 步骤310：获取多个样本的内容信息和多个样本的分类标签。

[0043] 在上述步骤中，样本的分类标签采用人工标记的方式获取，手动查看每个样本的内容信息，判定每个样本的分类标签并标记。

[0044] 步骤320：对多个样本的内容信息进行特征提取，得到多个样本的特征数据。

[0045] 在上述步骤中，对多个样本的内容信息进行预处理，得到预处理数据，对预处理数据进行向量化处理，得到多个样本的词向量数据，对词向量数据进行卷积池化处理，得到多个样本的特征数据。

[0046] 于一实施例中，对多个样本的内容信息进行预处理包括：对多个样本的内容信息进行分词处理，去除分词处理后的多个样本的内容信息中的停用词。分词处理就是将内容信息进行文本划分，形成一系列字词序列，去除停用词就是去除一些无实际意义的词，例如标点、数字、符号、“和”、“的”等。

[0047] 步骤330：根据多个样本的分类标签和特征数据，构建预设分类模型。

[0048] 于一实施例中，可以采用卷积神经网络对词向量数据进行卷积池化处理，得到多个样本的特征数据。卷积神经网络的输入为词向量矩阵，在卷积层使用多个不同尺寸的卷积核对词向量矩阵进行卷积计算，得到多个特征矩阵，然后在池化层使用最大池化对多个特征矩阵进行采样，保留值最大的特征矩阵，舍弃其他值较小的特征矩阵，得到特征数据。最后将特征数据输入至全连接层进行分类，并将分类结果与多个样本的分类标签进行对比，进行优化调整，直至分类结果与多个样本的分类标签一致，则得到训练后的预设分类模型。

[0049] 于一实施例中，预设分类模型的构建可以基于fasttext模型，将词向量数据直接相加求平均值，得到文本向量，再经过一个输出层得到分类结果，并将分类结果与多个样本的分类标签进行对比，进行优化调整，直至分类结果与多个样本的分类标签一致，则得到训练后的预设分类模型。

[0050] 如图4所示，其为本申请一实施例的账号分类装置400的结构示意图，该装置可应用于图1所示的电子设备100，包括：获取模块410、分类模块420、计算模块430和确定模块440。各个模块的原理关系如下：

[0051] 获取模块410，用于获取目标账号的发帖总数和每篇帖子的内容信息。详细内容参见上述实施例中步骤210的描述。

[0052] 分类模块420，用于将每篇帖子的内容信息分别输入预设分类模型，得到每篇帖子的至少一个分类标签。详细内容参见上述实施例中步骤220的描述。

[0053] 计算模块430，用于计算对应于每个分类标签的帖子的数量与发帖总数的比值。详细内容参见上述实施例中步骤230的描述。

[0054] 确定模块440，用于根据比值，确定目标账号的类型。详细内容参见上述实施例中步骤240的描述。

[0055] 于一实施例中，确定模块440用于：判断比值是否大于预设阈值；当比值大于预设

阈值时，将对应于比值的分类标签确定为目标账号的类型。详细内容参见上述实施例中步骤240的描述。

[0056] 如图5所示，其为本申请一实施例的账号分类装置400的结构示意图，该装置可应用于图1所示的电子设备100，包括：获取模块410、分类模块420、计算模块430、确定模块440和构建模块450。

[0057] 于一实施例中，构建模块450用于：获取多个样本的内容信息和多个样本的分类标签；对多个样本的内容信息进行特征提取，得到多个样本的特征数据；根据多个样本的分类标签和特征数据，构建预设分类模型。详细内容参见上述实施例中步骤310至步骤330的描述。

[0058] 于一实施例中，构建模块450具体用于：对多个样本的内容信息进行预处理，得到预处理数据；对预处理数据进行向量化处理，得到多个样本的词向量数据；对词向量数据进行卷积池化处理，得到多个样本的特征数据。详细内容参见上述实施例中步骤320的描述。

[0059] 于一实施例中，构建模块450具体用于：对多个样本的内容信息进行分词处理；去除分词处理后的多个样本的内容信息中的停用词。详细内容参见上述实施例中步骤320的描述。

[0060] 上述账号分类装置400的详细描述，请参见上述实施例中相关方法步骤的描述。

[0061] 本发明实施例还提供了一种电子设备可读存储介质，包括：程序，当其在电子设备上运行时，使得电子设备可执行上述实施例中方法的全部或部分流程。其中，存储介质可为磁盘、光盘、只读存储记忆体(Read-Only Memory，ROM)、随机存储记忆体(Random Access Memory，RAM)、快闪存储器(Flash Memory)、硬盘(Hard Disk Drive，缩写：HDD)或固态硬盘(Solid-State Drive，SSD)等。存储介质还可以包括上述种类的存储器的组合。

[0062] 以上仅为本申请的优选实施例而已，并不用于限制本申请。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

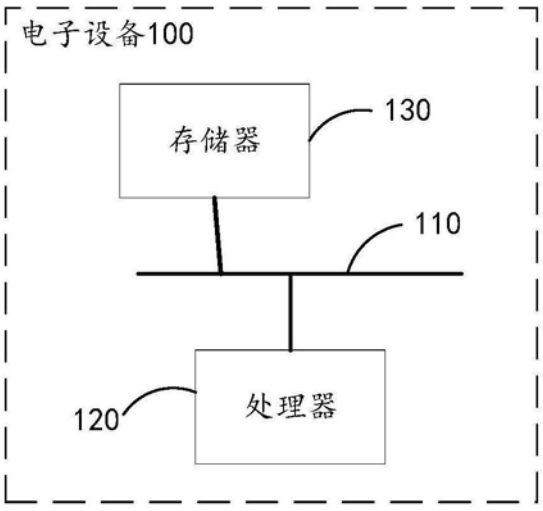

电子设备100
存储器
130
110
处理器
120

图1

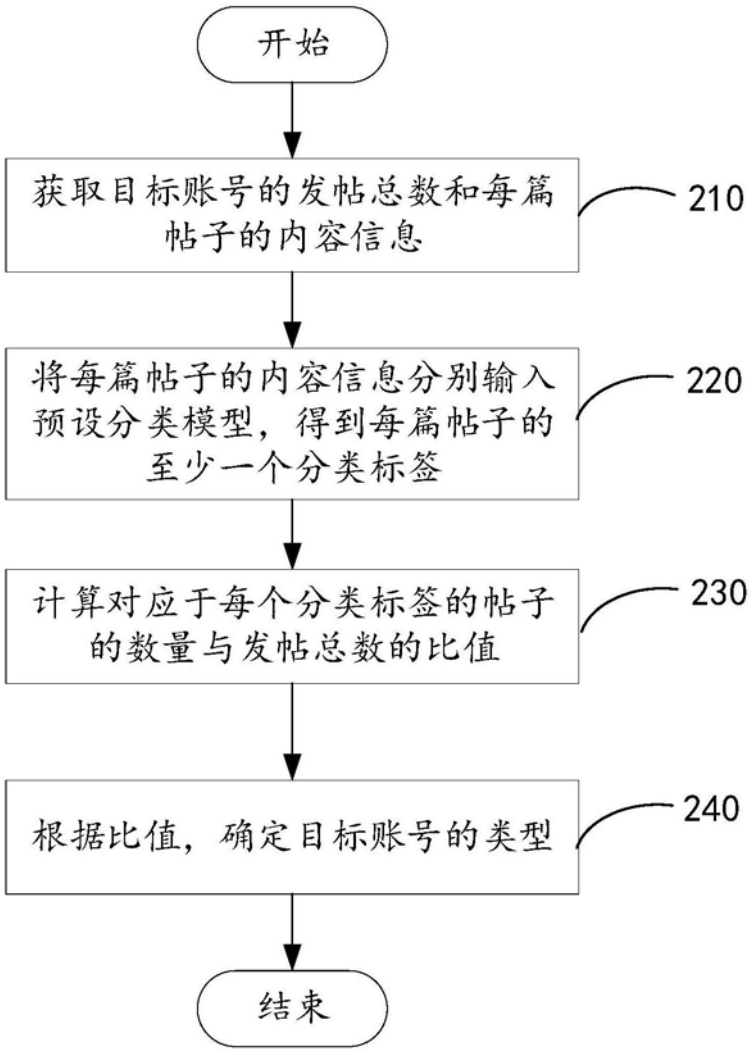

开始
获取目标账号的发帖总数和每篇帖子的内容信息
210
将每篇帖子的内容信息分别输入预设分类模型，得到每篇帖子的至少一个分类标签
220
计算对应于每个分类标签的帖子的数量与发帖总数的比值
230
根据比值，确定目标账号的类型
240
结束

图2

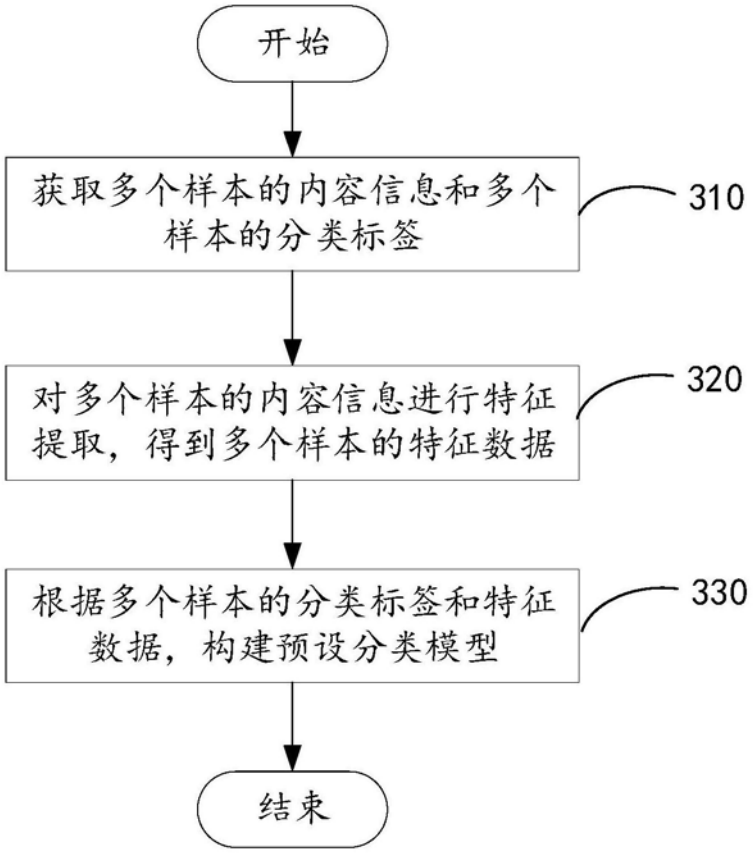

开始
获取多个样本的内容信息和多个样本的分类标签
310
对多个样本的内容信息进行特征提取，得到多个样本的特征数据
320
根据多个样本的分类标签和特征数据，构建预设分类模型
330
结束

图3

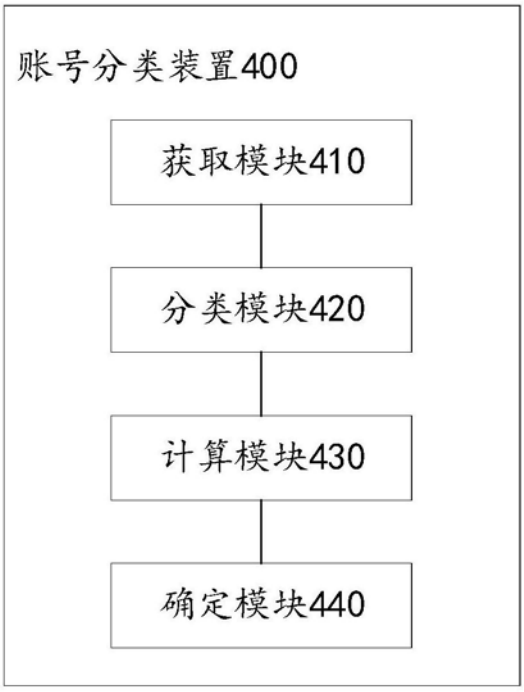

账号分类装置400
获取模块410
分类模块420
计算模块430
确定模块440

图4

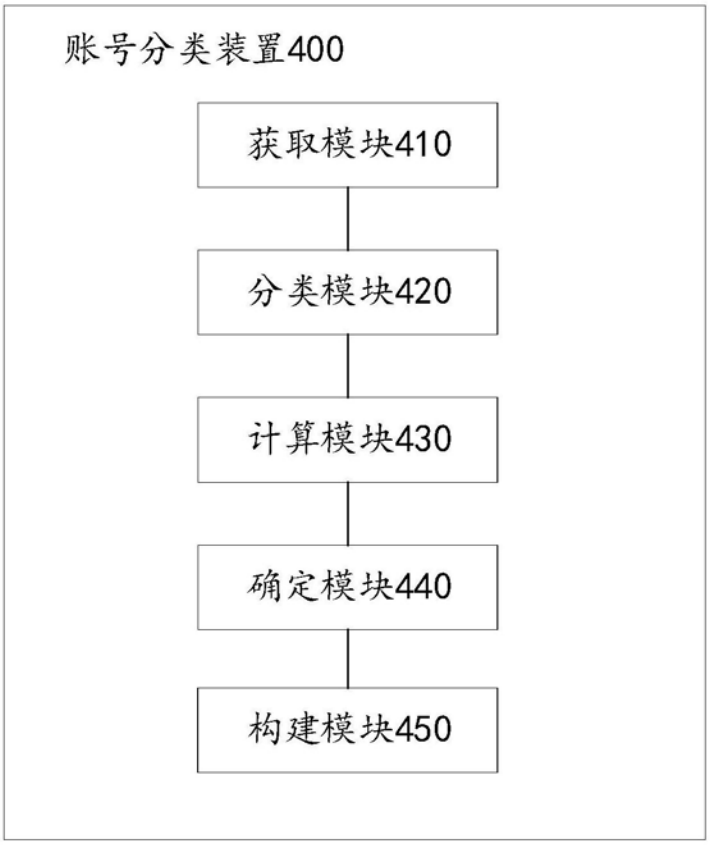


图5