(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 107301225 B
(45)授权公告日 2021.01.26

(21)申请号 201710469373.7

(22)申请日 2017.06.20

(65)同一申请的已公布的文献号
申请公布号 CN 107301225 A

(43)申请公布日 2017.10.27

(73)专利权人 挖财网络技术有限公司
地址 310012 浙江省杭州市西湖区华星路96号第18层

(72)发明人 尤志强 车曦 潘琪

(74)专利代理机构 北京博思佳知识产权代理有限公司 11415
代理人 林祥 李威

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/33(2019.01)

(56)对比文件
CN 105912716 A,2016.08.31
CN 104933074 A,2015.09.23
CN 104834747 A,2015.08.12
CN 105893354 A,2016.08.24
CN 106126605 A,2016.11.16
胡浩.海量短文本的主题挖掘及其可视化.《中国优秀硕士学位论文全文数据库信息科技辑》.2017,第2017卷(第03期),第I138-6032页.

审查员 侯昕煜

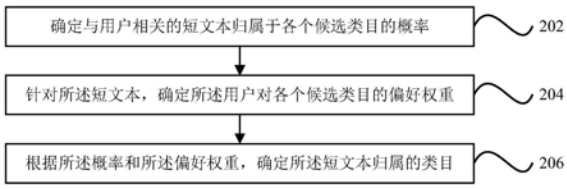
权利要求书3页 说明书15页 附图6页

(54)发明名称

短文本分类方法及装置

(57)摘要

本申请提供一种短文本分类方法及装置,该方法可以包括:确定与用户相关的短文本归属于各个候选类目的概率;针对所述短文本,确定所述用户对各个候选类目的偏好权重;根据所述概率和所述偏好权重,确定所述短文本归属的类目。通过本申请的技术方案,可以实现个性化的短文本分类操作,以提升对短文本的分类准确度。



CN 107301225 B

1.一种短文本分类方法,其特征在于,包括:

确定与用户相关的短文本归属于各个候选类目的概率,包括:通过多个类目召回模型分别计算所述短文本归属于各个候选类目的概率;所述类目召回模型由基于全量用户数据生成的训练样本进行训练得到;所述类目召回模型包括:第一类目召回模型,所述第一类目召回模型包括由所述全量用户数据对应的所有训练样本进行训练得到的卷积神经网络模型;第二类目召回模型,所述第二类目召回模型包括由所述全量用户数据中的小样本类目的用户数据对应的训练样本进行训练得到的长短期记忆网络模型或双向长短期记忆循环网络模型,其中所述小样本类目对应的用户数据的数据量小于预设数量;

针对所述短文本,确定所述用户对各个候选类目的偏好权重;

根据所述概率和所述偏好权重,确定所述短文本归属的类目。

2.根据权利要求1所述的方法,其特征在于,所述训练样本包括所述全量用户数据中的历史短文本转换得到的文本向量矩阵。

3.根据权利要求2所述的方法,其特征在于,所述文本向量矩阵由文本特征转换得到的文本向量组成,所述文本特征包括所述历史短文本被切分得到的分词。

4.根据权利要求3所述的方法,其特征在于,所述文本特征还包括:所述历史短文本归属的类目的信息。

5.根据权利要求3所述的方法,其特征在于,所述文本特征是在参照词向量集合和字向量集合的情况下,被转换为相应的文本向量;其中,所述词向量集合包括所述全量用户数据被切分得到的所有文本特征和对应的词向量之间的映射关系,所述字向量集合包括所述全量用户数据采用的文字的全量字和对应的字向量之间的映射关系。

6.根据权利要求2所述的方法,其特征在于,所述确定与用户相关的短文本归属于各个候选类目的概率,包括:

将所述短文本转换为相应的文本向量矩阵;

通过所述类目召回模型确定所述短文本对应的文本向量矩阵归属于各个候选类目的概率。

7.根据权利要求1所述的方法,其特征在于,所述卷积神经网络模型采用的激活函数包括:门控线性单元。

8.根据权利要求1所述的方法,其特征在于,所述卷积神经网络模型的池化(pooling)层与全连接(fully connected)层之间采用信息高速公路(highway)结构进行联通。

9.根据权利要求1所述的方法,其特征在于,所述针对所述短文本,确定所述用户对各个候选类目的偏好权重,包括:

通过对应于所述用户的类目偏好模型,确定所述用户对各个候选类目的偏好权重;其中,所述类目偏好模型被基于所述用户的个人用户数据而构建。

10.根据权利要求9所述的方法,其特征在于,所述类目偏好模型被基于朴素贝叶斯或贝叶斯网络而构建。

11.一种短文本分类装置,其特征在于,包括:

概率确定单元,确定与用户相关的短文本归属于各个候选类目的概率,包括:通过多个类目召回模型分别计算所述短文本归属于各个候选类目的概率;所述类目召回模型由基于全量用户数据生成的训练样本进行训练得到;所述类目召回模型包括:第一类目召回模型,

所述第一类目召回模型包括由所述全量用户数据对应的所有训练样本进行训练得到的卷积神经网络模型;第二类目召回模型,所述第二类目召回模型包括由所述全量用户数据中的小样本类目的用户数据对应的训练样本进行训练得到的长短期记忆网络模型或双向长短期记忆循环网络模型,其中所述小样本类目对应的用户数据的数据量小于预设数量;

权重确定单元,针对所述短文本,确定所述用户对各个候选类目的偏好权重;

类目确定单元,根据所述概率和所述偏好权重,确定所述短文本归属的类目。

12.根据权利要求11所述的装置,其特征在于,所述训练样本包括所述全量用户数据中的历史短文本转换得到的文本向量矩阵。

13.根据权利要求12所述的装置,其特征在于,所述文本向量矩阵由文本特征转换得到的文本向量组成,所述文本特征包括所述历史短文本被切分得到的分词。

14.根据权利要求13所述的装置,其特征在于,所述文本特征还包括:所述历史短文本归属的类目的信息。

15.根据权利要求13所述的装置,其特征在于,所述文本特征是在参照词向量集合和字向量集合的情况下,被转换为相应的文本向量;其中,所述词向量集合包括所述全量用户数据被切分得到的所有文本特征和对应的词向量之间的映射关系,所述字向量集合包括所述全量用户数据采用的文字的全量字和对应的字向量之间的映射关系。

16.根据权利要求12所述的装置,其特征在于,所述概率确定单元具体用于:

将所述短文本转换为相应的文本向量矩阵;

通过所述类目召回模型确定所述短文本对应的文本向量矩阵归属于各个候选类目的概率。

17.根据权利要求11所述的装置,其特征在于,所述卷积神经网络模型采用的激活函数包括:门控线性单元。

18.根据权利要求11所述的装置,其特征在于,所述卷积神经网络模型的池化(pooling)层与全连接(fully connected)层之间采用信息高速公路(highway)结构进行联通。

19.根据权利要求11所述的装置,其特征在于,所述权重确定单元具体用于:

通过对应于所述用户的类目偏好模型,确定所述用户对各个候选类目的偏好权重;其中,所述类目偏好模型被基于所述用户的个人用户数据而构建。

20.根据权利要求19所述的装置,其特征在于,所述类目偏好模型被基于朴素贝叶斯或贝叶斯网络而构建。

21.一种短文本分类装置,其特征在于,包括:

处理器;

用于存储处理器可执行指令的存储器;

其中,所述处理器被配置为实现如权利要求1-10中任一项所述的方法。

22.一种文本分类方法,其特征在于,包括:

确定与用户相关的文本归属于各个候选类目的概率,包括:通过多个类目召回模型分别计算所述文本归属于各个候选类目的概率;所述类目召回模型由基于全量用户数据生成的训练样本进行训练得到;所述类目召回模型包括:第一类目召回模型,所述第一类目召回模型包括由所述全量用户数据对应的所有训练样本进行训练得到的卷积神经网络模型;第

二类目召回模型，所述第二类目召回模型包括由所述全量用户数据中的小样本类目的用户数据对应的训练样本进行训练得到的长短期记忆网络模型或双向长短期记忆循环网络模型，其中所述小样本类目对应的用户数据的数据量小于预设数量；

针对所述文本，确定所述用户对各个候选类目的偏好权重；

根据所述概率和所述偏好权重，确定所述文本归属的类目。

23.一种文本分类装置，其特征在于，包括：

概率确定单元，确定与用户相关的文本归属于各个候选类目的概率，包括：通过多个类目召回模型分别计算所述文本归属于各个候选类目的概率；所述类目召回模型由基于全量用户数据生成的训练样本进行训练得到；所述类目召回模型包括：第一类目召回模型，所述第一类目召回模型包括由所述全量用户数据对应的所有训练样本进行训练得到的卷积神经网络模型；第二类目召回模型，所述第二类目召回模型包括由所述全量用户数据中的小样本类目的用户数据对应的训练样本进行训练得到的长短期记忆网络模型或双向长短期记忆循环网络模型，其中所述小样本类目对应的用户数据的数据量小于预设数量；

权重确定单元，针对所述文本，确定所述用户对各个候选类目的偏好权重；

类目确定单元，根据所述概率和所述偏好权重，确定所述文本归属的类目。

24.一种文本分类装置，其特征在于，包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器被配置为实现如权利要求22所述的方法。

短文本分类方法及装置

技术领域

[0001] 本申请涉及信息处理技术领域，尤其涉及一种短文本分类方法及装置。

背景技术

[0002] 在相关技术中，通过自然语言处理(NLP，Natural Language Processing)、计算语言学(CL，Computational Linguistics)等各类学科的发展，可以在一定程度上实现计算机对自然语言的理解和处理。

[0003] 然而，随着网络技术的不断发展，出现越来越多的短文本，由于短文本具有长度短、结构复杂以及变形词多等特点，导致计算机在缺乏上下文关联的情况下，越来越难以准确识别短文本的正确含义，从而无法实现对短文本的正确分类处理。

发明内容

[0004] 有鉴于此，本申请提供一种短文本分类方法及装置，可以实现个性化的短文本分类操作，以提升对短文本的分类准确度。

[0005] 为实现上述目的，本申请提供技术方案如下：

[0006] 根据本申请的第一方面，提出了一种短文本分类方法，包括：

[0007] 确定与用户相关的短文本归属于各个候选类目的概率；

[0008] 针对所述短文本，确定所述用户对各个候选类目的偏好权重；

[0009] 根据所述概率和所述偏好权重，确定所述短文本归属的类目。

[0010] 根据本申请的第二方面，提出了一种短文本分类装置，包括：

[0011] 概率确定单元，确定与用户相关的短文本归属于各个候选类目的概率；

[0012] 权重确定单元，针对所述短文本，确定所述用户对各个候选类目的偏好权重；

[0013] 类目确定单元，根据所述概率和所述偏好权重，确定所述短文本归属的类目。

[0014] 根据本申请的第三方面，提出了一种短文本分类装置，包括：

[0015] 处理器；

[0016] 用于存储处理器可执行指令的存储器；

[0017] 其中，所述处理器被配置为实现如第一方面所述的方法。

[0018] 根据本申请的第四方面，提出了一种文本分类方法，包括：

[0019] 确定与用户相关的文本归属于各个候选类目的概率；

[0020] 针对所述文本，确定所述用户对各个候选类目的偏好权重；

[0021] 根据所述概率和所述偏好权重，确定所述文本归属的类目。

[0022] 根据本申请的第五方面，提出了一种文本分类装置，包括：

[0023] 概率确定单元，确定与用户相关的文本归属于各个候选类目的概率；

[0024] 权重确定单元，针对所述文本，确定所述用户对各个候选类目的偏好权重；

[0025] 类目确定单元，根据所述概率和所述偏好权重，确定所述文本归属的类目。

[0026] 根据本申请的第六方面，提出了一种文本分类装置，包括：

[0027] 处理器；

[0028] 用于存储处理器可执行指令的存储器；

[0029] 其中，所述处理器被配置为实现如第四方面所述的方法。

[0030] 由以上技术方案可见，本申请在对与用户相关的短文本进行分类时，通过获取用户对各个候选类目的偏好权重，可以实现个性化的短文本分类操作，从而提升对短文本的分类准确度。

附图说明

[0031] 图1是本申请一示例性实施例提供的一种短文本分类系统的架构示意图。

[0032] 图2是本申请一示例性实施例提供的一种短文本分类方法的流程图。

[0033] 图3是本申请一示例性实施例提供的一种文本分类方法的流程图。

[0034] 图4-6是相关技术的一种记账应用的客户端界面的示意图。

[0035] 图7是本申请一示例性实施例提供的一种记账信息分类过程的示意图。

[0036] 图8是本申请一示例性实施例提供的一种训练类目召回模型的示意图。

[0037] 图9是本申请一示例性实施例提供的一种训练类目偏好模型的示意图。

[0038] 图10是本申请一示例性实施例提供的一种记账应用的记账界面的示意图。

[0039] 图11是本申请一示例性实施例提供的一种对记账信息进行智能分类的示意图。

[0040] 图12是本申请一示例性实施例提供的一种电子设备的结构示意图。

[0041] 图13是本申请一示例性实施例提供的一种短文本分类装置的框图。

[0042] 图14是本申请一示例性实施例提供的另一种电子设备的结构示意图。

[0043] 图15是本申请一示例性实施例提供的一种文本分类装置的框图。

具体实施方式

[0044] 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本申请相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的装置和方法的例子。

[0045] 本申请在对与用户相关的短文本进行分类时，通过获取用户对各个候选类目的偏好权重，可以实现个性化的短文本分类操作，从而提升对短文本的分类准确度。

[0046] 为对本申请进行进一步说明，提供下列实施例：

[0047] 图1是本申请一示例性实施例提供的一种短文本分类系统的架构示意图。如图1所示，该系统可以包括服务器11、网络12、若干电子设备，比如手机13、PC14等。

[0048] 服务器11可以为包含一独立主机的物理服务器，或者该服务器11可以为主机集群承载的虚拟服务器，或者该服务器11可以为云服务器。在运行过程中，服务器11可以运行某一应用的服务器侧的程序，以实现该应用的相关业务功能。

[0049] 手机13、PC14均为用户可以使用的一种类型的电子设备。实际上，用户显然还可以使用诸如下述类型的电子设备：平板设备、笔记本电脑、掌上电脑(PDAs，Personal Digital Assistants)、可穿戴设备(如智能眼镜、智能手表等)等，本申请并不对此进行限制。在运行过程中，该电子设备可以运行某一应用的客户端侧的程序，以实现该应用的相关业务功能。

[0050] 而对于手机13、PC14与服务器11之间进行交互的网络12，可以包括多种类型的有线或无线网络。在一实施例中，该网络12可以包括公共交换电话网络(Public Switched Telephone Network，PSTN)和因特网。

[0051] 因此，本申请的短文本分类方案可以应用于图1所示的实施例中，手机13或PC14等提供待分类的短文本或其他任意文本，并由服务器11实现相应的分类操作。当然，需要指出的是：在一些情况下，待分类的短文本或其他文本也可能并不需要由手机13、PC14等提供，而由服务器11通过其他方式获得、甚至由服务器11自行生成；以及，在一些情况下，还可能由手机13、PC14等对短文本或其他文本实施分类操作，而无需服务器11的配合；或者，还可能存在其他情况，但这些情况显然都属于对图1所示实施例的合理调整或变形，均属于本领域技术人员能够理解的关联方案，应当被包含于本申请的保护范围内。

[0052] 下面结合实施例，对本申请的短文本分类方案进行说明。

[0053] 图2是本申请一示例性实施例提供的一种短文本分类方法的流程图。如图2所示，该方法可以应用于诸如图1所示的服务器11、手机13或PC14等各类设备上，可以包括以下步骤：

[0054] 步骤202，确定与用户相关的短文本归属于各个候选类目的概率。

[0055] 在本实施例中，用户与短文本之间可以通过多种方式建立关联关系，以使得短文本被作为与用户相关的短文本。例如，该短文本可以包括：用户输入的短文本(比如用户输入的记账信息等)、用户接收的短文本、用户浏览的短文本、用户在浏览过程中选取的短文本、基于用户操作而生成的短文本(比如用户的消费流水信息等)等，本申请并不对此进行限制。

[0056] 在本实施例中，可以通过类目召回模型确定所述短文本归属于各个候选类目的概率；其中，所述类目召回模型由基于全量用户数据生成的训练样本进行训练得到。本申请中的用户数据是指已分类的历史用户数据，通过训练可以学习到历史用户数据的分类情况，以用于对后续产生的新数据进行合理分配。“全量用户数据”可以理解为所有用户产生的历史数据的集合，该历史数据可以包括历史短文本、历史短文本归属的类目等信息；当然，随着时间不断推进，“全量用户数据”包含的用户和数据都可能不断更新，那么本申请中的“全量用户数据”可以包含截止于任意时刻之前的历史数据。当训练样本对应的数据来源(即历史数据)越多时，往往能够反映出越多用户的分类需求，因而上述的任意时刻可以在时序上尽可能地靠后，使得训练处的类目召回模型能够尽可能地适用于更多用户和更多应用场景，具有较强的普遍适用性。

[0057] 在本实施例中，所述训练样本可以包括所述全量用户数据中的历史短文本转换得到的文本向量矩阵。其中，所述文本向量矩阵可以由文本特征转换得到的词向量组成，所述文本特征可以包括所述历史短文本被切分得到的分词，使得该文本向量矩阵可以对相应的历史短文本实现分布式表达(Distributed Representation)，从而有效克服短文本的表达稀疏问题，不仅有助于提升存储、计算效率，还能够避免发生过拟合。当然，在其他实施例中，训练样本也可以采用历史短文本对应的文本向量，比如该文本向量可以采用诸如独热表达(One-Hot Representation)等形式，或者采用历史短文本对应的偏旁部首(比如“吃饭”表征为“口”、“乞”、“饣”和“反”)、拼音(比如“吃饭”表征为“chi fan”)等更细粒度的文本向量形式，同样可以适用于本申请中个性化的短文本分类方案。

[0058] 进一步地，所述文本特征还可以包括：所述历史短文本归属的类目的信息，比如所述历史短文本归属的类目的信息可以被添加至所述历史短文本的前方和后方中至少之一，以构成所述文本特征。由于相关技术中的向量转换算法更加适用于长文本(或称为，非短文本)，因而可以通过添加类目的信息来增加文本特征的长度，使得该文本特征更加适应于相关技术中的向量转换算法，提升向量转换算法得到的文本向量的表达准确度。

[0059] 在本实施例中，所述文本特征是在参照词向量集合和字向量集合的情况下，被转换为相应的词向量；其中，所述词向量集合包括所述全量用户数据被切分得到的所有文本特征和对应的词向量之间的映射关系，所述字向量集合包括所述全量用户数据采用的文字的全量字和对应的字向量之间的映射关系。例如，对于词向量集合，可以将全量用户数据中的所有历史短文本分别进行切词处理，并根据得到的所有分词，生成上述的所有文本特征，然后将所有文本特征分别转换为相应的词向量，即可得到上述的词向量集合。对于字向量集合，当全量用户数据采用中文时，可以获得所有中文字，并转换得到所有中文字对应的字向量，即可得到上述的字向量集合。那么，即便全量用户数据发生更新，更新的历史短文本仍然可以通过参考上述的词向量集合和字向量集合，顺利转换得到相应的文本向量矩阵，从而既可以解决短文本的稀疏性问题、提升文本特征被转换为文本向量时的表达准确度，又能够增强泛化能力，以便于较好地应对未知数据。并且，当文本特征中包含类目的信息时，使得在根据上述的词向量集合和字向量集合对文本特征进行转换的过程中，该类目的信息可以作为监督信息而被应用于对该文本特征的无监督学习(即上述的“转换”)过程中，以提升向量转换算法得到的文本向量的表达准确度。

[0060] 在本实施例中，当类目召回模型的训练样本为历史短文本转换得到的文本向量矩阵时，对于待分类的短文本，可以将该短文本转换为相应的文本向量矩阵，并通过所述类目召回模型确定该短文本对应的文本向量矩阵归属于各个候选类目的概率。类似地，当类目召回模型的训练样本采用基于其他形式的特征时，可以通过将短文本转换为相应形式的特征，并由类目召回模型确定该特征归属于各个候选类目的概率。

[0061] 步骤204，针对所述短文本，确定所述用户对各个候选类目的偏好权重。

[0062] 在本实施例中，可以通过对应于所述用户的类目偏好模型，确定所述用户对各个候选类目的偏好权重；其中，所述类目偏好模型被基于所述用户的个人用户数据而构建，使得该类目偏好模型和偏好权重都能够有效地反映出用户的个性化分类需求，有助于提升对短文本的分类准确度，确保对短文本的分类结果更加贴近于用户的实际分类习惯。

[0063] 在本实施例中，所述类目偏好模型可以被基于朴素贝叶斯或贝叶斯网络而构建；当然，还可能通过其他方式构建该类目偏好模型，本申请并不对此进行限制。

[0064] 步骤206，根据所述概率和所述偏好权重，确定所述短文本归属的类目。

[0065] 在本实施例中，类目召回模型的数量可以仅为一个。对于与用户i相关的短文本，假定总共存在j个候选类目，类目召回模型可以分别确定该短文本归属于各个候选类目的概率Pj；以及，对于该短文本，假定用户i对各个候选类目的偏好权重为Dj，那么可以分别计算各个候选类目对应的得分Sj＝Pj×Dj，并将最高得分的候选类目确定为该短文本归属的类目。

[0066] 在本实施例中，类目召回模型的数量可以为多个，有助于融合各个类目召回模型的优势和特点，提升类目召回的准确度。对于与用户i相关的短文本，假定总共存在j个候选

类目、m个类目召回模型，每个类目召回模型可以分别确定短文本归属于各个候选类目的概率，比如第r个类目召回模型对应的概率为Prj，r∈[1，m]；以及，对于该短文本，假定用户i对各个候选类目的偏好权重为Dj，那么可以分别计算各个候选类目对应的得分Sj＝P1j×Dj+P2j×Dj……+Pmj×Dj，并将最高得分的候选类目确定为该短文本归属的类目。

[0067] 例如，类目召回模型可以包括第一类目召回模型和第二类目召回模型。其中，所述第一类目召回模型可以由所述全量用户数据对应的所有训练样本进行训练得到，比如可以采用基于卷积神经网络(Convolutional Neural Network，CNN)算法进行训练得到的卷积神经网络模型，以充分利用CNN算法对于大样本类目具有较好训练效果的特点。在一实施方式中，本申请中的CNN算法可以采用门控线性单元(Gated Linear Unit，GLU)作为激活函数，以使得CNN算法在保持非线性能力的基础上，通过提供线性路径来大幅缓解梯度消失问题，有助于加快CNN模型的收敛速度。在一实施方式中，虽然本申请构建的卷积神经网络为浅层网络(相对于深层神经网络)，但是可以在该浅层网络的池化(pooling)层与全连接(Fully Connected，FC)层之间采用深层神经网络采用的信息高速公路(Highway)结构进行联通，有助于加快CNN模型的收敛速度。在一实施方式中，本申请的卷积神经网络的全连接层可以带有Dropout(丢弃)结构和softmax分类器，利用CNN较强的非线性映射能力，在对于包含上百个候选类目的情况下，表现可以远优于相关技术中的机器学习算法。

[0068] 而所述第二类目召回模型由所述全量用户数据中的小样本类目的用户数据对应的训练样本进行训练得到，其中所述小样本类目对应的用户数据的数据量小于预设数量，从而能够解决各个候选类目的样本数量不均衡的问题，增加小样本类目的权重占比。例如，所述第二类目召回模型可以包括：长短期记忆网络(Long Short-Term Memory，LSTM)模型或双向长短期记忆循环网络(Bi-directional LSTM Recurrent Neural Network，Bi-LSTM)模型等。

[0069] 可见，通过同时采用上述的第一类目召回模型和第二类目召回模型，比如第一类目召回模型采用CNN模型、第二类目召回模型采用Bi-LSTM模型，可以融合不同算法的训练优势，不仅能够缓解数据非平衡造成的分类误差，而且有助于最终获得更好的分类效果。

[0070] 在本实施例中，当类目召回模型的数量为多个时，将待分类的短文本转换为相应的文本向量矩阵，并由各个类目召回模型分别对该文本向量矩阵进行处理，以得到相应的概率。其中，该文本向量矩阵按照细粒度划分时，可以包括词向量矩阵、字向量矩阵等不同类型，而各个类目召回模型可以对该词向量矩阵进行处理，也可以对字向量矩阵或其他细粒度的文本向量矩阵进行处理，本申请并不对此进行限制。例如，当采用上述的CNN模型和Bi-LSTM模型对短文本进行分类时，CNN模型可以对该短文本对应的词向量矩阵或字向量矩阵，Bi-LSTM模型也可以对该短文本对应的词向量矩阵或字向量矩阵；当然，不同的类目召回模型在对不同类型的文本向量矩阵进行处理时，可能存在一定的处理差异，比如CNN模型可能相对更加适合于对词向量矩阵进行处理、Bi-LSTM模型可能相对更加适合于对字向量矩阵进行处理，那么可以分别生成该短文本对应的词向量矩阵和字向量矩阵，并由CNN模型对词向量矩阵进行处理、Bi-LSTM模型对字向量矩阵进行处理，以使得到的概率具有更高的准确度。

[0071] 图3是本申请一示例性实施例提供的一种文本分类方法的流程图。如图3所示，该方法可以应用于诸如图1所示的服务器11、手机13或PC14等各类设备上，可以包括以下步

骤：

[0072] 步骤302，确定与用户相关的文本归属于各个候选类目的概率。

[0073] 在本实施例中，文本除了图2所示的短文本之外，还可以适用于其他任意类型的文本，比如长文本等，本申请并不对此进行限制。

[0074] 在本实施例中，用户与文本之间可以通过多种方式建立关联关系，以使得文本被作为与用户相关的文本。例如，该文本可以包括：用户输入的文本、用户接收的文本、用户浏览的文本、用户在浏览过程中选取的文本、基于用户操作而生成的文本等，本申请并不对此进行限制。

[0075] 在本实施例中，可以通过类目召回模型确定所述文本归属于各个候选类目的概率；其中，所述类目召回模型由基于全量用户数据生成的训练样本进行训练得到。本申请中的用户数据是指已分类的历史用户数据，通过训练可以学习到历史用户数据的分类情况，以用于对后续产生的新数据进行合理分配。“全量用户数据”可以理解为所有用户产生的历史数据的集合，该历史数据可以包括历史文本、历史文本归属的类目等信息；当然，随着时间不断推进，“全量用户数据”包含的用户和数据都可能不断更新，那么本申请中的“全量用户数据”可以包含截止于任意时刻之前的历史数据。当训练样本对应的数据来源(即历史数据)越多时，往往能够反映出越多用户的分类需求，因而上述的任意时刻可以在时序上尽可能地靠后，使得训练处的类目召回模型能够尽可能地适用于更多用户和更多应用场景，具有较强的普遍适用性。

[0076] 在本实施例中，所述训练样本可以包括所述全量用户数据中的历史文本转换得到的文本向量矩阵。其中，所述文本向量矩阵可以由文本特征转换得到的词向量组成，所述文本特征可以包括所述历史文本被切分得到的分词，使得该文本向量矩阵可以对相应的历史短文本实现分布式表达(Distributed Representation)，从而有效克服文本可能存在的表达稀疏问题，不仅有助于提升存储、计算效率，还能够避免发生过拟合。当然，在其他实施例中，训练样本也可以采用历史短文本对应的文本向量，比如该文本向量可以采用诸如独热表达(One-Hot Representation)等形式，或者采用历史短文本对应的偏旁部首(比如“吃饭”表征为“口”、“乞”、“饣”和“反”)、拼音(比如“吃饭”表征为“chi fan”)等更细粒度的文本向量形式，同样可以适用于本申请中个性化的短文本分类方案。

[0077] 进一步地，所述文本特征还可以包括：所述历史文本归属的类目的信息，比如所述历史文本归属的类目的信息可以被添加至所述历史文本的前方和后方中至少之一，以构成所述文本特征。由于相关技术中的向量转换算法更加适用于长文本，因而可以通过添加类目的信息来增加文本特征的长度，使得该文本特征更加适应于相关技术中的向量转换算法，提升向量转换算法得到的文本向量的表达准确度。

[0078] 在本实施例中，所述文本特征是在参照词向量集合和字向量集合的情况下，被转换为相应的词向量；其中，所述词向量集合包括所述全量用户数据被切分得到的所有文本特征和对应的词向量之间的映射关系，所述字向量集合包括所述全量用户数据采用的文字的全量字和对应的字向量之间的映射关系。例如，对于词向量集合，可以将全量用户数据中的所有历史文本分别进行切词处理，并根据得到的所有分词，生成上述的所有文本特征，然后将所有文本特征分别转换为相应的词向量，即可得到上述的词向量集合。对于字向量集合，当全量用户数据采用中文时，可以获得所有中文字，并转换得到所有中文字对应的字向

量,即可得到上述的字向量集合。那么,即便全量用户数据发生更新,更新的历史文本仍然可以通过参考上述的词向量集合和字向量集合,顺利转换得到相应的文本向量矩阵,从而既可以解决文本可能存在的稀疏性问题、提升文本特征被转换为文本向量时的表达准确度,又能够增强泛化能力,以便于较好地应对未知数据。并且,当文本特征中包含类目的信息时,使得在根据上述的词向量集合和字向量集合对文本特征进行转换的过程中,该类目的信息可以作为监督信息而被应用于对该文本特征的无监督学习(即上述的“转换”)过程中,以提升向量转换算法得到的文本向量的表达准确度。

[0079] 在本实施例中,当类目召回模型的训练样本为历史短文本转换得到的文本向量矩阵时,对于待分类的文本,可以将该文本转换为相应的文本向量矩阵,并通过所述类目召回模型确定该文本对应的文本向量矩阵归属于各个候选类目的概率。类似地,当类目召回模型的训练样本采用基于其他形式的特征时,可以通过将文本转换为相应形式的特征,并由类目召回模型确定该特征归属于各个候选类目的概率。

[0080] 步骤304,针对所述文本,确定所述用户对各个候选类目的偏好权重。

[0081] 在本实施例中,可以通过对应于所述用户的类目偏好模型,确定所述用户对各个候选类目的偏好权重;其中,所述类目偏好模型被基于所述用户的个人用户数据而构建,使得该类目偏好模型和偏好权重都能够有效地反映出用户的个性化分类需求,有助于提升对文本的分类准确度,确保对文本的分类结果更加贴近于用户的实际分类习惯。

[0082] 在本实施例中,所述类目偏好模型可以被基于朴素贝叶斯或贝叶斯网络而构建;当然,还可能通过其他方式构建该类目偏好模型,本申请并不对此进行限制。

[0083] 步骤306,根据所述概率和所述偏好权重,确定所述文本归属的类目。

[0084] 在本实施例中,类目召回模型的数量可以仅为一个。对于与用户i相关的文本,假定总共存在j个候选类目,类目召回模型可以分别确定该文本归属于各个候选类目的概率Pj;以及,对于该文本,假定用户i对各个候选类目的偏好权重为Dj,那么可以分别计算各个候选类目对应的得分Sj=Pj×Dj,并将最高得分的候选类目确定为该文本归属的类目。

[0085] 在本实施例中,类目召回模型的数量可以为多个,有助于融合各个类目召回模型的优势和特点,提升类目召回的准确度。对于与用户i相关的文本,假定总共存在j个候选类目、m个类目召回模型,每个类目召回模型可以分别确定文本归属于各个候选类目的概率,比如第r个类目召回模型对应的概率为Prj,r∈[1,m];以及,对于该文本,假定用户i对各个候选类目的偏好权重为Dj,那么可以分别计算各个候选类目对应的得分Sj=P1j×Dj+P2j×Dj……+Pmj×Dj,并将最高得分的候选类目确定为该文本归属的类目。

[0086] 例如,类目召回模型可以包括第一类目召回模型和第二类目召回模型。其中,所述第一类目召回模型可以由所述全量用户数据对应的所有训练样本进行训练得到,比如可以采用基于卷积神经网络(Convolutional Neural Network,CNN)算法进行训练得到的卷积神经网络模型,以充分利用CNN算法对于大样本类目具有较好训练效果的特点。在一实施方式中,本申请中的CNN算法可以采用门控线性单元(Gated Linear Unit,GLU)作为激活函数,以使得CNN算法在保持非线性能力的基础上,通过提供线性路径来大幅缓解梯度消失问题,有助于加快CNN模型的收敛速度。在一实施方式中,虽然本申请构建的卷积神经网络为浅层网络(相对于深层神经网络),但是可以在该浅层网络的池化(pooling)层与全连接(Fully Connected,FC)层之间采用深层神经网络采用的信息高速公路(Highway)结构进行

联通，有助于加快CNN模型的收敛速度。在一实施方式中，本申请的卷积神经网络的全连接层可以带有Dropout(丢弃)结构和softmax分类器，利用CNN较强的非线性映射能力，在对于包含上百个候选类目的情况下，表现可以远优于相关技术中的机器学习算法。

[0087] 而所述第二类目召回模型由所述全量用户数据中的小样本类目的用户数据对应的训练样本进行训练得到，其中所述小样本类目对应的用户数据的数据量小于预设数量，从而能够解决各个候选类目的样本数量不均衡的问题，增加小样本类目的权重占比。例如，所述第二类目召回模型可以包括：长短期记忆网络(Long Short-Term Memory，LSTM)模型或双向长短期记忆循环网络(Bi-directional LSTM Recurrent Neural Network，Bi-LSTM)模型等。

[0088] 可见，通过同时采用上述的第一类目召回模型和第二类目召回模型，比如第一类目召回模型采用CNN模型、第二类目召回模型采用Bi-LSTM模型，可以融合不同算法的训练优势，不仅能够缓解数据非平衡造成的分类误差，而且有助于最终获得更好的分类效果。

[0089] 在本实施例中，当类目召回模型的数量为多个时，将待分类的短文本转换为相应的文本向量矩阵，并由各个类目召回模型分别对该文本向量矩阵进行处理，以得到相应的概率。其中，该文本向量矩阵按照细粒度划分时，可以包括词向量矩阵、字向量矩阵等不同类型，而各个类目召回模型可以对该词向量矩阵进行处理，也可以对字向量矩阵或其他细粒度的文本向量矩阵进行处理，本申请并不对此进行限制。例如，当采用上述的CNN模型和Bi-LSTM模型对短文本进行分类时，CNN模型可以对该短文本对应的词向量矩阵或字向量矩阵，Bi-LSTM模型也可以对该短文本对应的词向量矩阵或字向量矩阵；当然，不同的类目召回模型在对不同类型的文本向量矩阵进行处理时，可能存在一定的处理差异，比如CNN模型可能相对更加适合于对词向量矩阵进行处理、Bi-LSTM模型可能相对更加适合于对字向量矩阵进行处理，那么可以分别生成该短文本对应的词向量矩阵和字向量矩阵，并由CNN模型对词向量矩阵进行处理、Bi-LSTM模型对字向量矩阵进行处理，以使得到的概率具有更高的准确度。

[0090] 为了便于理解，下面以记账应用为例，对本申请的技术方案进行说明。假定手机13或PC14上运行有记账应用客户端、服务器11上运行有记账应用服务端，其中手机13或PC14上的记账应用客户端登录有用户的注册账号，使得手机13或PC14可以接收用户输入的记账信息，并基于该注册账号与服务器11进行数据交互，使得服务器11可以对该记账信息进行智能分类，而无需用户手动选择记账信息的所属类目。

[0091] 例如，图4-6是相关技术的一种记账应用的客户端界面的示意图。在如图4所示的界面中，可以包含记账功能触发选项，比如该选项可以展示为图4所示的“记一笔”或其他任意内容。当检测到用户触发该记账功能选项时，可以转入图5所示的记账界面，以便用户输入记账信息，比如记账类型(如支出、收入、转账或借贷等)、金额、类目、账户等；其中，当检测到用户触发“类目”选项时，可以展示出图6所示的类目选择界面，并由用户对该类目选择界面示出的候选类目进行选择。

[0092] 然而，类目选择界面中包含的候选类目往往数量众多，比如餐饮、交通、购物、娱乐等大类，且每一大类下还包含若干小类，比如餐饮可以进一步包括早餐、午餐、晚餐、饮料水果等。实际上，为了便于用户详细记录以及后续的数据统计，记账应用提供的类目往往会达到上百种，导致用户每次记账时都需要实施复杂的类目选择操作，造成记账效率降低、打击

用户的记账积极性。

[0093] 而本申请提出了短文本分类方案，通过将该方案应用于对记账信息的处理，可以自动、准确地识别记账信息所属的类目，而无需用户手动选择，从而有助于简化用户记账操作、提升记账效率。下面结合图7-11，对基于本申请技术方案的记账信息分类过程进行详细说明。

[0094] 图7是本申请一示例性实施例提供的一种记账信息分类过程的示意图。如图7所示，对记账信息实施自动分类的过程，可以包含两个阶段：准备阶段和处理阶段；其中，准备阶段通过：①准备语料、②处理语料、③模型训练，可以得到用于对记账信息实施自动分类的模型，使得在处理阶段可以基于该模型对用户输入的记账信息进行分类。

[0095] 本申请的模型可以包括：类目召回模型和类目偏好模型。其中，类目召回模型用于计算各个候选类目对待识别的记账信息的召回概率；而类目偏好模型用于计算用户在面对该待识别的记账信息时，对各个候选类目的偏好权重，以便对记账信息的最终分类结果符合用户的个人习惯，实现对不同用户输入的记账信息的智能化、个性化地分类处理。

[0096] 为了得到上述的类目召回模型和类目偏好模型，需要采用不同的语料和算法实施模型训练，下面分别针对两个模型的训练过程进行描述：

[0097] 图8是本申请一示例性实施例提供的一种训练类目召回模型的示意图。如图8所示，类目召回模型的训练样本可以为基于全量用户记账文本生成的文本向量矩阵，即全量用户记账文本在向量空间中的表征信息。全量用户记账文本包括所有用户对应的历史记账数据，该历史记账数据中包含历史记账信息、该历史记账信息归属的类目等内容。通过对全量用户记账文本中的每条历史记账信息进行切词处理，可以获得相应的若干分词，比如将某一历史记账信息“我今天去饭店吃饭”切分为“我”、“今天”、“去”、“饭店”、“吃饭”共5个分词，并分别将每个分词转换为相应的文本向量，即可得到该历史记账信息对应的文本向量矩阵；比如，当每个分词被表达为8维文本向量(每个维度上的数值可以为该分词在相应维度上的权重)时，历史记账信息“我今天去饭店吃饭”整体可以被表达为相应的5×8矩阵。通过上述方式，使得每一历史记账信息均可以被表达为相应的文本向量矩阵，该方式属于分布式表达(Distributed Representation)，能够有效克服历史记账信息的文本可能存在的表达稀疏问题，不仅有助于提升存储、计算效率，还能够避免在训练类目召回模型的过程中发生过拟合。

[0098] 在本实施例中，可以利用word2vec(词转换至向量)工具或其他任意方式将历史记账信息对应的分词转换为文本向量。但是，由于word2vec工具本身更加适用于对长文本的转换，而记账信息往往为长度较短的短文本，容易引入较大的噪音、可能造成文本向量不准确的问题。为此，可以对上述分词进行一定处理，比如在历史记账信息的前方、后方中至少之一添加相应类目的信息；仍以上述历史记账信息“我今天去饭店吃饭”为例，假定该历史记账信息被用户记录至“午餐”类目，那么以该历史记账信息对应的分词“吃饭”为例，可以添加“午餐”类目以得到诸如“午餐吃饭午餐”(在分词“吃饭”的前方和后方同时添加类目“午餐”而得到)等修正后的分词，然后通过word2vec等词向量工具对该修正后的分词进行转换，以得到相应的文本向量。那么，通过上述方式对分词进行修正，可以增加分词的长度、以使得修正后的分词更加适应word2vec等词向量工具，有助于提升转换得到的文本向量的准确度。

[0099] 虽然可以直接通过word2vec等词向量工具将全量用户记账文本包含的历史记账信息转换为文本向量矩阵，以用于训练类目召回模型，但是由于“全量用户记账文本”并非静态数据集合，而是不断地更新其包含的数据，因而为了提升泛化能力、实现对“全量用户记账文本”不断增加的新数据的良好兼容，以及进一步解决记账信息的稀疏问题，还可以采用下述处理方式：

[0100] 如图8所示，一方面获得全量用户记账文本(全量用户记账文本可能不断发生更新，此处可以为任意时刻的版本)，并通过上述方式将该全量用户记账文本包含的历史记账信息进行分词后，通过word2vec等词向量工具将这些分词转换为相应的词向量，并生成相应的向量字典，即图8所示的字典1，该字典1包含上述的分词与词向量构成的信息对——(词，词向量)。另一方面，根据全量用户记账文本采用的文字类型，以中文为例，可以获得全量中文文本(例如来源于维基中文百科等)，并通过切字(即细粒度为字的切词处理)得到所有中文的单个汉字，然后通过word2vec等词向量工具将这些单个汉字转换为相应的字向量，并生成相应的向量字典，即图8所示的字典2，该字典2包含上述的单个汉字与字向量构成的信息对——(字，字向量)。

[0101] 然后，基于获得的上述字典1和字典2，可以通过word2vec等词向量工具对全量用户记账文本包含的每一历史记账信息进行转换，以获得相应的文本向量矩阵；并且，即便全量用户记账文本中出现更新数据时，仍然可以根据上述的字典1和字典2对该更新数据进行转换处理，就有很强的泛化能力和兼容性。其中，在对每一历史记账信息对应的分词进行转换之前，类似于上述实施例的方式，可以根据每一历史记账信息对应的类目，在每一历史记账信息的每个分词的前方和后方中至少之一添加相应的类目，得到修正后的分词；那么，当word2vec等词向量工具通过上述字典1和字典2对该修正后的分词进行转换操作时，虽然该转换过程属于无监督学习过程，但是添加的类目信息可以作为监督信息而被应用于该无监督学习过程，以使得到的文本向量能够更好、更准确地表达相应的分词。

[0102] 基于上述方式，可以将全量用户记账文本中的所有历史记账信息分别转换为相应的文本向量矩阵，并根据这些文本向量矩阵训练上述的类目召回模型。虽然相关技术中的模型训练算法都可以用于对上述的文本向量矩阵进行训练，但是训练结果可能存在一定的优劣，从而影响最终的记账信息的智能分类的准确度。

[0103] 实际上，每种模型训练算法都存在各自的优势和特点；因而在本实施例中，可以同时采用多种模型训练算法，以结合这些模型训练算法的优势、消除劣势，以实现对类目召回模型的优化。例如，本实施例可以采用CNN算法和Bi-LSTM算法对上述的文本向量矩阵进行训练；CNN算法和Bi-LSTM算法都可以自动化地提取特征，而无需人工构造特征，从而能够极大地简化模型训练的准备工作、提升模型训练效率，以及有助于极大地简化对类目召回模型的后期维护、更新等操作。

[0104] 对于CNN算法而言，CNN的基本结构包括特征提取层和特征映射层；特征提取层的每个神经元的输入与前一层的局部接受域相连、并提取该局部接受域的特征。一旦该局部接受域的特征被提取后，该特征与其它特征间的位置关系也随之确定下来。因此，CNN所需的特征都可以隐式地从训练样本中进行学习，而避免了显式的特征抽取，且无需人工参与。CNN通过卷积提取特征，将训练样本中符合条件(可以通过相应的激活值来判定；通常而言，激活值越大越可能符合条件)的部分筛选出来，以作为提取的特征。为了充分地提取特征，

可以添加多个卷积核;换言之,即采用长度不同的过滤器(filter)对作为训练样本的文本向量矩阵进行卷积,filter的宽度等于该文本向量矩阵的行宽度。然后,在池化层(如采用max-pooling操作,即对邻域内的特征点取最大)对每一filter提取的文本向量进行处理,使得每一filter得到对应的一个数字,而通过将这些filter对应的数字拼接起来,即可得到一个表征上述训练样本的向量,而基于CNN算法的类目召回模型可以基于该向量对待分类的记账信息进行分类预测。可见,上述过程中基于CNN算法,实际上是为了提取出训练样本中类似N-Gram的关键信息。

[0105] 对于Bi-LSTM算法而言,同样可以实现对特征的自动提取。并且,Bi-LSTM提取的特征,从某种意义上可以理解为捕获变长且双向的“n-gram”信息。

[0106] 在本实施例中,可以向CNN算法和Bi-LSTM算法提供相同的训练样本,比如全量用户记账文本对应的所有文本向量矩阵,以用于分别训练得到相应的类目召回模型。但是,由于CNN算法和Bi-LSTM算法具有不同特性,也可以对提供的训练样本进行一定调整;例如,由于CNN算法在训练样本的数量较多的情况下具有更好(相对于训练样本的数量较少的情况)的训练效果,而Bi-LSTM算法在训练样本的数量较少的情况下具有更好(相对于训练样本的数量较多的情况)的训练效果,因而可以通过CNN算法对全量的训练样本进行训练,而仅向Bi-LSTM算法提供小样本类目(即该类目下的训练样本数量小于预设数量)对应的训练样本,以融合两种算法的各自优势;尤其是,通过采用Bi-LSTM算法对小样本类目的训练样本进行训练,能够显著提升对小样本类目的预测准确度。

[0107] 进一步地,在通过CNN算法进行模型训练的过程中,为了加快模型收敛速度,可以对相关技术中的CNN算法实施下述改进:

[0108] 1)更换激活函数。在相关技术中,CNN算法通常采用的激活函数为Relu函数。而在本实施例中,激活函数可以采用GLU函数,从而在保持非线性能力的基础上,通过提供线性路径来大幅缓解梯度消失的问题,有助于加快模型收敛速度。

[0109] 2)采用Highway结构。在相关技术中,Highway结构被应用于深层神经网络中;而在本实施例中,将Highway结构应用于浅层神经网络的池化层与全连接层之间,能够显著加快模型收敛速度。

[0110] 当然,本实施例的CNN可以单独更换激活函数,或者单独采用Highway结构,也可以同时单独更换激活函数和采用Highway结构,两者并不存在必然的依赖关系,本申请并不对此进行限制。

[0111] 此外,在本申请的CNN中,其全连接层可以带有Dropout(丢弃)结构和softmax分类器,从而能够通过CNN较强的非线性映射能力,在面对候选类目的数量达到上百个的情况下,使得表现仍然可以远优于相关技术中的机器学习算法。

[0112] 因此,通过上述过程可以分别通过CNN算法和Bi-LSTM算法训练得到相应的类目召回模型,比如可以分别称为CNN模型和Bi-LSTM模型,以用于在后续过程中对记账信息进行自动化的智能分类。

[0113] 并且,本申请在模型训练过程中,可以采用诸如Tensorflow等带有checkpoint(检查点)功能的系统,使得当全量用户数据发生更新、导致训练样本更新时,可以基于当前版本的类目召回模型对应的checkpoint数据,对该类目召回模型进行增量学习和版本更新,而无需对训练样本进行重新训练。

[0114] 图9是本申请一示例性实施例提供的一种训练类目偏好模型的示意图。如图9所示，通过对任意用户i的历史记账数据进行训练，可以得到对应于该用户i的类目偏好模型；那么，通过结合上述的类目召回模型和该类目偏好模型，使得用户i输入记账信息时，可以实现适用于该用户i的个性化分类操作，以尽可能地贴近于该用户i对记账信息的分类习惯。

[0115] 如图9所示，获取用户i的历史记账数据，该历史记账数据中可以包括历史记账信息和用户i为该历史记账信息手动划分的类目。对历史记账信息进行切分得到相应的分词，并对该分词和对应的类目进行处理；其中，根据采用的对类目偏好模型的训练算法，对该分词和类目的处理方式可能存在不同。

[0116] 例如，当采用朴素贝叶斯(Naive Bayesian)算法时，可以根据每一分词和相应的类目，分别计算相应的P(分词|类目)、P(分词)、P(类目)等概率；其中，P(分词|类目)表示该分词被划分至该类目的概率，P(分词)表示该分词出现的概率，P(类目)表示该类目出现的概率。然后，基于朴素贝叶斯算法可以训练得到相应的类目偏好模型，该类目偏好模型为朴素贝叶斯模型。

[0117] 当采样贝叶斯网络(Bayesian Network)算法时，可以根据各个分词和相应的类目，构建基于连接强度特征扩展的贝叶斯网络，以形成上述的类目偏好模型；其中，采用的贝叶斯网络算法例如可以为KDB(k-dependence Bayesian network classifiers)算法等。

[0118] 图10是本申请一示例性实施例提供的一种记账应用的记账界面的示意图。如图10所示，当记账应用采用本申请的短文本分类方案时，记账界面可以包括“记账信息”选项，以供用户输入记账信息，比如“朋友聚餐”等；相比于图5所示的相关技术中的记账界面而言，图10所示的记账界面无需提供“类目”选项，记账应用可以根据用户输入的记账信息实现自动化、智能化的分类处理。例如，当用户i在诸如图10所示的记账界面输入记账信息时，该记账信息可以被记账应用的客户端上传至服务端，并由服务端通过图11所示的分类过程，对该记账信息进行分类；当然，在一些实施例中，记账应用的客户端可能直接对记账信息进行分类，本申请并不对此进行限制。

[0119] 图11是本申请一示例性实施例提供的一种对记账信息进行智能分类的示意图。如图11所示，针对用户i输入的记账信息，分别由训练得到的类目召回模型和类目偏好模型进行处理，并根据类目召回模型输出的概率信息和类目偏好模型输出的权重信息，计算记账信息对应的最终分类结果。

[0120] 首先，介绍概率信息的计算过程。基于上述实施例，训练得到的类目召回模型可以包括：CNN模型和Bi-LSTM模型。由于CNN模型和Bi-LSTM模型采用的训练样本均为历史记账信息对应的文本向量矩阵，因而对于用户i输入的记账信息，也应当转换为相应的文本向量矩阵，并由CNN模型和Bi-LSTM模型分别对该文本向量矩阵进行处理。例如，可以将用户i输入的记账信息分别按照词和字的细粒度进行转换，得到相应的词向量矩阵和字向量矩阵，并将词向量矩阵交由CNN模型进行处理、将字向量矩阵交由Bi-LSTM模型进行处理。

[0121] 假定记账应用中存在n种候选类目，那么训练得到的每一类目召回模型均包含对应于这n种候选类目的分类器，即图11所示的分类器1、分类器2……分类器n。那么，当这些分类器对用户i输入的记账信息对应的文本向量矩阵进行处理后，即可分别得到相应的分类结果1、分类结果2……分类结果n，用于表示该记账信息分别归属于这n种候选类目的概

率。因此，当分别采用CNN模型和Bi-LSTM模型对用户i输入的记账信息进行处理时，可以分别得到CNN模型计算出的概率$P_{CNN}1$、$P_{CNN}2$……$P_{CNN}n$，以及Bi-LSTM模型计算出的概率$P_{LSTM}1$、$P_{LSTM}2$……$P_{LSTM}n$。

[0122] 然后，介绍偏好权重的计算过程。假定类目偏好模型为贝叶斯网络，由于该贝叶斯网络是基于用户i的历史记账数据训练得到，因而该贝叶斯网络可以表现出用户i的历史分类习惯，从而确定出该用户i将当前输入的记账信息划分至各个候选类目的概率，并将该概率表现为上述的偏好权重，比如该偏好权重可以为D1、D2……Dn。

[0123] 根据CNN模型计算出的概率$P_{CNN}1$、$P_{CNN}2$……$P_{CNN}n$，Bi-LSTM模型计算出的概率$P_{LSTM}1$、$P_{LSTM}2$……$P_{LSTM}n$，以及类目偏好模型计算出的偏好权重D1、D2……Dn，可以计算出每一候选类目对应的分值S1、S2……Sn；其中：

[0124] $Sj=P_{CNN}j\times Dj+P_{LSTM}j\times Dj$，且$j\in[1,n]$

[0125] 然后，对各个候选类目对应的分值进行排序处理；其中，对于分值最大的候选类目，可以被选取为用户i输入的记账信息的最终分类结果。

[0126] 图12示出了根据本申请的一示例性实施例的电子设备的示意结构图。请参考图12，在硬件层面，该电子设备包括处理器1202、内部总线1204、网络接口1206、内存1208以及非易失性存储器1210，当然还可能包括其他业务所需要的硬件。处理器1202从非易失性存储器1210中读取对应的计算机程序到内存1208中然后运行，在逻辑层面上形成短文本分类装置。当然，除了软件实现方式之外，本申请并不排除其他实现方式，比如逻辑器件抑或软硬件结合的方式等等，也就是说以下处理流程的执行主体并不限定于各个逻辑单元，也可以是硬件或逻辑器件。

[0127] 请参考图13，在软件实施方式中，该短文本分类装置可以包括：

[0128] 概率确定单元1301，确定与用户相关的短文本归属于各个候选类目的概率；

[0129] 权重确定单元1302，针对所述短文本，确定所述用户对各个候选类目的偏好权重；

[0130] 类目确定单元1303，根据所述概率和所述偏好权重，确定所述短文本归属的类目。

[0131] 可选的，所述概率确定单元1301具体用于：

[0132] 通过类目召回模型确定所述短文本归属于各个候选类目的概率；其中，所述类目召回模型由基于全量用户数据生成的训练样本进行训练得到。

[0133] 可选的，所述训练样本包括所述全量用户数据中的历史短文本转换得到的文本向量矩阵。

[0134] 可选的，所述文本向量矩阵由文本特征转换得到的文本向量组成，所述文本特征包括所述历史短文本被切分得到的分词。

[0135] 可选的，所述文本特征还包括：所述历史短文本归属的类目的信息。

[0136] 可选的，所述文本特征是在参照词向量集合和字向量集合的情况下，被转换为相应的文本向量；其中，所述词向量集合包括所述全量用户数据被切分得到的所有文本特征和对应的词向量之间的映射关系，所述字向量集合包括所述全量用户数据采用的文字的全量字和对应的字向量之间的映射关系。

[0137] 可选的，所述概率确定单元1301具体用于：

[0138] 将所述短文本转换为相应的文本向量矩阵；

[0139] 通过所述类目召回模型确定所述短文本对应的文本向量矩阵归属于各个候选类

目的概率。

[0140] 可选的，所述类目召回模型的数量为多个；所述概率确定单元1301具体用于：

[0141] 通过多个类目召回模型分别计算所述短文本归属于各个候选类目的概率。

[0142] 可选的，所述类目召回模型包括：

[0143] 第一类目召回模型，所述第一类目召回模型由所述全量用户数据对应的所有训练样本进行训练得到；

[0144] 第二类目召回模型，所述第二类目召回模型由所述全量用户数据中的小样本类目的用户数据对应的训练样本进行训练得到，其中所述小样本类目对应的用户数据的数据量小于预设数量。

[0145] 可选的，所述第一类目召回模型包括：卷积神经网络模型。

[0146] 可选的，所述卷积神经网络模型采用的激活函数包括：门控线性单元。

[0147] 可选的，所述卷积神经网络模型的池化层与全连接层之间采用信息高速公路结构进行联通。

[0148] 可选的，所述第二类目召回模型包括：长短期记忆网络模型或双向长短期记忆循环网络模型。

[0149] 可选的，所述权重确定单元1302具体用于：

[0150] 通过对应于所述用户的类目偏好模型，确定所述用户对各个候选类目的偏好权重；其中，所述类目偏好模型被基于所述用户的个人用户数据而构建。

[0151] 可选的，所述类目偏好模型被基于朴素贝叶斯或贝叶斯网络而构建。

[0152] 图14示出了根据本申请的一示例性实施例的电子设备的示意结构图。请参考图14，在硬件层面，该电子设备包括处理器1402、内部总线1404、网络接口1406、内存1408以及非易失性存储器1410，当然还可能包括其他业务所需要的硬件。处理器1402从非易失性存储器1410中读取对应的计算机程序到内存1408中然后运行，在逻辑层面上形成文本分类装置。当然，除了软件实现方式之外，本申请并不排除其他实现方式，比如逻辑器件抑或软硬件结合的方式等等，也就是说以下处理流程的执行主体并不限定于各个逻辑单元，也可以是硬件或逻辑器件。

[0153] 请参考图15，在软件实施方式中，该文本分类装置可以包括：

[0154] 概率确定单元1501，确定与用户相关的文本归属于各个候选类目的概率；

[0155] 权重确定单元1502，针对所述文本，确定所述用户对各个候选类目的偏好权重；

[0156] 类目确定单元1503，根据所述概率和所述偏好权重，确定所述文本归属的类目。

[0157] 上述实施例阐明的系统、装置、模块或单元，具体可以由计算机芯片或实体实现，或者由具有某种功能的产品来实现。一种典型的实现设备为计算机，计算机的具体形式可以是个人计算机、膝上型计算机、蜂窝电话、相机电话、智能电话、个人数字助理、媒体播放器、导航设备、电子邮件收发设备、游戏控制台、平板计算机、可穿戴设备或者这些设备中的任意几种设备的组合。

[0158] 在一个典型的配置中，计算机包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。

[0159] 内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的

示例。

[0160] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体(transitory media)，如调制的数据信号和载波。

[0161] 还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

[0162] 在本申请使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本申请。在本申请和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本文中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[0163] 应当理解，尽管在本申请可能采用术语第一、第二、第三等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本申请范围的情况下，第一信息也可以被称为第二信息，类似地，第二信息也可以被称为第一信息。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[0164] 以上所述仅为本申请的较佳实施例而已，并不用以限制本申请，凡在本申请的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本申请保护的范围之内。

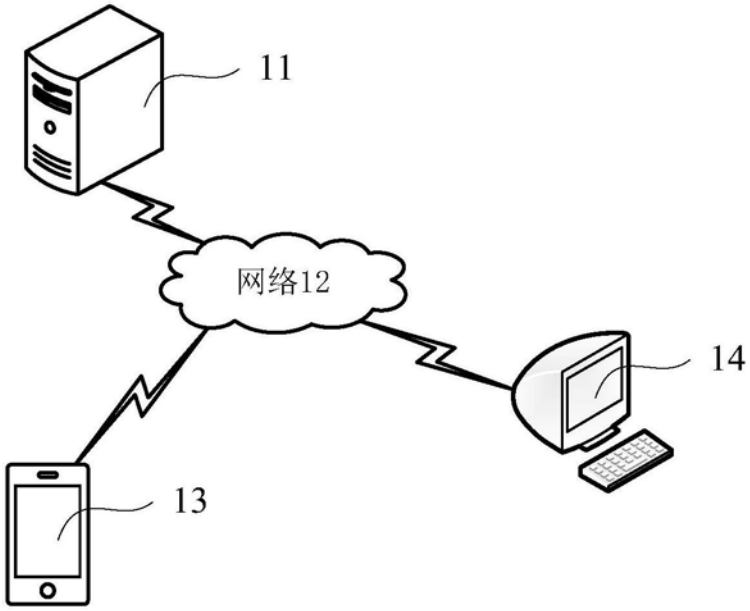

11
网络12
14
13

图1

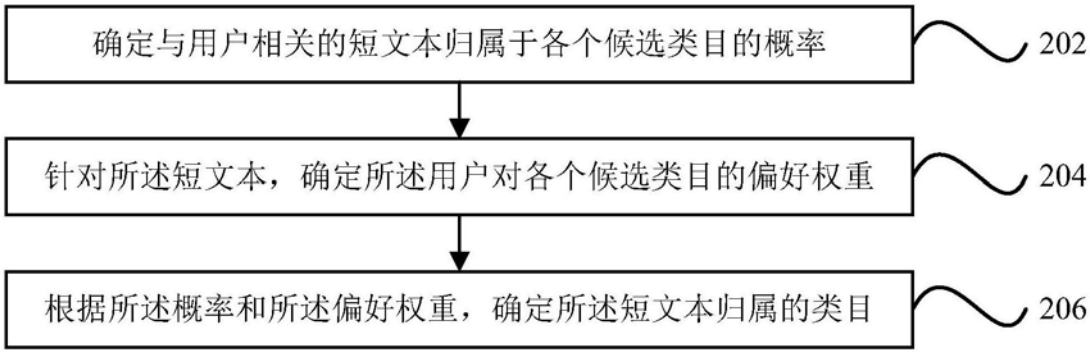

确定与用户相关的短文本归属于各个候选类目的概率
202
针对所述短文本，确定所述用户对各个候选类目的偏好权重
204
根据所述概率和所述偏好权重，确定所述短文本归属的类目
206

图2

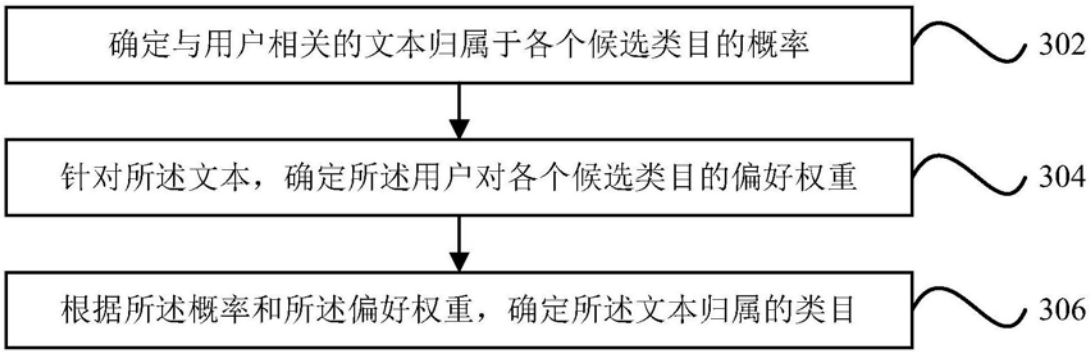

确定与用户相关的文本归属于各个候选类目的概率
302
针对所述文本，确定所述用户对各个候选类目的偏好权重
304
根据所述概率和所述偏好权重，确定所述文本归属的类目
306

图3


XX记账
¥x.xx
本月收入
¥x.xx
本月支出
¥x.xx
本月差额
记一笔
今日支出
未消费
统计信息
XXXX

图4


支出
收入
转账
借贷
金额：500.00
类目：
账户：现金
备注：请输入备注信息

图5


选择类目
餐饮
交通
购物
娱乐
医教
居家
投资
人情
生意
早餐
午餐
晚餐
饮料水果
零食

图6

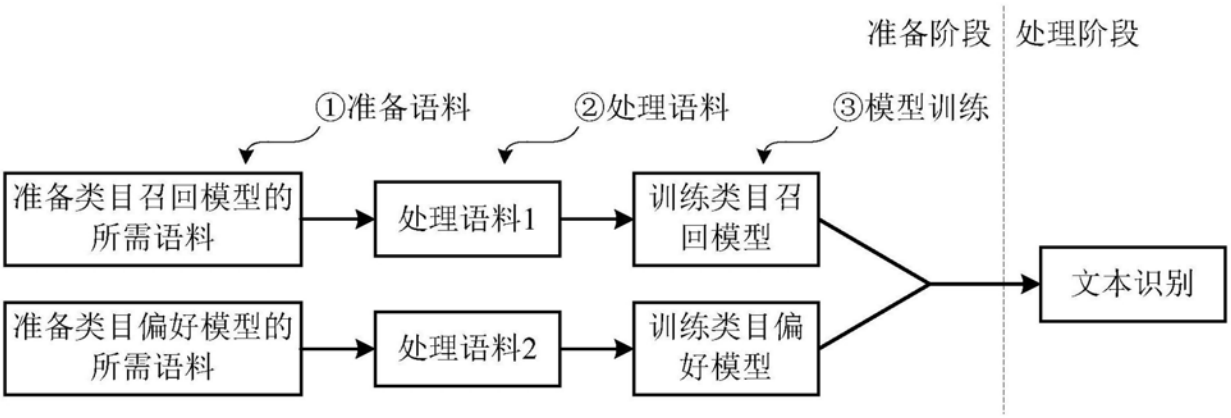

准备阶段
处理阶段
①准备语料
②处理语料
③模型训练
准备类目召回模型的所需语料
处理语料1
训练类目召回模型
准备类目偏好模型的所需语料
处理语料2
训练类目偏好模型
文本识别

图7

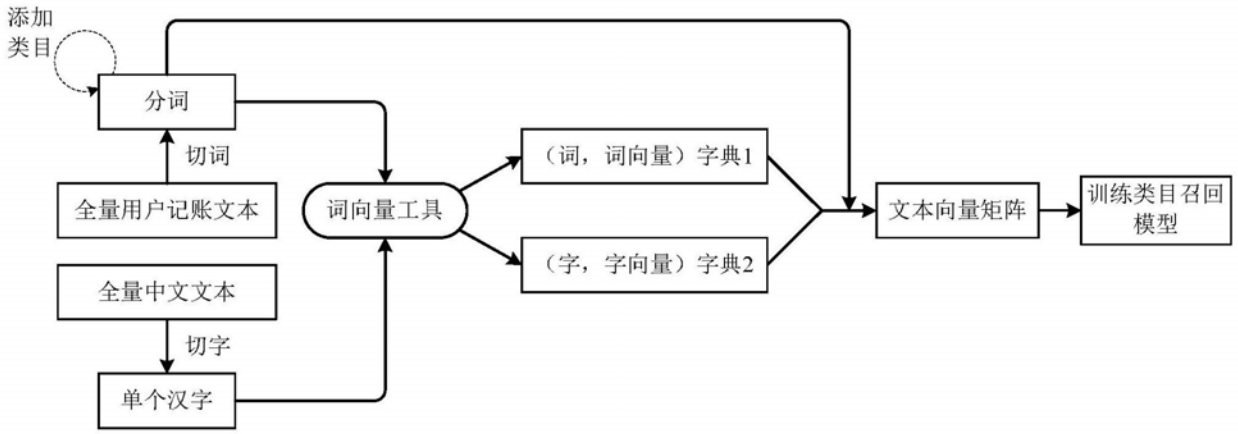

添加
类目
分词
切词
全量用户记账文本
全量中文文本
切字
单个汉字
词向量工具
（词，词向量）字典1
（字，字向量）字典2
文本向量矩阵
训练类目召回
模型

图8

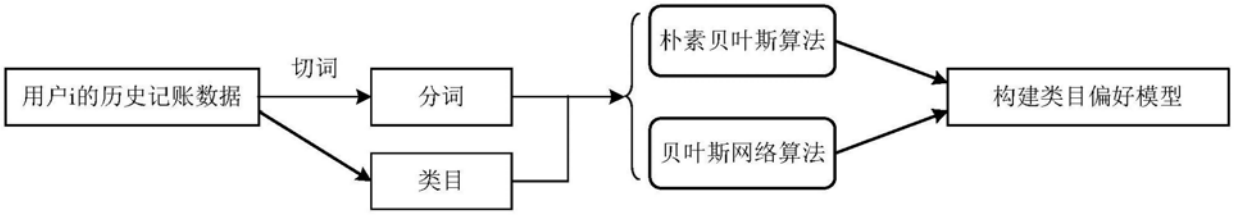

用户i的历史记账数据
切词
分词
类目
朴素贝叶斯算法
贝叶斯网络算法
构建类目偏好模型

图9


支出
收入
转账
借贷
金额：500.00
记账信息：请输入记账信息

图10

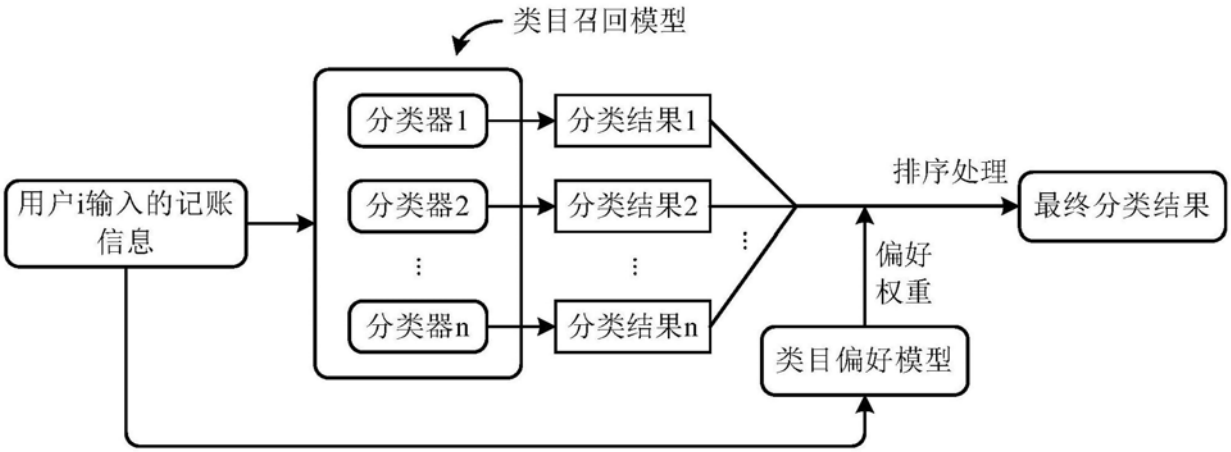

类目召回模型
分类器1
分类器2
分类器n
分类结果1
分类结果2
分类结果n
用户i输入的记账信息
排序处理
最终分类结果
偏好权重
类目偏好模型

图11

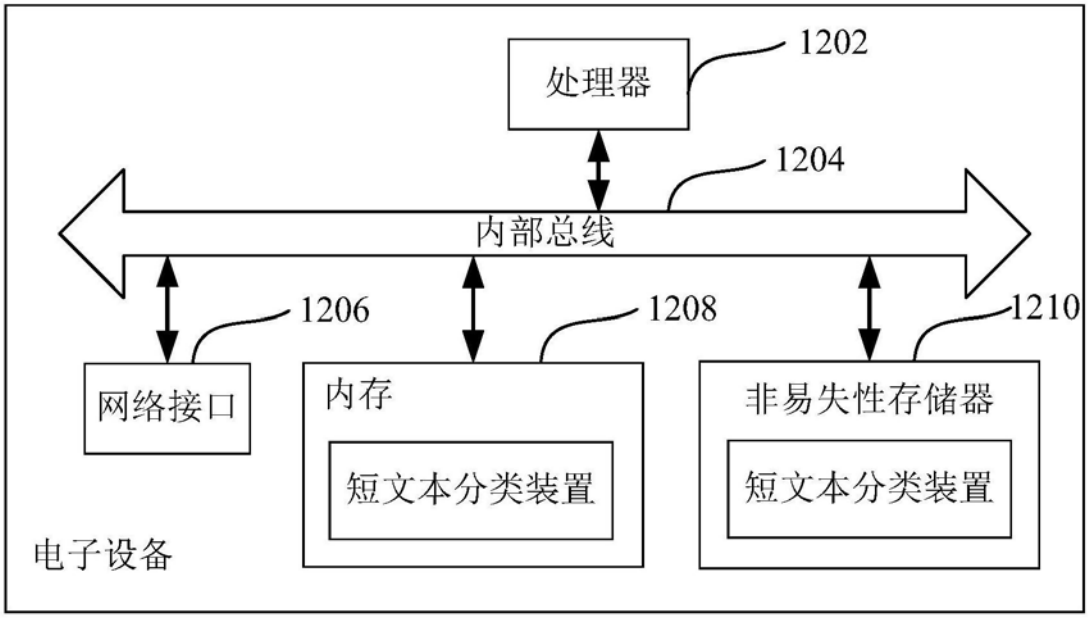

处理器
1202
1204
内部总线
1206
1208
1210
网络接口
内存
短文本分类装置
非易失性存储器
短文本分类装置
电子设备

图12

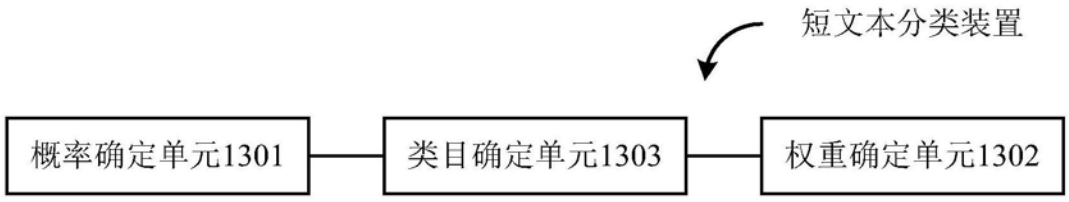

短文本分类装置
概率确定单元1301
类目确定单元1303
权重确定单元1302

图13

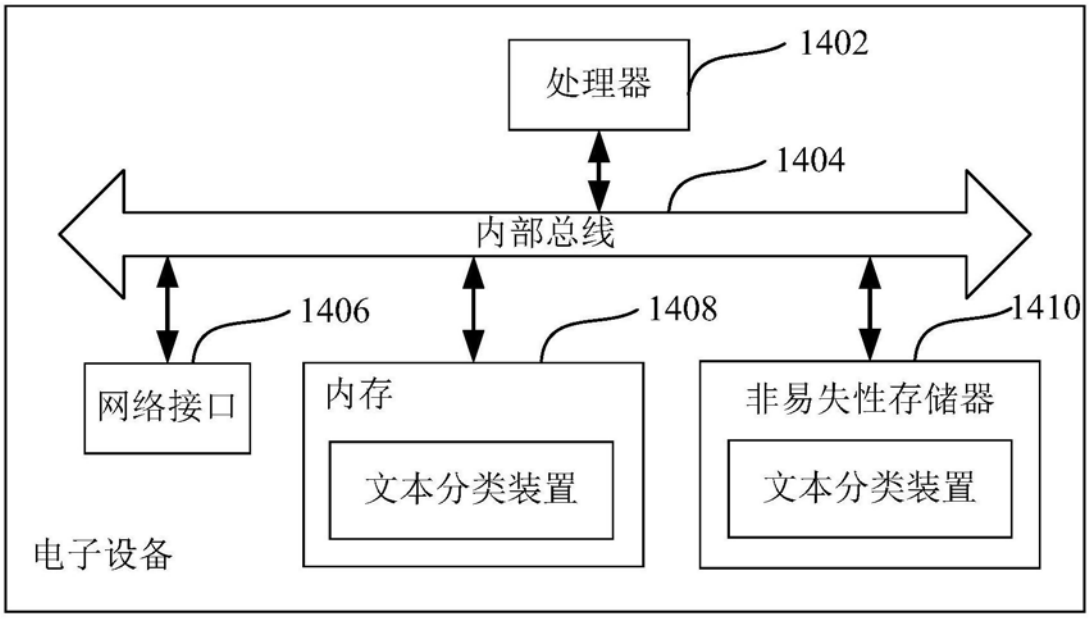

处理器
1402
1404
内部总线
1406
1408
1410
网络接口
内存
文本分类装置
非易失性存储器
文本分类装置
电子设备

图14

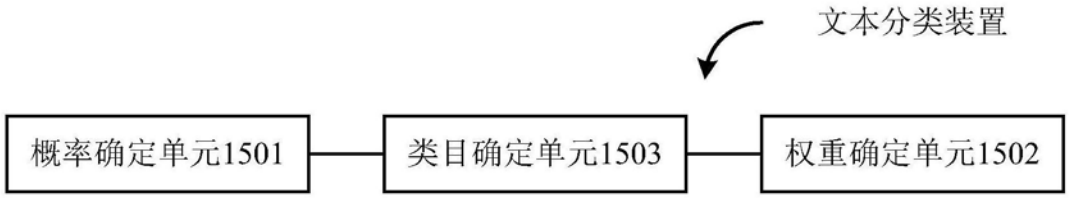

文本分类装置
概率确定单元1501
类目确定单元1503
权重确定单元1502

图15