



(12) 发明专利申请

(10) 申请公布号 CN 116340831 A

(43) 申请公布日 2023. 06. 27

(21) 申请号 202310596032.1

G06Q 10/08 (2023.01)

(22) 申请日 2023.05.24

(71) 申请人 京东科技信息技术有限公司

地址 100176 北京市大兴区经济技术开发区
科创十一街18号院2号楼6层601

(72) 发明人 张智慧 王阳阳 邹波 俞晓光
陈蒙 何晓冬

(74) 专利代理机构 北京品源专利代理有限公司
11332

专利代理人 蔡舒野

(51) Int. Cl.

G06F 18/24 (2023.01)

G06F 18/22 (2023.01)

G06F 18/23 (2023.01)

G06F 18/214 (2023.01)

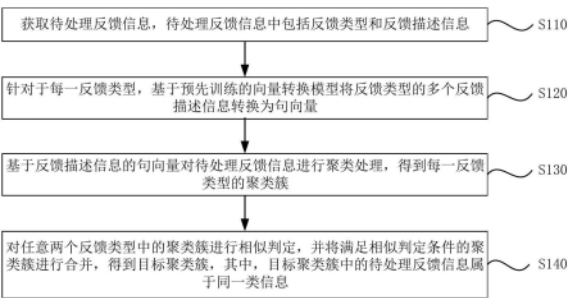
权利要求书2页 说明书12页 附图4页

(54) 发明名称

一种信息分类方法、装置、电子设备及存储
介质

(57) 摘要

本发明公开了一种信息分类方法、装置、电子设备及存储介质。包括：获取待处理反馈信息，待处理反馈信息中包括反馈类型和反馈描述信息；针对于每一反馈类型，基于预先训练的向量转换模型将反馈类型的多个反馈描述信息转换为句向量；基于反馈描述信息的句向量对待处理反馈信息进行聚类处理，得到每一反馈类型的聚类簇；对任意两个反馈类型中的聚类簇进行相似判定，并将满足相似判定条件的聚类簇进行合并，得到目标聚类簇，其中，目标聚类簇中的待处理反馈信息属于同一类信息。本方案实现了自动将满足相似判定条件的反馈信息合并到同一问题类型，避免了人工干预反馈问题分类的问题，提高信息分类的准确性和高效性。



1. 一种信息分类方法,其特征在于,包括:

获取待处理反馈信息,所述待处理反馈信息中包括反馈类型和反馈描述信息;

针对于每一所述反馈类型,基于预先训练的向量转换模型将所述反馈类型的多个反馈描述信息转换为句向量;基于所述反馈描述信息的句向量对所述待处理反馈信息进行聚类处理,得到每一所述反馈类型的聚类簇;

对任意两个反馈类型中的聚类簇进行相似判定,并将满足相似判定条件的聚类簇进行合并,得到目标聚类簇,其中,所述目标聚类簇中的待处理反馈信息属于同一类信息。

2. 根据权利要求1所述的方法,其特征在于,所述向量转换模型的训练方法包括:

构建初始模型,所述初始模型中包括第一向量转换模型、第二向量转换模型和相似判定模型;

获取样本数据,所述样本数据中包括第一描述信息、第二描述信息和相似度标签;

基于所述样本数据对所述初始模型进行训练,在满足训练结束条件时,将训练好的所述第一向量转换模型或所述第二向量转换模型确定为训练好的向量转换模型,其中,所述第一向量转换模型和所述第二向量转换模型在训练过程中权重共享。

3. 根据权利要求2所述的方法,其特征在于,所述第一向量转换模型用于将第一描述信息转换为第一句向量,所述第二向量转换模型用于将第二描述信息转换为第二句向量,所述相似判定模型用于生成所述第一句向量和所述第二句向量的相似度预测结果。

4. 根据权利要求2所述的方法,其特征在于,所述第一向量转换模型和所述第二向量转换模型分别包括:词向量转换模块、至少两个词向量特征提取模块和特征拼接模块。

5. 根据权利要求1所述的方法,其特征在于,所述基于所述反馈描述信息的句向量对所述待处理反馈信息进行聚类处理,得到每一所述反馈类型的聚类簇,包括:

遍历各所述句向量,分别计算两个句向量之间的相似度数据,并基于所述相似度数据确定所述待处理反馈信息所属的聚类簇。

6. 根据权利要求1所述的方法,其特征在于,所述对任意两个反馈类型中的聚类簇进行相似判定,并将满足相似判定条件的聚类簇进行合并,包括:

确定第一反馈类型的第一聚类簇中的第一反馈描述信息和第二反馈类型的第二聚类簇中的第二反馈描述信息;

对所述第一反馈描述信息的句向量和所述第二反馈描述信息的句向量进行相似度处理,得到相似度数据;

基于相似判定条件对所述相似度数据进行判定,在所述相似度数据满足所述相似判定条件的情况下,将所述第一聚类簇和所述第二聚类簇进行合并。

7. 根据权利要求1所述的方法,其特征在于,所述方法还包括:

设置同类判定阈值,基于所述同类判定阈值对所述目标聚类簇中的信息数进行判定;

释放不满足所述同类判定阈值的目标聚类簇。

8. 一种信息分类装置,其特征在于,包括:

待处理反馈信息获取模块,用于获取待处理反馈信息,所述待处理反馈信息中包括反馈类型和反馈描述信息;

聚类簇确定模块,用于针对于每一反馈类型,基于预先训练的向量转换模型将所述反馈类型的多个反馈描述信息转换为句向量;基于所述反馈描述信息的句向量对所述待处理

反馈信息进行聚类处理,得到每一所述反馈类型的聚类簇;

目标聚类簇确定模块,用于对任意两个反馈类型中的聚类簇进行相似判定,并将满足相似判定条件的聚类簇进行合并,得到目标聚类簇,其中,所述目标聚类簇中的待处理反馈信息属于同一类信息。

9. 一种电子设备,其特征在于,所述电子设备包括:

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计算机程序,所述计算机程序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权利要求1-7中任一项所述的信息分类方法。

10. 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,所述计算机指令用于使处理器执行时实现权利要求1-7中任一项所述的信息分类方法。

一种信息分类方法、装置、电子设备及存储介质

技术领域

[0001] 本发明涉及信息处理技术领域,尤其涉及一种信息分类方法、装置、电子设备及存储介质。

背景技术

[0002] 通过服务热线,群众或者顾客可以向相关部门或者公司进行问题反馈,在相关工作人员接收到问题反馈后,会把收到的问题记录在各自的工单系统中。

[0003] 对于接收到的问题记录进行分析,会发现存在多人对某个事件或问题进行了相同的反馈的现象,对于此类问题,通常是通过为每一个事件或者问题设置对应的问题类别进行分类,人为地根据问题类别从所有反馈的问题中找出相同类别的问题后进行跟进处理。

[0004] 通过上述方案提出的根据问题类别分类群众反馈的问题或者事件,对问题分类体系的构建要求高,还需要人工干预,耗费大量的人力和时间,问题分类效率低。

发明内容

[0005] 本发明提供了一种信息分类方法、装置、电子设备及存储介质,以解决信息分类的准确性低和高效性低的问题。

[0006] 根据本发明的一方面,提供了一种信息分类方法,包括:

[0007] 获取待处理反馈信息,待处理反馈信息中包括反馈类型和反馈描述信息;

[0008] 针对于每一反馈类型,基于预先训练的向量转换模型将反馈类型的多个反馈描述信息转换为句向量;基于反馈描述信息的句向量对待处理反馈信息进行聚类处理,得到每一反馈类型的聚类簇;

[0009] 对任意两个反馈类型中的聚类簇进行相似判定,并将满足相似判定条件的聚类簇进行合并,得到目标聚类簇,其中,目标聚类簇中的待处理反馈信息属于同一类信息。

[0010] 可选的,向量转换模型的训练方法包括:

[0011] 构建初始模型,初始模型中包括第一向量转换模型、第二向量转换模型和相似判定模型;获取样本数据,样本数据中包括第一描述信息、第二描述信息和相似度标签;基于样本数据对初始模型进行训练,在满足训练结束条件时,将训练好的第一向量转换模型或第二向量转换模型确定为训练好的向量转换模型,其中,第一向量转换模型和第二向量转换模型在训练过程中权重共享。

[0012] 可选的,第一向量转换模型用于将第一描述信息转换为第一句向量,第二向量转换模型用于将第二描述信息转换为第二句向量,相似判定模型用于生成第一句向量和第二句向量的相似度预测结果;

[0013] 第一向量转换模型和第二向量转换模型分别包括:词向量转换模块、至少两个词向量特征提取模块和特征拼接模块。

[0014] 可选的,基于反馈描述信息的句向量对待处理反馈信息进行聚类处理,得到每一反馈类型的聚类簇,包括:遍历各句向量,分别计算两个句向量之间的相似度数据,并基于

相似度数据确定待处理反馈信息所属的聚类簇。

[0015] 可选的,对任意两个反馈类型中的聚类簇进行相似判定,并将满足相似判定条件的聚类簇进行合并,包括:确定第一反馈类型的第一聚类簇中的第一反馈描述信息和第二反馈类型的第二聚类簇中的第二反馈描述信息;对第一反馈描述信息的句向量和第二反馈描述信息的句向量进行相似度处理,得到相似度数据;基于相似判定条件对相似度数据进行判定,在相似度数据满足相似判定条件的情况下,将第一聚类簇和第二聚类簇进行合并。

[0016] 可选的,方法还包括:设置同类判定阈值,基于同类判定阈值对目标聚类簇中的信息数进行判定;释放不满足同类判定阈值的目标聚类簇。

[0017] 根据本发明的另一方面,提供了一种信息分类装置,包括:

[0018] 待处理反馈信息获取模块,用于获取待处理反馈信息,待处理反馈信息中包括反馈类型和反馈描述信息;

[0019] 聚类簇确定模块,用于针对于每一反馈类型,基于预先训练的向量转换模型将反馈类型的多个反馈描述信息转换为句向量;基于反馈描述信息的句向量对待处理反馈信息进行聚类处理,得到每一反馈类型的聚类簇;

[0020] 目标聚类簇确定模块,用于对任意两个反馈类型中的聚类簇进行相似判定,并将满足相似判定条件的聚类簇进行合并,得到目标聚类簇,其中,目标聚类簇中的待处理反馈信息属于同一类信息。

[0021] 根据本发明的另一方面,提供了一种电子设备,电子设备包括:

[0022] 至少一个处理器;以及

[0023] 与至少一个处理器通信连接的存储器;其中,

[0024] 存储器存储有可被至少一个处理器执行的计算机程序,计算机程序被至少一个处理器执行,以使至少一个处理器能够执行本发明任一实施例的信息分类方法。

[0025] 根据本发明的另一方面,提供了一种计算机可读存储介质,计算机可读存储介质存储有计算机指令,计算机指令用于使处理器执行时实现本发明任一实施例的信息分类方法。

[0026] 本发明实施例的技术方案,通过对待处理反馈信息进行聚类处理,得到每一反馈类型的聚类簇,进而基于聚类簇进行相似度判定,实现了自动将满足相似判定条件的反馈信息合并到同一问题类型,避免了人工干预反馈问题分类的问题,提高了信息分类的准确性和高效性。

[0027] 应当理解,本部分所描述的内容并非旨在标识本发明的实施例的关键或重要特征,也不用于限制本发明的范围。本发明的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0028] 为了更清楚地说明本发明实施例中的技术方案,下面将对实施例描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,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,还可以根据这些附图获得其他的附图。

[0029] 图1是本发明实施例提供的一种信息分类方法的流程图;

[0030] 图2是本发明实施例提供的一种向量转换模型结构示意图;

- [0031] 图3是本发明提供的优选实施例的信息分类方法的流程图；
- [0032] 图4是适用于优选实施例的句向量模型的结构示意图；
- [0033] 图5本发明实施例提供的一种信息分类装置的结构示意图；
- [0034] 图6是实现本发明实施例的信息分类方法的电子设备的结构示意图。

具体实施方式

[0035] 为了使本技术领域的人员更好地理解本发明方案，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的实施例，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，都应当属于本发明保护的范围。

[0036] 需要说明的是，本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本发明的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0037] 图1是本发明实施例提供的一种信息分类方法的流程图，本实施例可适用于信息分类的情况，该方法可以由信息分类装置来执行，该信息分类装置可以采用硬件和/或软件的形式实现，该信息分类装置可配置于计算机等电子设备中。如图1所示，该方法包括：

[0038] S110、获取待处理反馈信息，待处理反馈信息中包括反馈类型和反馈描述信息。

[0039] 其中，待处理反馈信息具体可以理解为是用户反馈的问题或事件的记录数据中的信息，记录数据的内容包含但不限于反馈类别、描述信息等，可以将记录数据存储在数据管理系统或者服务器中。用户反馈的问题内容可根据业务场景确定，其中，业务场景包括但不限于货物交易、物流配送等。待处理反馈信息的获取方式包括但不限于通过客服热线方式获取的用户反馈提取得到，或者，通过订单评价或者售后反馈等信息提取得到。可选的，根据不同业务场景，待处理反馈信息还可以包括业务场景的其他关联信息，以货物交易为例，待处理反馈信息还可以包括交易货物类型、数量等，以物流配送为例，待处理反馈信息还可以包括配送地址、配送订单信息等。具体的，待处理反馈信息的获取方式可以是获取用户反馈数据，提取用户反馈数据的摘要作为描述信息，识别用户反馈数据的反馈类别，形成结构化数据，即待处理反馈信息。可选的，将获取的大量用户反馈数据分发至处理端，接收各处理端发送的用户反馈数据。

[0040] 反馈类型具体可以理解为是对用户反馈数据进行分类而设置的类别，一般是根据用户反馈数据选择的相匹配的类别，以物流配送为例，反馈类型可包含但不限于物流异常、订单异常等，反馈类型是根据实际情况预先设置并存储在系统中的，需要说明的是，如果遇到接收到的用户反馈数据所属的反馈类型不存在的情况，可以直接添加新增反馈类型并存储在系统中。反馈描述信息具体可以理解为是会话摘要，是对反馈问题的主要内容的总结信息，即对反馈类型的简要说明，可通过提取关键信息实现，比如反馈类型为物流反馈，其

对应的反馈描述信息可以为送错地址。

[0041] 具体的,获取待处理反馈信息可以从系统或者服务器中直接读取,还可以通过导入设备将记录反馈信息文本导入到系统中后获取,读取反馈类型和反馈描述信息。

[0042] S120、针对于每一反馈类型,基于预先训练的向量转换模型将反馈类型的多个反馈描述信息转换为句向量。

[0043] S130、基于反馈描述信息的句向量对待处理反馈信息进行聚类处理,得到每一反馈类型的聚类簇。

[0044] 其中,聚类处理具体可以理解为是按照数据对象的内容的相似度将内容分组为多个类或者簇的过程,同类对象间具有较高的相似性,不同类的对象间具有较大的差异性,本实施例中的数据对象指的是待处理反馈信息。通过预设的聚类算法对待处理反馈信息进行聚类,聚类算法包含但不限于K-means算法、Single-Pass增量聚类算法、围绕中心划分(Partitioning Around Mediods,PAM)算法,还可以通过聚类模型对待处理反馈信息进行聚类,此处不作限定。需要理解的是,属于同一反馈类型的待处理反馈信息的反馈描述信息可能不相同,所以,每一反馈类型的待处理反馈信息中可以包含多种反馈描述信息。聚类簇具体可以理解为是由相似的待处理反馈信息所构成的信息集合,可以理解的是,同一类簇内的反馈描述信息彼此相似,与其他类簇中的反馈描述信息相异。

[0045] 具体的,可以根据反馈类型对待处理反馈信息进行粗分类,进一步的将属于同一反馈类型的待处理反馈信息的多个反馈描述信息进行聚类处理,通过文本聚类算法对反馈描述信息进行进一步的聚类处理,进而得到每一反馈类型的聚类簇。

[0046] 示例性的,在物流配送场景下,可以根据物流异常、开具发票、订单异常等反馈类型进行分类,不同的反馈类型记录中存储对应的反馈描述信息,如表1所示的反馈信息记录数据,将表中的记录数据作为待处理反馈信息,待处理反馈信息包含记录编号、反馈类型和反馈描述信息。

[0047] 表1 反馈信息记录

[0048]	记录编号	反馈类型	反馈描述信息
	1	物流异常	反馈超过配送时效了,但是没有收到货物
	2	物流异常	反馈送错站点
	3	开具发票	发票信息填写错误,需要重新修改发票
	4	订单异常	咨询订单为啥一直不发货
	5	开具发票	申请补开发票
	6	物流异常	反馈到时间了,但是货物依然没有配送

[0049] 对于上表中的反馈信息依据反馈类型进行分类之后,发现记录1、记录2、记录6可以归类为反馈类型为“物流异常”,但是继续查看反馈信息描述中的内容,可以发现记录1和记录6反馈的是同样的问题,而记录2与记录1、记录6反馈的是不同的问题,需要进一步通过

聚类算法对反馈类型的多个反馈描述信息进行聚类处理,将相似的反馈信息聚类在一起,得到每一反馈类型的聚类簇。

[0050] 在本实施例中,在根据反馈类型对待处理反馈信息进行分类,分别对每一反馈类型的多个反馈描述信息对待处理反馈信息进行聚类处理,得到每一个反馈类型对应的聚类簇,有助于提高对反馈描述信息进行识别和分类,提高了反馈信息识别的准确性和高效性。

[0051] 具体的,通过预先训练的向量转换模型对每一反馈类型中的反馈描述信息进行处理,分别得到每一反馈描述信息对应的句向量。

[0052] 可选的,向量转换模型的训练方法包括:步骤1:构建初始模型,初始模型中包括第一向量转换模型、第二向量转换模型和相似判定模型;步骤2:获取样本数据,样本数据中包括第一描述信息、第二描述信息和相似度标签;步骤3:基于样本数据对初始模型进行训练,在满足训练结束条件时,将训练好的第一向量转换模型或第二向量转换模型确定为训练好的向量转换模型,其中,第一向量转换模型和第二向量转换模型在训练过程中权重共享。

[0053] 可选的,第一向量转换模型用于将第一描述信息转换为第一句向量,第二向量转换模型用于将第二描述信息转换为第二句向量,相似判定模型用于生成第一句向量和第二句向量的相似度预测结果;第一向量转换模型和第二向量转换模型分别包括:词向量转换模块、至少两个词向量特征提取模块和特征拼接模块。

[0054] 其中,构建的初始模型结构如图2所示,初始模型中包含第一向量转换模型、第二向量转换模型和相似判定模型。样本数据具体可以理解为是反馈描述信息集,可以从系统或者服务器中直接读取反馈记录中的反馈描述信息。对样本数据进行句向量转换,完成对句向量模型的训练。训练结束条件可以预先设置在模型中,比如可以设置将所有的样本数据训练完之后便可结束训练并确定向量转换模型,还可以设置训练结果的准确满足预设的准确阈值便可结束样本训练。词向量转换模块用于将语句进行处理,得到语句中包含的词对应的词向量的模块,可以通过word2vec、fastText、glove、NLP、BERT(Bidirectional Encoder Representation from Transformers,双向Transformer的Encoder)模型等方法进行处理得到语句的词向量。词向量特征提取模块用于提取词向量对应的特征,可以通过词向量的平均值模块、最大值模块、或最小值模块得到,此处不作限定。

[0055] 示例性的,样本数据信息表的形式可以如表2所示,包含第一描述信息、第二描述信息和相似度标签,其中,相似度标签可以理解为是表征第一描述信息和第二描述信息是否相似的标签,可以通过数字0和1进行表示第一描述信息和第二描述信息是否相似,0表示不相似,1表示相似。

[0056] 表2 样本数据信息表

	第一描述信息	第二描述信息	相似度标签
[0057]	反馈超过配送时效了,但是没有收到货物	超过配送时效了,未收到货物	1
	反馈到时间了,但是货物依然没有配送	反馈超过配送时效了,但是没有收到货物	1
	反馈商品存在问题,需要退货	买错商品了,需要退换货	0

[0058] 具体的,根据上述模型训练步骤训练词向量转换模型,构建初始模型,获取反馈记录中的大量反馈描述信息作为训练集,例如可以选取20000条记录作为训练集。将任意两条反馈描述信息作为一条训练记录,分别记录在如表2所示的第一描述信息和第二描述信息列,并将训练结果存储在对应的相似度标签列。将第一描述信息和第二描述信息列中的描述信息作为初始模型的输入参数,分别通过第一向量转换模型和第二向量转换模型对第一描述信息和第二描述信息列中的描述信息进行处理,由向量转换模型中的词向量转换模块处理,得到第一描述信息和第二描述信息对应的词向量,其中,词向量转换可以通过BERT模型进行处理,需要说明的是,各个词向量转换模型中权重处理采用权重共享的方式。由特征提取模块对得到的词向量进行特征提取处理,可以选取至少两种词向量特征提取模块,比如词向量最大值模块、词向量平均值模块等。通过特征拼接模块将由至少两个词向量特征提取模块提取的词向量对应的向量特征进行拼接,例如可以通过CONCAT函数进行拼接,将拼接得到的向量作为句向量,第一向量转换模型得到的是第一句向量,第二向量转换模型得到的是第二句向量。通过相似判定模型对第一句向量和第二句向量进行处理,生成第一句向量和第二句向量的相似度预测结果。

[0059] 进一步的,还可以任意选取一定数量的反馈信息描述数据对作为测试样本,测试向量转换模型的准确性,比如可以选取4000条记录。

[0060] 在本实施例中,通过构建初始模型,基于获取的大量的样本信息进行模型训练,得到预先训练好的向量转换模型,通过向量转换模型对语句进行向量转换,得到句向量,提高了对语句的向量转换速度,有助于提高反馈描述信息的识别和分类效率。

[0061] 通过句向量转换模型对反馈描述信息进行处理得到反馈描述信息的句向量,对反馈描述信息的句向量进行聚类处理,通过聚类算法确定相似度,得到每一反馈类型中的聚类簇,即得到每一种反馈类型中的具有一定相似性的反馈信息描述的分类结果。

[0062] 可选的,遍历各句向量,分别计算两个句向量之间的相似度数据,并基于相似度数据确定待处理反馈信息所属的聚类簇。

[0063] 具体的,遍历由句向量转换模型输出的各个句向量,分别计算两个句向量之间的相似度数据,可以通过计算两个句向量的之间的差值的绝对值,设置相似度阈值,如果差值绝对值小于相似度阈值,可以判定两个句向量相似,反之不相似,将相似的反馈信息归为一类反馈类型,即每一类反馈类型中的反馈描述信息集合作为待处理反馈信息所属的聚类

簇。

[0064] 示例性的,在获得句向量转换模型之后,可以对根据反馈类型进行分类得到不同类型及其对应的反馈描述信息进行处理,生成各自反馈描述信息的句向量。句向量之间的距离采用向量的余弦相似度,规定阈值超过0.85为相似。计算所有句向量两两之间的相似度,然后依次遍历,比如句向量A与句向量B相似度为0.9,句向量A与句向量C相似度为0.6,句向量A与句向量D相似度为0.88,且句向量A、句向量B、句向量D之间两两相似度都大于0.85,则句向量A、句向量B、句向量D为一个类别,最终找出包含最多样本的簇,在这个簇中,因为每个的句向量之间,两两之间的相似度都是大于0.85,所以,句向量A、句向量B、句向量D所属的整个簇中反馈描述信息表达的意图是一致的。也因此可以把这个簇中句向量所代表的反馈信息当作多人同诉事件。

[0065] 在本实施例中,对任意两个句向量进行相似度判定,将相似的反馈描述信息归为一类,即构成一个聚类簇,对所有的句向量完成相似度判定后,得到待处理反馈信息所属的聚类簇,对于同一反馈类型中的反馈信息描述进行相似度判定,提高了反馈描述信息的识别和分类的精确性。

[0066] 本实施例的技术方案,通过对待处理反馈信息进行聚类处理,通过预训练的词向量对反馈描述信息进行词向量转换,对词向量进行相似度判定,得到相似度数据,基于相似度数据确定待处理反馈信息所属的聚类簇,实现了对于描述同一反馈的反馈描述信息合并到同一问题类型,避免了人工干预反馈问题识别和分类的问题,使得根据反馈描述信息进行归类更加地准确,提高了信息分类的准确性和高效性,提高反馈类型识别的参考价值。

[0067] S140、对任意两个反馈类型中的聚类簇进行相似判定,并将满足相似判定条件的聚类簇进行合并,得到目标聚类簇,其中,目标聚类簇中的待处理反馈信息属于同一类信息。

[0068] 其中,相似判定具体可以理解为是对不同聚类簇中的待处理反馈信息进行相似度计算,可以通过比较聚类簇之间的相似度,将相似度较大的两聚类簇中的待处理反馈信息判定为一个类别。通过相似度算法对不同反馈类型中的任意两个聚类簇进行处理,相似度算法可以包含但不限于余弦距离算法、马氏距离算法等。

[0069] 目标聚类簇具体可以理解为是对不同聚类簇进一步进行相似判定并将满足相似判定条件的聚类簇合并得到的集合,不同的目标聚类簇之间不满相似判定条件,即目标聚类簇为无法继续进行合并的聚类簇。其中,相似判定条件是一种相似判定规则,可以通过设置相似度阈值作为相似判定条件,比如设置相似度阈值可以为0.85。

[0070] 具体的,选取任意两种反馈类型中的聚类簇,并分别从聚类簇中任意选取一个反馈信息描述并对其进行相似度判定,相似判定条件为设置相似度阈值,比如设置为0.85,通过比较聚类簇之间的相似度是否大于等于相似度阈值,如果满足,则可以判定两个聚类簇中的反馈信息属于同一种情况,可以将两个聚类簇进行合并处理,反之,保持两个聚类簇独立。将所有的聚类簇进行相似判定处理后,得到目标聚类簇。需要说明的是,目标聚类簇中的待处理反馈信息属于同一类信息。

[0071] 可选的,对任意两个反馈类型中的聚类簇进行相似判定,并将满足相似判定条件的聚类簇进行合并,包括:确定第一反馈类型的第一聚类簇中的第一反馈描述信息和第二反馈类型的第二聚类簇中的第二反馈描述信息;对第一反馈描述信息的句向量和第二反馈

描述信息的句向量进行相似度处理,得到相似度数据;基于相似判定条件对相似度数据进行判定,在相似度数据满足相似判定条件的情况下,将第一聚类簇和第二聚类簇进行合并。

[0072] 其中,第一反馈类型和第二反馈类型中的第一和第二仅是用于区分获取的任意两个反馈类型,并无其他特殊含义,同样地,第一反馈描述信息和第一聚类簇、第二反馈描述信息和第二聚类簇中的第一和第二也是用于区分不同的反馈描述信息和聚类簇,只需要与上述第一反馈类型和第二反馈类型相对应即可。句向量具体可以理解为表征整个句子语义的向量,把句子语义投影到一个N维向量空间上,通过包含但不限于fastText、glove、word2vec等方法对语句进行处理,得到语句包含的各个词向量,并通过求得词向量平均值得到该语句对应的句向量,还可以通过句向量模型进行向量转换处理并得到对应的句向量,此处对于获取句向量的方法不做限定。

[0073] 具体的,遍历选取两个不同类型中的聚类簇,分别作为第一反馈类型的第一聚类簇和第二反馈类型的第二聚类簇,分别在选取的两个聚类簇中任意选取一个反馈描述信息,分别作为第一反馈描述信息和第二反馈描述信息,即得到了第一反馈类型的第一聚类簇中的第一反馈描述信息和第二反馈类型的第二聚类簇中的第二反馈描述信息。通过预训练的句向量转换模型对第一反馈描述信息和第二反馈描述信息进行处理,得到对应的第一句向量和第二句向量,即得到第一反馈描述信息的句向量和第二反馈描述信息的句向量。对上述获得大两个句向量进行相似度处理,可以通过包含但不限于闵科夫斯基距离算法、余弦距离算法、马氏距离算法等算法进行相似度计算,得到两个句向量的相似度数据。通过设置相似度阈值的方法确定相似判定条件,如果相似度数据满足相似度阈值,则可以将第一聚类簇和第二聚类簇进行合并处理,示例性的,可以通过拼接函数直接将第一聚类簇和第二聚类簇进行合并。

[0074] 进一步的,方法还包括:设置同类判定阈值,基于同类判定阈值对目标聚类簇中的信息数进行判定;释放不满足同类判定阈值的目标聚类簇。

[0075] 具体的,可以通过设置同类判定阈值从目标聚类簇中过滤掉包含反馈信息数量较少的聚类簇,比如设置同类判定阈值为10,那么聚类簇对应的反馈类型中的反馈记录数量少于同类判定阈值,那么可以忽略掉该反馈类型即释放掉该类型对应的目标聚类簇。

[0076] 示例性的,在得到目标聚类簇之后,可以把同一目标聚类簇中的反馈信息记录作为一个多人同诉问题,在反馈信息处理过程中,会出现多人同诉问题记录数量较少的情况,可以通过设定同类判定阈值将记录较少的反馈类型过滤掉,得到满足同类判定阈值的目标聚类簇。

[0077] 本实施例的技术方案,通过对待处理反馈信息进行聚类处理,得到每一反馈类型的聚类簇,进而基于聚类簇进行相似度判定,实现了自动将满足相似判定条件的反馈信息合并到同一问题类型,避免了人工干预反馈问题识别和分类的问题,提高了信息分类的准确性和高效性。还引入设置同类判定阈值,释放不满足同类判定阈值的目标聚类簇,有助于过滤掉参考价值掉的反馈类型,提高反馈类型识别的参考价值。

[0078] 在上述实施例的基础上,本发明实施例还提供了一种信息分类方法的优选实例,参见图3,对于待处理反馈信息的处理方法的流程为:

[0079] 步骤1:根据反馈类型对待处理反馈信息进行类别筛选,得到初步的粗筛结果,即多人同诉信息集合。

[0080] 步骤2:先通过预训练句向量模型对粗筛结果中的异常反馈描述信息进行句向量转换,得到各异常反馈描述信息对应的句向量,完成句向量之间的相似度匹配,可以采用向量的余弦相似度算法计算句向量之间的相似度,可以设置相似度阈值为0.85,若相似度大于等于0.85可以判定两个句向量相似,根据同样的算法计算所有句向量两两之间的相似度,然后依次遍历,例如句向量A与句向量B相似度为0.9,句向量A与句向量C相似度为0.6,句向量A与句向量D相似度为0.88,且句向量A、句向量B、句向量D之间两两相似度都大于0.85,则句向量A、句向量B、句向量D为同一种反馈类型,最终找出包含最多样本的簇,在这个簇中,因为两两句向量之间的相似度都是大于0.85,所以整个簇中反馈描述信息的意图是一致的,进而可以把这个簇中的句向量所代表的反馈信息当作一个多人同诉事件。

[0081] 步骤3:扩展多人同诉事件,在步骤2中,通过文本聚类方法找到了相同反馈类型中的多人同诉信息,需要说明的是,在记录反馈信息的时候,会出现问题类型的归纳错误,可能会有其他的反馈相同的问题被归到了别的反馈类型,因此需要召回这些归类错误的反馈信息一并归入至正确的多人同诉中;扩展多人同诉事件的具体方法可以通过遍历其他的反馈类型的反馈信息,并将其他的反馈类型的反馈描述信息生成句向量,然后从多人同诉事件中选一个反馈描述信息与其他反馈类型的反馈描述信息对应的句向量进行相似度判断,如果两者相似度大于0.9,则将其他的反馈类型的反馈描述信息归入到多人同诉事件中。进一步查找全部反馈类型的多人同诉事件,得到准确的多人同诉事件的集合。

[0082] 步骤4:设置数量阈值对反馈信息数量进行过滤,需要说明的是,有的反馈类型对应的反馈记录较少,可以设置数量阈值将这种反馈类型的反馈记录进行过滤,例如:在一天的反馈记录中,如果某种反馈类型的反馈记录少于10,那么此种反馈类型及其对应的反馈记录可以忽略,最终得到满足数量阈值的全部的多人同诉事件。将结果返回。

[0083] 其中,在上述优选实施例的步骤2中使用的句向量模型的具体结构图如图4所示,包含词向量提取模块、词向量平均值模块、词向量最大值模块、词向量拼接模块、句向量确定模块,其中词向量提取模块中可以采用BERT模型完成语句中的词向量的提取,词向量平均值模块和词向量最大值模块用于计算词向量提取模块输出的词向量的最大值和平均值,词向量拼接模块可以采用CONCAT函数完成词向量的平均值和最大值的拼接,句向量确定模块可以采用Embedding模型对词向量拼接模块输出的结果进行处理,得到语句对应的句向量,例如将语句A和语句B输入至句向量模型中,分别对语句A和语句B进行处理,得到对应的句向量A和句向量B,然后根据相似度判定模块计算句向量A和句向量B的相似度,并通过分类模块将相似度满足相似度阈值的句向量分为一类。需要说明的是,在对语句A和语句B进行词向量提取的时候,分别采用的BERT模型的权重是共享的。

[0084] 图5是本发明实施例提供的一种信息分类装置的结构示意图。如图5所示,该装置包括:

[0085] 待处理反馈信息获取模块310,用于获取待处理反馈信息,待处理反馈信息中包括反馈类型和反馈描述信息;

[0086] 聚类簇确定模块320,用于针对于每一反馈类型,基于预先训练的向量转换模型将反馈类型的多个反馈描述信息转换为句向量;基于反馈描述信息的句向量对待处理反馈信息进行聚类处理,得到每一反馈类型的聚类簇;

[0087] 目标聚类簇确定模块330,用于对任意两个反馈类型中的聚类簇进行相似判定,并

将满足相似判定条件的聚类簇进行合并,得到目标聚类簇,其中,目标聚类簇中的待处理反馈信息属于同一类信息。

[0088] 可选的,该装置还包括:模型训练模块,用于:

[0089] 构建初始模型,初始模型中包括第一向量转换模型、第二向量转换模型和相似判定模型;

[0090] 获取样本数据,样本数据中包括第一描述信息、第二描述信息和相似度标签;

[0091] 基于样本数据对初始模型进行训练,在满足训练结束条件时,将训练好的第一向量转换模型或第二向量转换模型确定为训练好的向量转换模型,其中,第一向量转换模型和第二向量转换模型在训练过程中权重共享。

[0092] 第一向量转换模型用于将第一描述信息转换为第一句向量,第二向量转换模型用于将第二描述信息转换为第二句向量,相似判定模型用于生成第一句向量和第二句向量的相似度预测结果;

[0093] 第一向量转换模型和第二向量转换模型分别包括:词向量转换模块、至少两个词向量特征提取模块和特征拼接模块。

[0094] 可选的,聚类簇确定模块320还用于:

[0095] 遍历各句向量,分别计算两个句向量之间的相似度数据,并基于相似度数据确定待处理反馈信息所属的聚类簇。

[0096] 可选的,目标聚类簇确定模块330,具体用于:

[0097] 对任意两个反馈类型中的聚类簇进行相似判定,并将满足相似判定条件的聚类簇进行合并,包括:

[0098] 确定第一反馈类型的第一聚类簇中的第一反馈描述信息和第二反馈类型的第二聚类簇中的第二反馈描述信息;

[0099] 对第一反馈描述信息的句向量和第二反馈描述信息的句向量进行相似度处理,得到相似度数据;

[0100] 基于相似判定条件对相似度数据进行判定,在相似度数据满足相似判定条件的情况下,将第一聚类簇和第二聚类簇进行合并。

[0101] 该装置还包括:目标聚类簇判定模块,用于:设置同类判定阈值,基于同类判定阈值对目标聚类簇中的信息数进行判定;释放不满足同类判定阈值的目标聚类簇。

[0102] 本发明实施例所提供的信息分类装置可执行本发明任意实施例所提供的信息分类方法,具备执行方法相应的功能模块和有益效果。

[0103] 图6是本发明实施例提供的一种电子设备的结构示意图。电子设备10旨在表示各种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备(如头盔、眼镜、手表等)和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的和/或者要求的本发明的实现。

[0104] 如图6所示,电子设备10包括至少一个处理器11,以及与至少一个处理器11通信连接的存储器,如只读存储器(ROM)12、随机访问存储器(RAM)13等,其中,存储器存储有可被至少一个处理器执行的计算机程序,处理器11可以根据存储在只读存储器(ROM)12中的计

算机程序或者从存储单元18加载到随机访问存储器(RAM)13中的计算机程序,来执行各种适当的动作和处理。在RAM 13中,还可存储电子设备10操作所需的各种程序和数据。处理器11、ROM 12以及RAM 13通过总线14彼此相连。输入/输出(I/O)接口15也连接至总线14。

[0105] 电子设备10中的多个部件连接至I/O接口15,包括:输入单元16,例如键盘、鼠标等;输出单元17,例如各种类型的显示器、扬声器等;存储单元18,例如磁盘、光盘等;以及通信单元19,例如网卡、调制解调器、无线通信收发机等。通信单元19允许电子设备10通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0106] 处理器11可以是各种具有处理和计算能力的通用和/或专用处理组件。处理器11的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的处理器、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。处理器11执行上文所描述的各个方法和处理,例如信息分类方法。

[0107] 在一些实施例中,信息分类方法可被实现为计算机程序,其被有形地包含于计算机可读存储介质,例如存储单元18。在一些实施例中,计算机程序的部分或者全部可以经由ROM 12和/或通信单元19而被载入和/或安装到电子设备10上。当计算机程序加载到RAM 13并由处理器11执行时,可以执行上文描述的信息分类方法的一个或多个步骤。备选地,在其他实施例中,处理器11可以通过其他任何适当的方式(例如,借助于固件)而被配置为执行信息分类方法。

[0108] 本文中以上描述的系统和技术各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0109] 用于实施本发明的信息分类方法的计算机程序可以采用一个或多个编程语言的任何组合来编写。这些计算机程序可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器,使得计算机程序当由处理器执行时使流程图和/或框图中所规定的功能/操作被实施。计算机程序可以完全在机器上执行、部分地在机器上执行,作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0110] 本发明实施例还提供了一种计算机可读存储介质,计算机可读存储介质存储有计算机指令,计算机指令用于使处理器执行一种信息分类方法,该方法包括:

[0111] 获取待处理反馈信息,待处理反馈信息中包括反馈类型和反馈描述信息;

[0112] 针对于每一反馈类型,基于预先训练的向量转换模型将反馈类型的多个反馈描述信息转换为句向量;基于反馈描述信息的句向量对待处理反馈信息进行聚类处理,得到每一反馈类型的聚类簇;

[0113] 对任意两个反馈类型中的聚类簇进行相似判定,并将满足相似判定条件的聚类簇进行合并,得到目标聚类簇,其中,目标聚类簇中的待处理反馈信息属于同一类信息。

[0114] 在本发明的上下文中,计算机可读存储介质可以是有形的介质,其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的计算机程序。计算机可读存储介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备,或者上述内容的任何合适组合。备选地,计算机可读存储介质可以是机器可读信号介质。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0115] 为了提供与用户的交互,可以在电子设备上实施此处描述的系统和技术,该电子设备具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给电子设备。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0116] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)、区块链网络和互联网。

[0117] 计算系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器,又称为云计算服务器或云主机,是云计算服务体系中的一项主机产品,以解决了传统物理主机与VPS服务中,存在的管理难度大,业务扩展性弱的缺陷。

[0118] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本发明中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本发明的技术方案所期望的结果,本文在此不进行限制。

[0119] 上述具体实施方式,并不构成对本发明保护范围的限制。本领域技术人员应该明白的是,根据设计要求和因素,可以进行各种修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等,均应包含在本发明保护范围之内。

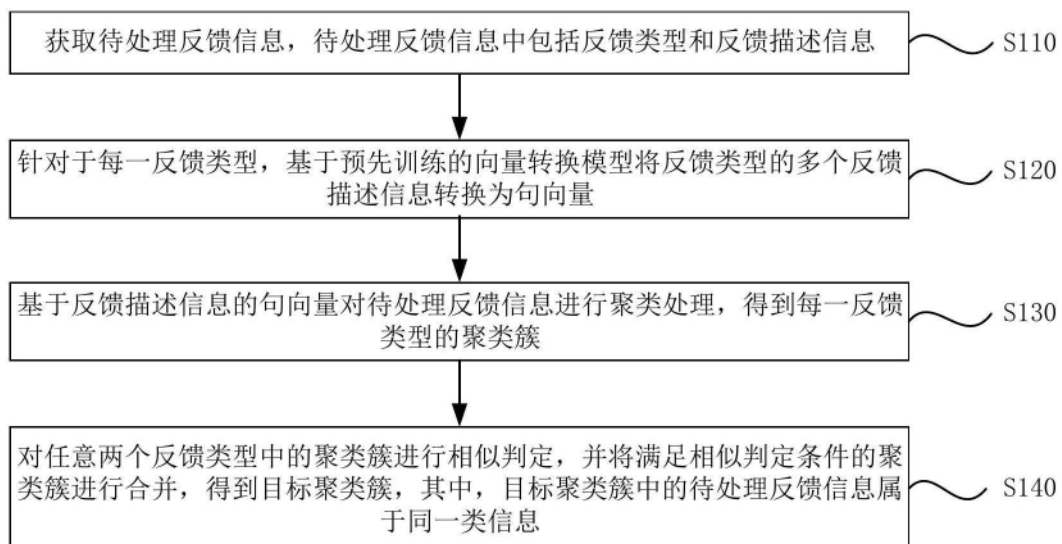


图1

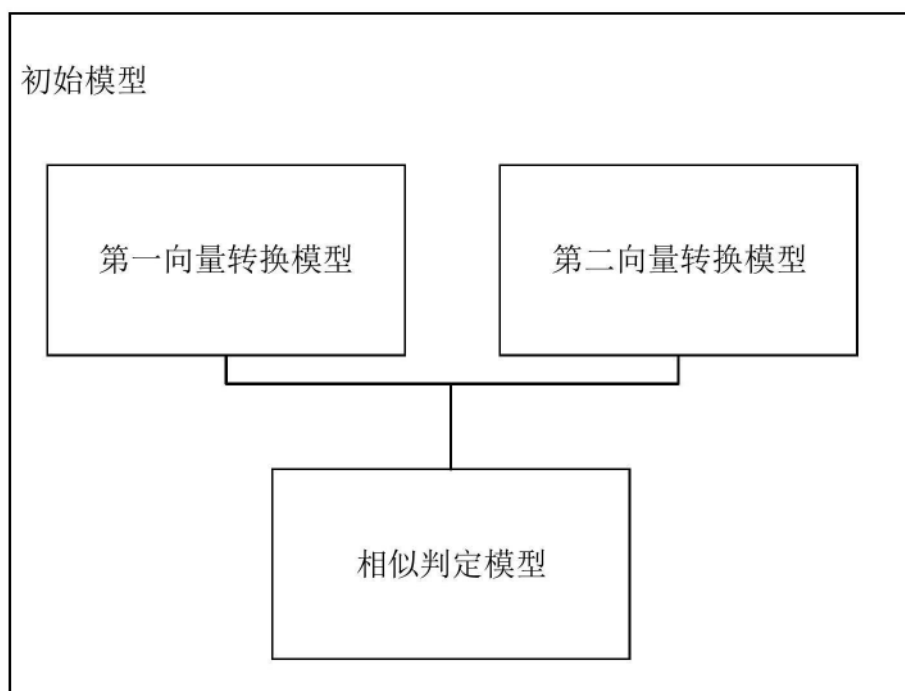


图2

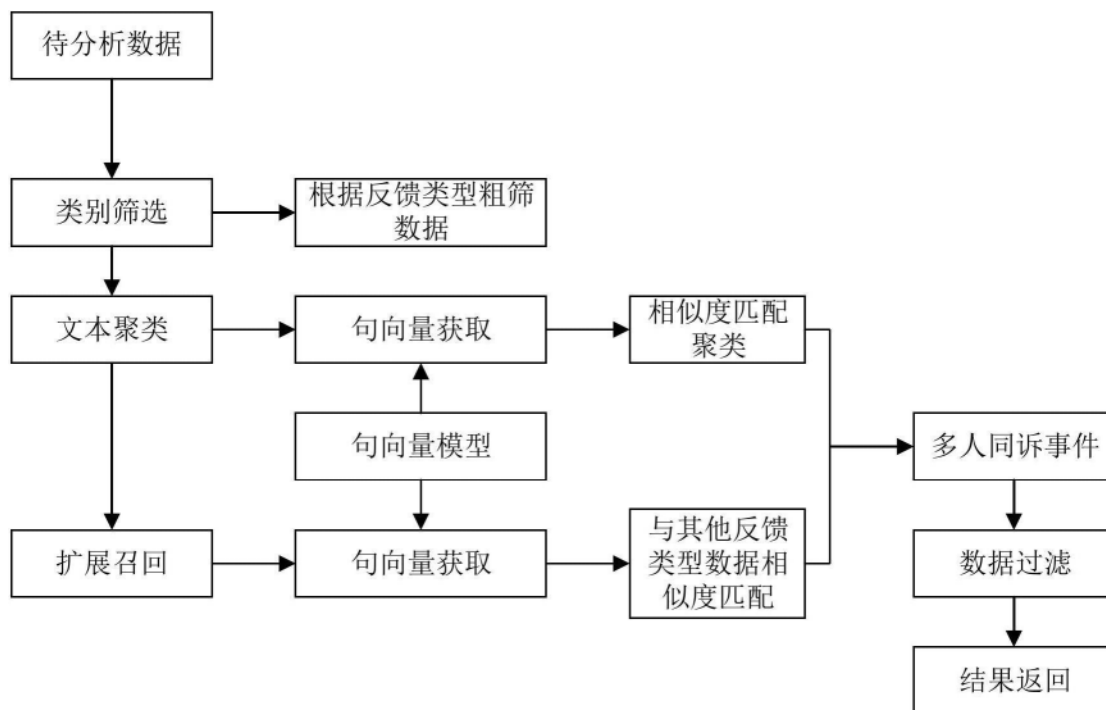


图3

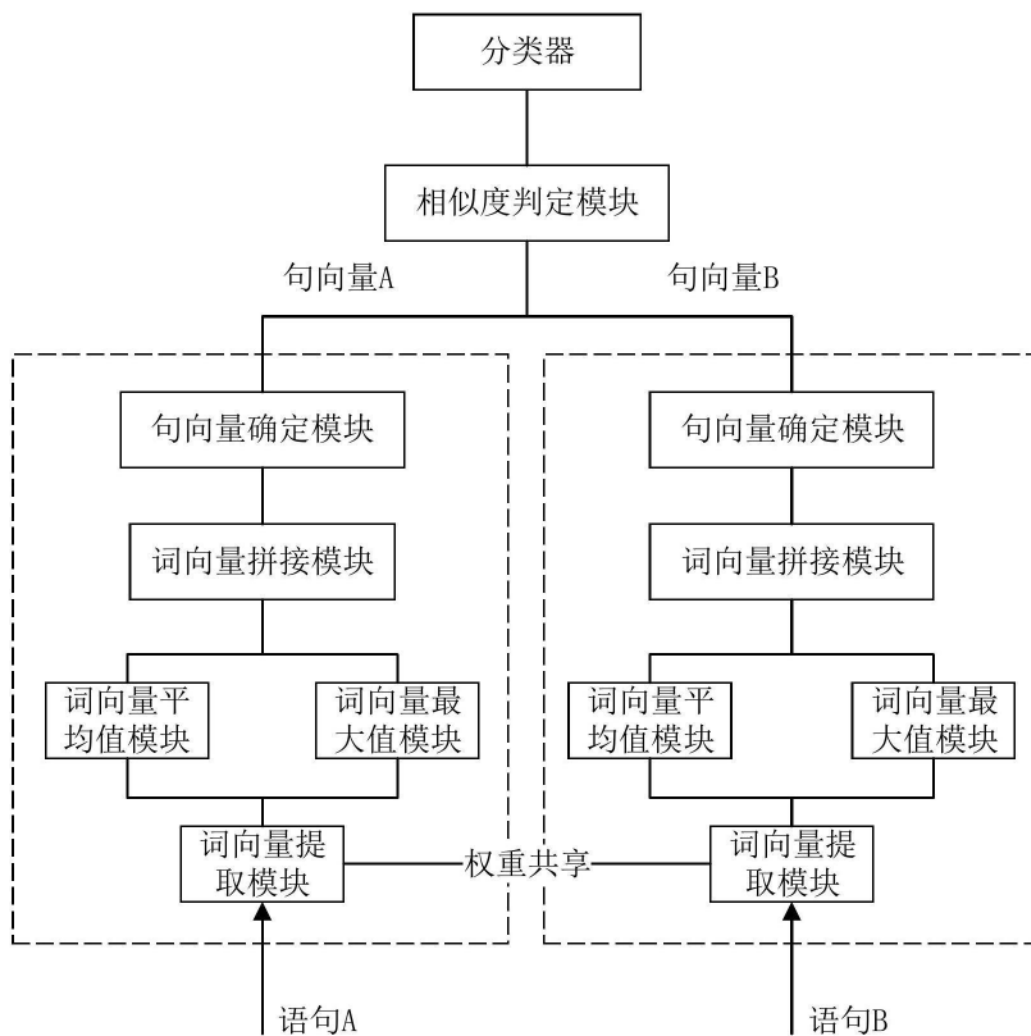


图4

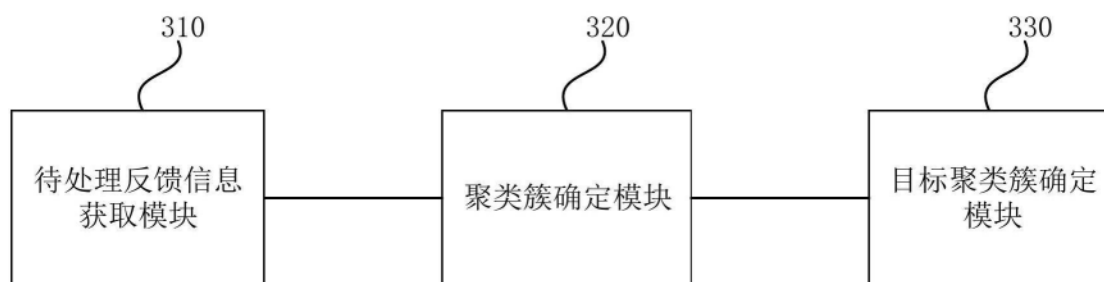


图5

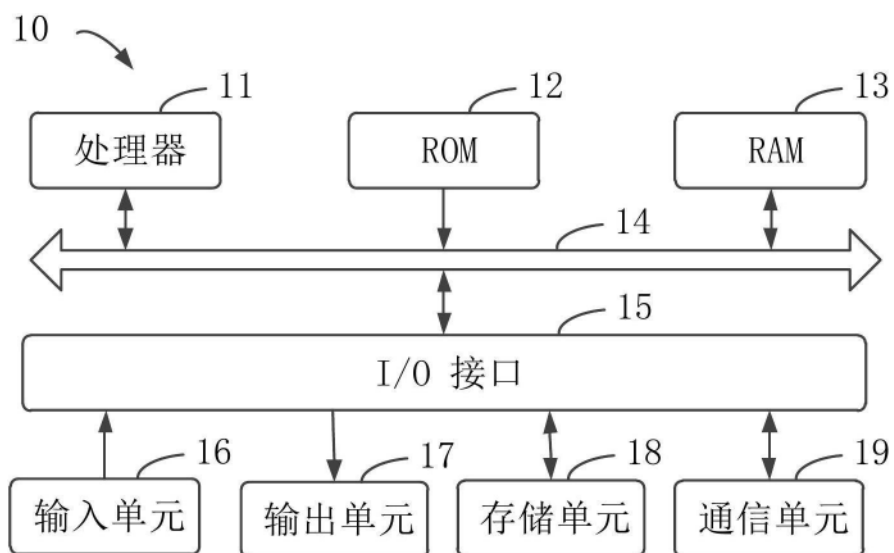


图6