(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117493571 A
(43)申请公布日 2024.02.02

(21)申请号 202311516555.7

(22)申请日 2023.11.14

(71)申请人 京东科技信息技术有限公司
地址 100176 北京市大兴区北京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 张振羽 宋阳 陈蒙

(74)专利代理机构 北京世辉律师事务所 16093
专利代理师 罗利娜

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06F 18/214(2023.01)
G06F 18/24(2023.01)

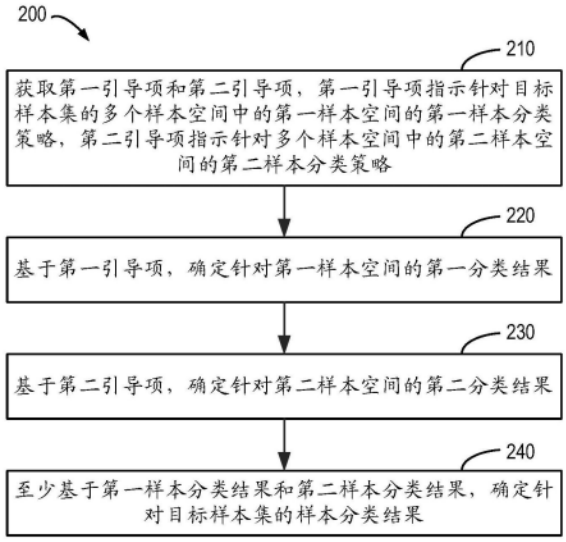
权利要求书2页 说明书12页 附图5页

(54)发明名称

用于分类的方法、装置、设备和可读介质

(57)摘要

本公开的实施例提供用于分类的方法、装置、设备和可读介质。该方法包括:获取第一引导项和第二引导项,第一引导项指示针对目标样本集的多个样本空间中的第一样本空间的第一样本分类策略,第二引导项指示针对多个样本空间中的第二样本空间的第二样本分类策略;基于第一引导项,确定针对第一样本空间的第一分类结果;基于第二引导项,确定针对第二样本空间的第二分类结果;以及至少基于第一样本分类结果和第二样本分类结果,确定针对目标样本集的样本分类结果。以此方式,可以通过指示用户分类策略的引导项对样本进行分类,并且可以自动确定分类结果,无需人力成本,且对于意图分类体系的构建效率较高。



CN 117493571 A

1.一种用于分类的方法,包括:

获取第一引导项和第二引导项,所述第一引导项指示针对目标样本集的多个样本空间中的第一样本空间的第一样本分类策略,所述第二引导项指示针对所述多个样本空间中的第二样本空间的第二样本分类策略;

基于所述第一引导项,确定针对所述第一样本空间的第一分类结果;

基于所述第二引导项,确定针对所述第二样本空间的第二分类结果;以及

至少基于所述第一分类结果和所述第二分类结果,确定针对所述目标样本集的样本分类结果。

2.根据权利要求1所述的方法,其中所述目标样本集包括多个子样本集,所述多个样本空间是基于对所述多个子样本集进行采样所得到的。

3.根据权利要求2所述的方法,其中所述多个样本空间基于以下过程被确定:

确定所述多个子样本集中各个子样本集之间的相似度;以及

通过将相似度高于阈值的至少两个子样本集进行合并,确定所述多个样本空间。

4.根据权利要求3所述的方法,其中将相似度高于阈值的至少两个子样本集进行合并使得所述多个子样本集的数目从第一数目改变为第二数目,所述第一数目大于所述第二数目。

5.根据权利要求1所述的方法,其中所述第一样本分类策略和/或所述第二样本分类策略至少指示待分类的类别数目。

6.根据权利要求1所述的方法,其中确定针对所述目标样本集的样本分类结果包括:

确定针对所述目标样本集的目标分类数目;以及

基于所述目标分类数目,至少对所述第一分类结果和所述第二分类结果进行聚类以确定针对所述目标样本集的样本分类结果。

7.根据权利要求6所述的方法,其中所述目标样本集的样本分类结果包括目标分类,所述目标分类包括针对目标样本空间的分类结果的至少部分,并且所述方法还包括:

获取针对所述目标分类的目标引导项;以及

向模型提供所述目标引导项,以指示所述模型基于所述目标引导项确定所述目标分类的描述信息,所述描述信息包括所述目标分类的标识和/或所述目标分类的代表样本。

8.一种用于分类的装置,包括:

引导项获取模块,被配置为获取第一引导项和第二引导项,所述第一引导项指示针对目标样本集的多个样本空间中的第一样本空间的第一样本分类策略,所述第二引导项指示针对所述多个样本空间中的第二样本空间的第二样本分类策略;

第一结果确定模块,被配置为基于所述第一引导项,确定针对所述第一样本空间的第一分类结果;

第二结果确定模块,被配置为基于所述第二引导项,确定针对所述第二样本空间的第二分类结果;以及

分类结果确定模块,被配置为至少基于所述第一分类结果和所述第二分类结果,确定针对所述目标样本集的样本分类结果。

9.一种电子设备,包括:

至少一个处理单元;以及

至少一个存储器,所述至少一个存储器被耦合到所述至少一个处理单元并且存储用于由所述至少一个处理单元执行的指令,所述指令在由所述至少一个处理单元执行时使所述电子设备执行根据权利要求1至7任一项所述的方法。

10.一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序可由处理器执行以实现根据权利要求1至7任一项所述的方法。

用于分类的方法、装置、设备和可读介质

技术领域

[0001] 本公开的示例实施例总体涉及计算机技术领域,并且更具体地,涉及用于分类的方法、装置、设备和计算机可读存储介质。

背景技术

[0002] 随着信息技术的飞速发展,越来越多的应用或平台等均提供问答功能,给广大用户带来了诸多便利。具有问答功能的应用或平台可以基于智能对话机器人(Intelligent Robot)向用户提供问答服务。在进行对话中,智能对话机器人首先需要对用户的提问进行理解,然后基于理解的用户意图采取对应的应答策略。人们期望能够提高对对话进行意图理解的准确性。

发明内容

[0003] 在本公开的第一方面,提供一种用于分类的方法。该方法包括:获取第一引导项和第二引导项,第一引导项指示针对目标样本集的多个样本空间中的第一样本空间的第一样本分类策略,第二引导项指示针对多个样本空间中的第二样本空间的第二样本分类策略;基于第一引导项,确定针对第一样本空间的第一分类结果;基于第二引导项,确定针对第二样本空间的第二分类结果;以及至少基于第一样本分类结果和第二样本分类结果,确定针对目标样本集的样本分类结果。

[0004] 在本公开的第二方面,提供一种用于分类的装置。该装置包括:引导项获取模块,被配置为获取第一引导项和第二引导项,第一引导项指示针对目标样本集的多个样本空间中的第一样本空间的第一样本分类策略,第二引导项指示针对多个样本空间中的第二样本空间的第二样本分类策略;第一结果确定模块,被配置为基于第一引导项,确定针对第一样本空间的第一分类结果;第二结果确定模块,被配置为基于第二引导项,确定针对第二样本空间的第二分类结果;以及分类结果确定模块,被配置为至少基于第一样本分类结果和第二样本分类结果,确定针对目标样本集的样本分类结果。

[0005] 在本公开的第三方面,提供了一种电子设备。该电子设备包括至少一个处理单元;以及至少一个存储器,至少一个存储器被耦合到至少一个处理单元并且存储用于由至少一个处理单元执行的指令,指令在由至少一个处理单元执行时使电子设备执行本公开第一方面的方法。

[0006] 在本公开的第四方面,提供了一种计算机可读存储介质。该计算机可读存储介质上存储有计算机程序,其可由处理器执行以执行根据本公开的第一方面的方法。

[0007] 应当理解,发明内容部分中所描述的内容并非旨在限定本公开的实施例的关键特征或重要特征,也不用于限制本公开的范围。本公开的其它特征将通过以下的描述而变得容易理解。

附图说明

[0008] 在下文中,结合附图并参考以下详细说明,本公开各实现方式的上述和其他特征、优点及方面将变得更加明显。在附图中,相同或相似的附图标记表示相同或相似的元素,其中:

[0009] 图1示出了本公开的实施例能够在其中实现的示例环境的示意图;

[0010] 图2示出了根据本公开的一些实施例的用于分类的过程的流程图;

[0011] 图3A示出了根据本公开的一些实施例的第一引导项的示例的示意图;

[0012] 图3B示出了根据本公开的一些实施例的第一分类结果的示例的示意图;

[0013] 图3C示出了根据本公开的一些实施例的目标引导项的示例的示意图;

[0014] 图3D示出了根据本公开的一些实施例的针对目标分类的分类信息的示例的示意图;

[0015] 图4示出了根据本公开的一些实施例的分类的流程的示例的示意图;

[0016] 图5示出了根据本公开的一些实施例的用于分类的装置的示意性结构框图;以及

[0017] 图6示出了其中可以实施本公开的一个或多个实施例的计算设备的框图。

具体实施方式

[0018] 下面将参照附图更详细地描述本公开的实施例。虽然附图中示出了本公开的一些实施例,然而应当理解的是,本公开可以通过各种形式来实现,而且不应该被解释为限于这里阐述的实施例,相反,提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是,本公开的附图及实施例仅用于示例性作用,并非用于限制本公开的保护范围。

[0019] 在本公开的实施例的描述中,术语“包括”及其类似用语应当理解为开放性包含,即“包括但不限于”。术语“基于”应当理解为“至少部分地基于”。术语“一个实施例”或“该实施例”应当理解为“至少一个实施例”。术语“一些实施例”应当理解为“至少一些实施例”。下文还可能包括其它明确的和隐含的定义。

[0020] 如本文中所使用的,术语“模型”可以从训练数据中学习到相应的输入与输出之间的关联,从而在训练完成后可以针对给定的输入,生成对应的输出。模型的生成可以基于机器学习技术。深度学习是一种机器学习算法,通过使用多层处理单元来处理输入和提供相应输出。在本文中,“模型”也可以被称为“机器学习模型”、“机器学习网络”、“神经网络”或“网络”,这些术语在本文中可互换地使用。

[0021] “神经网络”是一种基于深度学习的机器学习网络。神经网络能够处理输入并且提供相应输出,其通常包括输入层和输出层以及在输入层与输出层之间的一个或多个隐藏层。在深度学习应用中使用的神经网络通常包括许多隐藏层,从而增加网络的深度。神经网络的各个层按顺序相连,从而前一层的输出被提供作为后一层的输入,其中输入层接收神经网络的输入,而输出层的输出作为神经网络的最终输出。神经网络的每个层包括一个或多个节点(也称为处理节点或神经元),每个节点处理来自上一层的输入。

[0022] 通常,机器学习大致可以包括三个阶段,即训练阶段、测试阶段和应用阶段(也称为推理阶段)。在训练阶段,给定的模型可以使用大量的训练数据进行训练,不断迭代更新参数值,直到模型能够从训练数据中获得一致的满足预期目标的推理。通过训练,模型可以被认为能够从训练数据中学习从输入到输出之间的关联(也称为输入到输出的映射)。训练

后的模型的参数值被确定。在测试阶段,将测试输入应用到训练后的模型,测试模型是否能够提供正确的输出,从而确定模型的性能。在应用阶段,模型可以被用于基于训练得到的参数值,对实际的输入进行处理,确定对应的输出。

[0023] 需要说明的是,本公开的技术方案中,所涉及的用户个人信息的获取,存储和应用等,均符合相关法律法规的规定,且不违背公序良俗。

[0024] 可以理解的是,在使用本公开各实施例公开的技术方案之前,均应当根据相关法律法规通过适当的方式对本公开所涉及个人信息的类型、使用范围、使用场景等告知用户并获得用户的授权。

[0025] 例如,在响应于接收到用户的主动请求时,向用户发送提示信息,以明确地提示用户,其请求执行的操作将需要获得和使用到用户的个人信息,从而使得用户可以根据提示信息来自主地选择是否向执行本公开技术方案的操作的电子设备、应用程序、服务器或存储介质等软件或硬件提供个人信息。

[0026] 作为一种可选的但非限制性的实现方式,响应于接收到用户的主动请求,向用户发送提示信息的方式,例如可以是弹出窗口的方式,弹出窗口中可以以文字的方式呈现提示信息。此外,弹出窗口中还可以承载供用户选择“同意”或“不同意”向电子设备提供个人信息的选择控件。

[0027] 如上文所讨论的,具有问答功能的应用或平台可以基于智能对话机器人向用户提供问答服务。在进行对话中,智能对话机器人首先需要对用户的提问进行理解,然后基于理解的用户意图采取对应的应答策略。因此,用户意图识别模型是智能对话机器人中一个非常重要的核心模块。但是意图分类体系的构建通常需要人工介入从数据中总结分析归类。此外,不同的对话场景中的意图分类体系往往都不相同,例如旅游行程咨询场景主要涉及车票、门票等问题的意图分类体系,而电商购物场景又涉及发货快递、三包政策等问题的意图体系。意图分类体系会随着对话数据而变化,经常需要更改,这会导致大量重复且高额的人力劳动。传统上往往利用文本聚类算法,通过将对话数据聚类,构建多个簇,然后自动抽取一些辅助提示信息或者人工观察每个簇的数据来总结意图分类体系。

[0028] 传统上的意图分类体系构建方法包括纯人工方案和半自动方案。纯人工方案需要大量的领域专家分析很多真实的场景数据,进而基于人工的经验总结出一套分类体系,成本十分昂贵。半自动方案主要基于文本编码与聚类组合的模式,进行分类体系的半自动化总结。半自动方案首先使用一些常见的文本编码模型将文本转化为向量。然后,编码后的向量被使用聚类算法来进行聚类,以形成多个簇。然后通过对每个簇的内容进行抽取分析,再人工总结出最终的分类体系。目前常见的聚类算法包括K-means算法、AP谱聚类算法、基于深度对比学习的聚类算法等。这些聚类算法的聚类划分粒度主要依靠一些超参数控制,这会导致粒度粗细难以和具体场景下的分类体系构建需求对齐。一方面,这些算法都没有对齐人们对应用场景的业务需求,导致生成的知识分类体系和世界业务需求差距较大,需要大量人工参与总结。另一方面,受限于基础技术的水平,当前无法直接基于聚类结果为每个知识点生成一个名称和定义说明,需要人工辅助做大量的分类定义。这都会使得意图分类体系的构建需要耗费大量人力成本,且构建效率较差。

[0029] 有鉴于此,本公开的实施例提供了一种用于分类的方法。该方法包括:获取指示针对目标样本集的多个样本空间中的第一样本空间的第一样本分类策略的第一引导项和指

示针对多个样本空间中的第二样本空间的第二样本分类策略的第二引导项。基于第一引导项,确定针对第一样本空间的第一分类结果。基于第二引导项,确定针对第二样本空间的第二分类结果。至少基于第一样本分类结果和第二样本分类结果,确定针对目标样本集的样本分类结果。以此方式,本公开的实施例可以通过指示用户分类策略的引导项对样本进行分类,并且可以自动确定分类结果,无需人力成本,且对于意图分类体系的构建效率较高。

[0030] 图1示出了本公开的实施例能够在其中实现的示例环境100的示意图。如图1所示,环境100可以包括电子设备110。

[0031] 电子设备110可以确定样本集合102对应的分类结果112。样本集合102可以是存储在电子设备110本地的文本,也可以是电子设备110实时获取到的文本。样本集合102例如可以为问题样本集合。在这种情况下,样本集合102中包括用户在不同场景下的多个问题样本。可以理解,样本集合102还可以是对话样本集合、文本样本集合等等,本公开对此不做限定。下文为了方便表述,将以样本集合102为问题样本集合为例进行示例性说明。

[0032] 在一些实施例中,电子设备110获取样本集合102后,可以基于样本集合102中所包括的多个问题各自对应的用户意图,对多个问题进行聚类以得到分类结果112。分类结果112中不同类别的问题所指示的用户意图不同。电子设备110可以通过任意适当的方式确定分类结果,例如通过聚类算法、通过经训练的机器学习模型等等。

[0033] 电子设备110可以包括具有计算能力的任何计算系统,例如各种计算设备/系统、终端设备、服务端设备等。终端设备可以是任意类型的移动终端、固定终端或便携式终端,包括移动手机、台式计算机、膝上型计算机、笔记本计算机、上网本计算机、平板计算机、媒体计算机、多媒体平板、掌上电脑、便携式游戏终端、VR/AR设备、个人通信系统(Personal Communication System,PCS)设备、个人导航设备、个人数字助理(Personal Digital Assistant,PDA)、音频/视频播放器、数码相机/摄像机、定位设备、电视接收器、无线电广播接收器、电子书设备、游戏设备或者前述各项的任意组合,包括这些设备的配件和外设或者其任意组合。服务端设备可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络、以及大数据和人工智能平台等基础云计算服务的云服务器。服务端设备例如可以包括计算系统/服务器,诸如大型机、边缘计算节点、云环境中的计算设备,等等。

[0034] 应当理解,仅出于示例性的目的描述环境100中各个元素的结构和功能,而不暗示对于本公开的范围的任何限制。

[0035] 以下将继续参考附图描述本公开的一些示例实施例。

[0036] 图2示出了根据本公开的一些实施例的用于分类的过程200的流程图。过程200可以被实现在电子设备110处。为便于讨论,将参考图1的环境100来描述过程200。

[0037] 在框210,电子设备110获取第一引导项和第二引导项,第一引导项指示针对目标样本集的多个样本空间中的第一样本空间的第一样本分类策略,第二引导项指示针对多个样本空间中的第二样本空间的第二样本分类策略。目标样本集例如可以为问题样本集,其包括大量的问题样本。问题样本例如可以为“我的电脑为什么变卡了”、“我的电脑打游戏发烫”等等。这里的第一样本空间与第二样本空间可以是多个样本空间中任意的两个不同的样本空间。

[0038] 这里的多个样本空间可以是电子设备110基于目标样本集所自行确定的,也可以是其他设备确定好进而发送给电子设备110的。为方便表述,下文将以电子设备110自行确定多个样本空间为例进行示例性说明。在一些实施例中,电子设备110获取到目标样本集后,可以首先确定目标样本集所包括的多个子样本集,进而基于这多个子样本集,确定多个样本空间。

[0039] 在一些实施例中,电子设备110可以先通过经训练的语言模型来对目标样本集所包括的多个问题样本进行文本语义辅助分析,模型可以输出每个问题样本对应的文本语义表征。在一些实施例中,为保证聚类的准确性,电子设备110还可以利用经训练的机器学习模型(例如SimCSE模型),将每个问题样本编码为一个固定维度的向量。

[0040] 电子设备110进而可以基于每个问题样本对应的文本语义表征,利用聚类算法对目标样本集中的样本进行第一次聚类以确定目标样本集所包括的多个子样本集。第一次聚类的目的是将样本按照语义进行粗分类,以方便后续基于语义进行分层采样。这里的聚类算法包括但不限于K-means算法、AP谱聚类算法、基于深度对比学习的聚类算法等等。示例性的,电子设备110可以利用K-means算法对目标样本集所包括的多个样本进行聚类,聚类结果即为目标样本集所包括的多个子样本集。

[0041] 在一些实施例中,为保证最终确定的分类结果的准确性,这里的多个子样本集的数目应大于用户最终期望的分类数目。分类数目也可以被称之为聚类簇数目。若第一次聚类时的聚类簇数目为K1,用户最终期望得到的聚类簇数目为M(可以基于具体应用场景确定任意适当的数目),则K1应该在1.5*M到2M之间。例如,若用户期望将目标样本划分为10个分类,则在第一次聚类时,用户可以向电子设备110输入相应指令以指示电子设备110将目标样本集中的样本聚类成15个聚类簇(也即将目标样本集划分成15个子样本集)。

[0042] 电子设备110进而可以计算多个聚类簇中各个聚类簇的中心点的相似度以构造一个以聚类中心点为节点,节点之间的相似度为边的邻接图。电子设备110进而可以基于这个临接图得到语义独立子空间。具体而言,电子设备110可以采用广度优先算法(BFS)算法从第一个节点出发,将相似度大于阈值的节点都囊括到一个子图里。电子设备110进而针对剩余节点不断重复这个过程,最终将所有节点都分配到各自子图里。每个子图的聚类中心点及其对应簇的问题样本就形成了一个个语义相对而言比较独立的子空间。由于这些子空间依旧有很多样本数据,为了满足语言模型进行处理时的输入长度需求,电子设备110还可以对每个子空间进行一定次数(例如S次)不放回的随机采样,每次获取第二数目的(例如D个)样本。由此,电子设备110可以为每个语义独立的子空间构建了一个包含S组含有D个样本的代表数据集。

[0043] 简单来说,电子设备110可以确定多个子样本中各个子样本集之间的相似度,并通过将相似度高于阈值(例如0.7或者高于0.7且超过1的任意数值)的至少两个子样本集进行合并以更新多个子样本集。电子设备110进而可以基于更新后的多个子样本集,确定多个样本空间。可以理解,将相似度高于阈值的至少两个子样本集进行合并可以使得多个子样本集的数目从第一数目改变为第二数目,第一数目大于第二数目。例如,更新前存在5个子样本集,则更新后可能仅存在3个子样本集。电子设备110进而可以对更新后的多个子样本集进行多次(例如5次)不放回的随机采样以确定多个样本空间,每次采样的数目可以为任意适当的数目(例如80-100的任意数目)。具体地,电子设备110对更新后的多个子样本集的各

个子样本集分别进行采样以从子样本集中获取一定数据的样本,进而基于采样得到的样本构建多个样本空间。多个样本空间与更新后的多个子样本集是一一对应的关系。需要注意的是,这里的多个样本空间中不同样本空间彼此语义独立,也可以理解为不同样本空间中的样本的应用场景不同。在一些实施例中,在确定多个样本空间时,电子设备110还可以确定多个样本空间各自的应用场景。例如,第一样本空间对应于“商品有质量问题”这个应用场景,第二样本空间对应于“操作问题”这个应用场景等等。

[0044] 进一步地,电子设备110可以获取针对多个样本空间的多个引导项(prompt),这多个引导项至少包括针对第一样本空间的第一引导项和针对第二样本空间的第二引导项。这里的多个引导项与多个样本空间例如可以是一一对应的关系。每个引导项均指示针对目标样本集的多个样本空间中的相应样本空间的样本分类策略。例如,第一引导项可以指示针对目标样本集的多个样本空间中的第一样本空间的第一样本分类策略,第二引导项可以指示针对多个样本空间中的第二样本空间的第二样本分类策略等等。这里的样本分类策略至少指示待分类的类别数目。也即,引导项应至少指示对应样本空间待分类的类别数目(也即聚类簇数目)。在一些实施例中,这样的引导项还可以指示模型生成每个类别(也可以称为每个类簇)的名称定义和代表性问题。

[0045] 在一些实施例中,电子设备110中可以预先存储有用于生成这样的引导项的引导项模板。电子设备110可以基于这样的引导项模板来生成针对不同样本空间的多个引导项。可以理解,电子设备110生成的针对第二样本空间的第二引导项与针对第一样本空间的第一引导项可以为基于相同引导项模板所生成的不同引导项。备选地或者附加地,这样的多个引导项还可以是电子设备110通过其他方式获取的。例如,可以是基于用户输入所获取的,或者,从其他设备处通过通信连接所获取的。

[0046] 在框220,电子设备110基于第一引导项,确定针对第一样本空间的第一分类结果。在框230,电子设备110基于第二引导项,确定针对第二样本空间的第二分类结果。

[0047] 电子设备110获取到针对多个样本空间的多个引导项后,可以基于这多个引导项,确定多个样本空间各自的分类结果。例如,电子设备110基于第一引导项,确定针对第一样本空间的第一分类结果。电子设备110基于第二引导项,确定针对第二样本空间的第二分类结果。等等。

[0048] 针对每个样本空间,相应的引导项将被提供给经训练的机器学习模型以便机器学习模型基于引导项对样本空间中的样本进行分类。由于目标样本集是问题样本集,这里的机器学习模型例如可以为语言模型或者大语言模型。

[0049] 参考图3A以及图3B。图3A示出了根据本公开的一些实施例的第一引导项300A的示例的示意图。图3B示出了根据本公开的一些实施例的第一分类结果300B的示例的示意图。第一引导项300A例如可以是电子设备110基于引导项模板所生成的针对第一样本空间的引导项。

[0050] 如图3A所示,第一引导项300A中包括与第一样本空间相应的应用场景“商品有质量问题”。这代表第一样本空间中所包括的问题样本均为“商品有质量问题”这一场景下的问题样本。

[0051] 第一引导项300A还包括超参数301(例如5-8),超参数用于指示对样本空间进行分类的分类数目。关于超参数301的数目的确定,在一些实施例中,电子设备110可以基于相应

的样本空间所包括的潜在分类个数(NC)来确定。这里的潜在分类个数即为样本空间所对应的更新后的样本子集所包括的原始样本子集的数目。电子设备110可以确定超参数所指示的数目为潜在分类个数的1.3到2倍,且为了保证样本的完整,取值时应向下取整。例如,目标样本集在第一次聚类后被划分为5个样本子集A、B、C、D、E,其中A和B之间的相似度高于阈值,D和E之间的相似度高于阈值,则更新后的多个样本子集为样本子集AB(A+B)、C、DE(D+E)。电子设备110可以确定与这三个样本子集各自对应的样本空间1、2、3,若样本空间1对应样本子集AB,则可以确定样本空间1的潜在分类数目为2。因此,样本空间1对应的引导项中的超参数所指示的数目应为2的1.3到2倍,且由于向下取整,超参数所指示的数目应为2至4。

[0052] 第一引导项300A中的“为每个簇生成简明的簇定义和2个语义完整的代表性问法”可以指示模型生成每个类别的名称定义和生成2个代表性问题,这两个代表性问题可以是第一样本空间中的问题样本,也可以是模型自行生成的新的问题样本。

[0053] 第一引导项300A中的“相似的簇应该合并为一个,确保簇的数目尽可能少”可以指示模型生成针对更细粒度的不同应用场景的多个类别。也即,这多个类别对应的大的场景虽然均为“商品有质量问题”,但这多个类别对应的更细粒度的应用场景彼此不同。例如,第一类别对应“游戏性能问题”这一应用场景,第二类别对应“软件闪退问题”这一应用场景。

[0054] 第一引导项300A中的“严格按照“#簇1:<10字内的定义>。Q1:<20字内的问法>Q2:<20字内的问法>”这种格式,确保每个簇在一行内输出。”用于指示模型生成的名称定义和生成2个代表性问题的格式要求。第一引导项300A中的多个问题样本302即为第一样本空间中所包含的多个问题样本。

[0055] 第一引导项300A被提供给语言模型后,语言模型可以基于第一引导项生成针对第一样本空间的第一分类结果300B。第一分类结果300B中包含5个簇,也即包含5个分类。每个分类均具有相应的名称定义以及2个代表性问题。例如,第一分类结果300B中的簇1(也即分类1)的10字内的名称定义为“声音画面卡顿问题”,其对应的2个代表性问题(且这两个代表性问题的字数都为20字以内)分别为“Q1:声音画面都会卡一下,怎么回事?”和“Q2:打游戏弹出的东西为什么会出现声音画面卡顿?”。

[0056] 可以理解,模型获取到针对第二样本空间的第二引导项后,可以生成与第一分类结果300B相似的第二分类结果。

[0057] 以此方式,电子设备110可以基于多个引导项,通过模型确定多个样本空间对应的多个样本分类结果。每个样本分类结果中均包含相应样本空间的至少一个分类。这样的针对多个样本空间的多个分类结果可以被视为最终生成的意图分类体系的子空间知识体系。

[0058] 在框240,电子设备110至少基于第一分类结果和第二分类结果,确定针对目标样本集的样本分类结果。

[0059] 在一些实施例中,在确定针对目标样本集的样本分类结果之前,电子设备110需要确定针对目标样本集的目标分类数目。目标分类数目即为用户最终期望的针对目标样本集的分类数目M。电子设备110可以基于目标分类数目,采用任意适当的聚类方法(例如采用K-means算法)对与多个样本空间相应的多个分类结果进行聚类以确定针对目标样本集的样本分类结果。这里的聚类即为与目标样本集相关联的第二次聚类。第二次聚类的聚类簇数目为用户最终期望的针对目标样本集的分类数目M。可以理解,在电子设备110可以直接获

取多个样本空间的情况下,电子设备110也可以仅执行第二次聚类而无需执行第一次聚类。

[0060] 在一些实施例中,在执行第二次聚类之前,电子设备110还可以将多个样本空间分别对应的多个分类结果聚合起来以得到一个聚合结果。例如,可以将样本空间对应的分类结果中的所有<分类簇名称,问法Q1,问法Q2>这样的数据结果,按照“分类簇名称”聚合起来,形成针对样本空间的<分类名,问题列表>这样的聚合结果,其中“问题列表”囊括不超过10个问题。

[0061] 对于针对不同样本空间的多个聚合结果,可以将这多个聚合结果基于“分类名”和“问题列表拼接”拼接为一个文本。电子设备110进而可以使用任意适当的编码方法(例如基于大模型的编码方法)将这个文本编码为向量表征。因为拼接后的文本语义较为复杂,普通的小模型(例如SimCSE等)难以有效编码。电子设备110进而可以使用聚类算法在这个表征上进行第二次聚类以使得多个独立空间的分类结果被归并到了M个分类下面。也即电子设备110通过第二次聚类所确定的针对目标样本集的样本分类结果中的每个类簇包含多个<分类,问题列表>的数据,每个类簇中的数据都是同一应用场景下语义比较相似的问题数据。

[0062] 可以发现,针对目标样本集进行第二次聚类所获取到的样本分类结果中,每个分类均包含多个样本空间中至少一个样本空间对应的分类结果的至少部分。例如,多个样本空间包括样本空间1、2、3,其中样本空间1对应的分类结果包含5个聚类簇,样本空间2对应的分类结果包含3个聚类簇,样本空间3对应的分类结果包含4个聚类簇,则三个样本空间对应的多个分类结果共包含12个聚类簇。若用户期望的针对目标样本集的分类数目为10,则电子设备110在第二次聚类时需要将这12个聚类簇进行聚类以得到10个最终的聚类簇。可以理解,10个最终的聚类簇中的第一聚类簇应包括这12个聚类簇中的至少一个聚类簇,也即针对目标样本集的样本分类结果的第一分类中应至少包括针对目标样本空间的分类结果的至少部分(即针对目标样本空间的分类结果中的至少一个聚类簇),这里的目标样本空间可以是多个样本空间中的任意样本空间。

[0063] 在一些实施例中,电子设备110还可以获取目标引导项。目标引导项可以被提供给模型以指示模型生成与目标引导项相对应的目标分类的描述信息。图3C示出了根据本公开的一些实施例的目标引导项300C的示例的示意图。如图3C所示,针对样本分类结果中的目标分类(目标分类可以是样本分类结果中的多个聚类簇中的任意一个),电子设备110可以获取到与目标分类相对应的目标引导项300C。目标引导项300C中的“请你根据分类描述和相关问法示例信息,将下面的分类总结为一个简明具体的分类,并生成20个代表性的示例问法”可以指示模型对目标分类所包含的针对样本空间的至少一个分类互补性总结归纳并生成名称定义,以及生成20个示例问题。同样的,这20个示例问题可以是目标分类所包括的问题样本,也可以是模型自行生成的新的问题样本。目标引导项300C中的分类列表303示出了目标分类所包括的全部问题样本中的部分问题样本。可以理解,在目标引导项300C针对另一分类时,分类列表303中所包括的问题样本也随之发生改变。

[0064] 模型可以基于目标引导项300C生成针对目标分类的描述信息。描述信息包括目标分类的标识(例如名称定义)以及代表样本(也即代表性问题,还可以被称之为代表问法)。图3D示出了根据本公开的一些实施例的针对目标分类的分类信息300D的示例的示意图。如图3D所示,分类信息300D包括目标分类的名称定义“电脑性能问题”,以及多个代表问法。

[0065] 通过类似的方式,电子设备110可以确定针对目标样本集的样本分类结果所包括的多个分类中每个分类对应的标识以及代表样本。

[0066] 图4示出了根据本公开的一些实施例的分类的流程400的示例的示意图。如图4所示,电子设备110获取到问题样本集合410后,可以利用语言模型415确定问题样本集合410所包括的多个问题样本各自对应的语义信息。电子设备110进而可以基于语义信息对问题样本集合410中的多个问题样本进行第一次聚类以得到多个子样本集420。多个子样本集420包括子样本集1、子样本集2、子样本集3、子样本集4以及子样本集5。电子设备110确定各个子样本集与其他子样本集之间的相似度,并将相似度超过阈值的至少两个子样本集确定为一个子样本集。由此,多个子样本集420可以被视为包括子样本集12(子样本集1和子样本集2)、子样本集3以及子样本集45(包括子样本集4和子样本集5)。电子设备110进而通过对子样本集12、子样本集3以及子样本集45分别进行采样以确定与三者分别对应的多个独立空间430(也即多个样本空间)。多个独立空间430中包括独立空间A、独立空间B和独立空间C。电子设备110可以通过语言模型415来分别确定多个独立空间430分别对应的分类结果。例如,独立空间A对应的分类结果包括A_1、A_2、……、A_{nA}共nA个分类,独立空间B对应的分类结果包括B_1、B_2、……、B_{nB}共nB个分类,独立空间C对应的分类结果包括C_1、C_2、……、C_{nC}共nC个分类,其中nA、nB和nC是三个可以相同也可以不同的数值。电子设备110可以将多个独立空间430对应的多个分类结果进行拼接以得到拼接结果440。电子设备110进而对拼接结果440进行第二次聚类以获取针对问题样本集合410的样本分类结果450。电子设备110还可以利用语言模型415确定样本分类结果450中每个分类所对应的描述信息(包括标识以及代表样本)。

[0067] 以此方式,电子设备110可以获取针对问题样本集合410的、包括相应的分类信息的样本分类结果450。这样的样本分类结果450也可以被视为与问题样本集合410相关联的意图分类体系结果。此外,采用这样的方式确定问题样本集合410的样本分类结果450一方面可以避免由于语言模型生成长度有限、而样本数据量过大所导致的难以完全加载问题(由于模型生成长度普遍为2000-4000,而样本数据往往为百万级,因此难以直接利用语言模型的推理、理解能力进行分类体系构建。)。另一方面,如果直接进行全样本采样,模型每次都看到的是来自于全局的样本,用户无法控制具体的子分类数粒度。此外,每次都让大模型见到很多来自相同潜在分类的数据,会导致生成结果重复度变高,影响最终的分类质量。因此,本公开所述的方式可以规避上述问题的同时,可以利用语言模型进行分类体系构建。

[0068] 综上所述,本公开的实施例可以通过指示用户分类策略的引导项对样本进行分类,并且可以自动确定分类结果,无需人力成本,且对于意图分类体系的构建效率较高。

[0069] 本公开的实施例还提供了用于实现上述方法或过程的相应装置。图5示出了根据本公开的某些实施例的用于分类的装置500的示意性结构框图。装置500可以被实现为或者被包括在电子设备110中。装置500中的各个模块/组件可以由硬件、软件、固件或者它们的任意组合来实现。

[0070] 如图5所示,装置500包括引导项获取模块510,被配置为获取第一引导项和第二引导项,第一引导项指示针对目标样本集的多个样本空间中的第一样本空间的第一样本分类策略,第二引导项指示针对多个样本空间中的第二样本空间的第二样本分类策略。装置500还包括第一结果确定模块520,被配置为基于第一引导项,确定针对第一样本空间的第一分

类结果。装置500还包括第二结果确定模块530,被配置为基于第二引导项,确定针对第二样本空间的第二分类结果。装置500还包括分类结果确定模块540,被配置为至少基于第一分类结果和第二分类结果,确定针对目标样本集的样本分类结果。

[0071] 在一些实施例中,目标样本集包括多个子样本集,多个样本空间是基于对多个子样本集进行采样所得到的。

[0072] 在一些实施例中,装置500还包括:样本空间确定模块,被配置为确定多个子样本集中各个子样本集之间的相似度;以及通过将相似度高于阈值的至少两个子样本集进行合并,确定多个样本空间。

[0073] 在一些实施例中,将相似度高于阈值的至少两个子样本集进行合并使得多个子样本集的数目从第一数目改变为第二数目,第一数目大于第二数目。

[0074] 在一些实施例中,第一样本分类策略和/或第二样本分类策略至少指示待分类的类别数目。

[0075] 在一些实施例中,分类结果确定模块540包括:分类数目确定模块,被配置为确定针对目标样本集的目标分类数目;以及聚类模块,被配置为基于目标分类数目,至少对第一分类结果和第二分类结果进行聚类以确定针对目标样本集的样本分类结果。

[0076] 在一些实施例中,目标样本集的样本分类结果包括目标分类,目标分类包括针对目标样本空间的分类结果的至少部分,装置500还包括:目标引导项获取模块,被配置为获取针对目标分类的目标引导项;以及描述信息确定模块,被配置为向模型提供目标引导项,以指示模型基于目标引导项确定目标分类的描述信息,描述信息包括目标分类的标识和/或目标分类的代表样本。

[0077] 装置500中所包括的单元和/或模块可以利用各种方式来实现,包括软件、硬件、固件或其任意组合。在一些实施例中,一个或多个单元和/或模块可以使用软件和/或固件来实现,例如存储在存储介质上的机器可执行指令。除了机器可执行指令之外或者作为替代,装置500中的部分或者全部单元和/或模块可以至少部分地由一个或多个硬件逻辑组件来实现。作为示例而非限制,可以使用的示范类型的硬件逻辑组件包括现场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准品(ASSP)、片上系统(SOC)、复杂可编程逻辑器件(CPLD),等等。

[0078] 应当理解,以上方法中的一个或多个步骤可以由适当的电子设备或电子设备的组合来执行。这样的电子设备或电子设备的组合例如可以包括图1中的电子设备110。

[0079] 图6示出了其中可以实施本公开的一个或多个实施例的电子设备600的框图。应当理解,图6所示出的电子设备600仅仅是示例性的,而不应当构成对本文所描述的实施例的功能和范围的任何限制。图6所示出的电子设备600可以用于实现图1的电子设备110。

[0080] 如图6所示,电子设备600是通用电子设备的形式。电子设备600的组件可以包括但不限于一个或多个处理器或处理单元610、存储器620、存储设备630、一个或多个通信单元640、一个或多个输入设备650以及一个或多个输出设备660。处理单元610可以是实际或虚拟处理器并且能够根据存储器620中存储的程序来执行各种处理。在多处理器系统中,多个处理单元并行执行计算机可执行指令,以提高电子设备600的并行处理能力。

[0081] 电子设备600通常包括多个计算机存储介质。这样的介质可以是电子设备600可访问的任何可以获得的介质,包括但不限于易失性和非易失性介质、可拆卸和不可拆卸介质。

存储器620可以是易失性存储器(例如寄存器、高速缓存、随机访问存储器(RAM))、非易失性存储器(例如,只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、闪存)或它们的某种组合。存储设备630可以是可拆卸或不可拆卸的介质,并且可以包括机器可读介质,诸如闪存驱动、磁盘或者任何其他介质,其可以能够用于存储信息和/或数据并且可以在电子设备600内被访问。

[0082] 电子设备600可以进一步包括另外的可拆卸/不可拆卸、易失性/非易失性存储介质。尽管未在图6中示出,可以提供用于从可拆卸、非易失性磁盘(例如"软盘")进行读取或写入的磁盘驱动和用于从可拆卸、非易失性光盘进行读取或写入的光盘驱动。在这些情况中,每个驱动可以由一个或多个数据介质接口被连接至总线(未示出)。存储器620可以包括计算机程序产品625,其具有一个或多个程序模块,这些程序模块被配置为执行本公开的各种实施例的各种方法或动作。

[0083] 通信单元640实现通过通信介质与其他电子设备进行通信。附加地,电子设备600的组件的功能可以以单个计算集群或多个计算机器来实现,这些计算机器能够通过通信连接进行通信。因此,电子设备600可以使用与一个或多个其他服务器、网络个人计算机(PC)或者另一个网络节点的逻辑连接来在联网环境中进行操作。

[0084] 输入设备650可以是一个或多个输入设备,例如鼠标、键盘、追踪球等。输出设备660可以是一个或多个输出设备,例如显示器、扬声器、打印机等。电子设备600还可以根据需要通过通信单元640与一个或多个外部设备(未示出)进行通信,外部设备诸如存储设备、显示设备等,与一个或多个使得用户与电子设备600交互的设备进行通信,或者与使得电子设备600与一个或多个其他电子设备通信的任何设备(例如,网卡、调制解调器等)进行通信。这样的通信可以经由输入/输出(I/O)接口(未示出)来执行。

[0085] 根据本公开的示例性实现方式,提供了一种计算机可读存储介质,其上存储有计算机可执行指令,其中计算机可执行指令被处理器执行以实现上文描述的方法。根据本公开的示例性实现方式,还提供了一种计算机程序产品,计算机程序产品被有形地存储在非瞬态计算机可读介质上并且包括计算机可执行指令,而计算机可执行指令被处理器执行以实现上文描述的方法。

[0086] 这里参照根据本公开实现的方法、装置、设备和计算机程序产品的流程图和/或框图描述了本公开的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0087] 这些计算机可读程序指令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单元,从而生产出一种机器,使得这些指令在通过计算机或其他可编程数据处理装置的处理单元执行时,产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作,从而,存储有指令的计算机可读介质则包括一个制造品,其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0088] 可以把计算机可读程序指令加载到计算机、其他可编程数据处理装置、或其他设备上,使得在计算机、其他可编程数据处理装置或其他设备上执行一系列操作步骤,以产生计算机实现的过程,从而使得在计算机、其他可编程数据处理装置、或其他设备上执行的指

令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0089] 附图中的流程图和框图显示了根据本公开的多个实现的系统、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、程序段或指令的一部分,模块、程序段或指令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。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,两个连续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动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] 以上已经描述了本公开的各实现,上述说明是示例性的,并非穷尽性的,并且也不限于所公开的各实现。在不偏离所说明的各实现的范围和精神的情况下,对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择,旨在最好地解释各实现的原理、实际应用或对市场中的技术的改进,或者使本技术领域的其他普通技术人员能理解本文公开的各个实现方式。

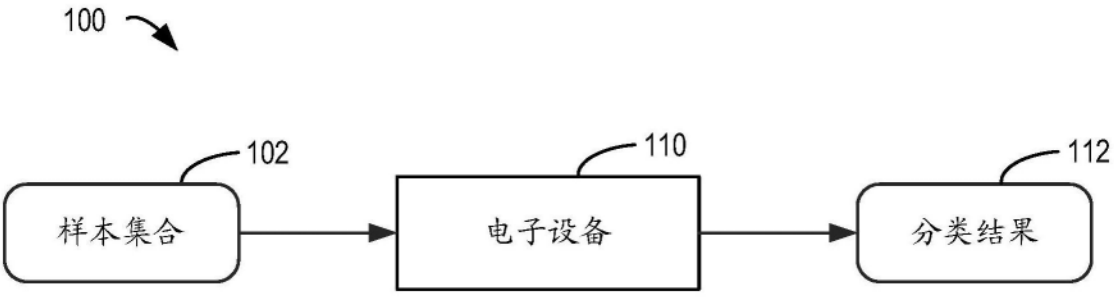

100
102
样本集合
110
电子设备
112
分类结果

图1

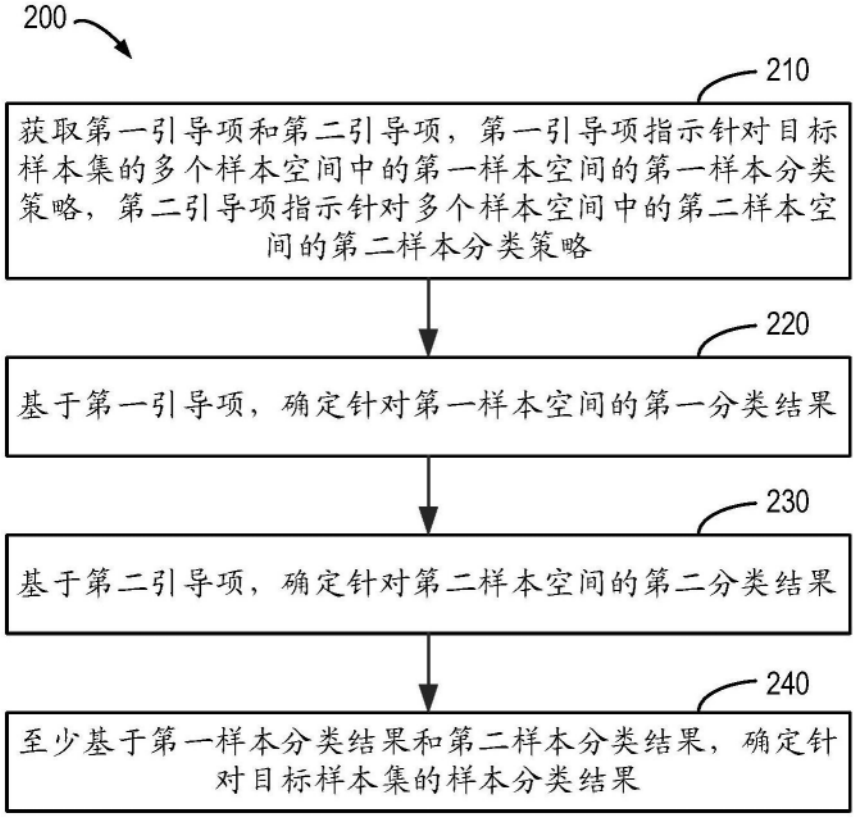

200
210
获取第一引导项和第二引导项，第一引导项指示针对目标样本集的多个样本空间中的第一样本空间的第一样本分类策略，第二引导项指示针对多个样本空间中的第二样本空间的第二样本分类策略
220
基于第一引导项，确定针对第一样本空间的第一分类结果
230
基于第二引导项，确定针对第二样本空间的第二分类结果
240
至少基于第一样本分类结果和第二样本分类结果，确定针对目标样本集的样本分类结果

图2

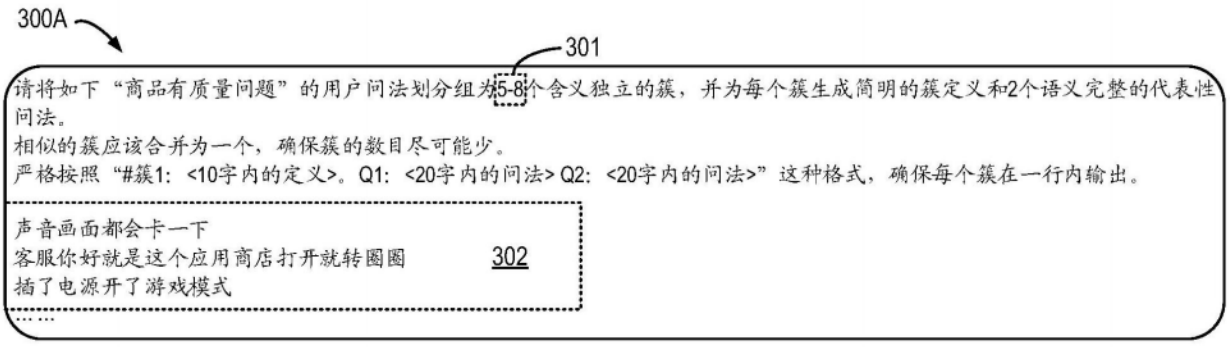

300A
301
请将如下“商品有质量问题”的用户问法划分组为5-8个含义独立的簇，并为每个簇生成简明的簇定义和2个语义完整的代表性问法。
相似的簇应该合并为一个，确保簇的数目尽可能少。
严格按照“#簇1：<10字内的定义>。Q1：<20字内的问法> Q2：<20字内的问法>”这种格式，确保每个簇在一行内输出。
声音画面都会卡一下
客服你好就是这个应用商店打开就转圈圈
插了电源开了游戏模式
302
……

图3A

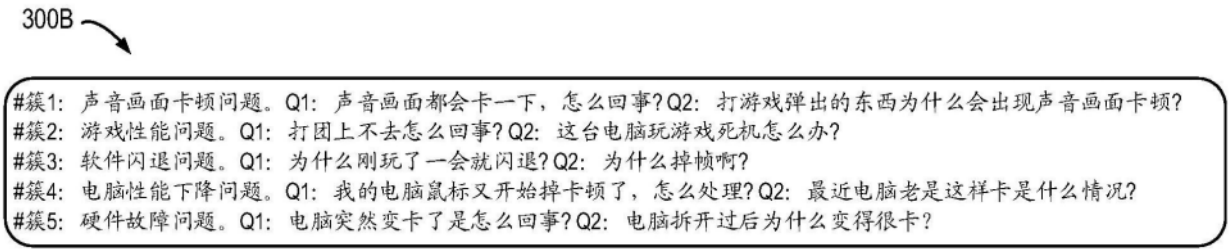

300B
#簇1：声音画面卡顿问题。Q1：声音画面都会卡一下，怎么回事？Q2：打游戏弹出的东西为什么会出现声音画面卡顿？
#簇2：游戏性能问题。Q1：打团上不去怎么回事？Q2：这台电脑玩游戏死机怎么办？
#簇3：软件闪退问题。Q1：为什么刚玩了一会就闪退？Q2：为什么掉帧啊？
#簇4：电脑性能下降问题。Q1：我的电脑鼠标又开始掉卡顿了，怎么处理？Q2：最近电脑老是这样卡是什么情况？
#簇5：硬件故障问题。Q1：电脑突然变卡了是怎么回事？Q2：电脑拆开过后为什么变得很卡？

图3B

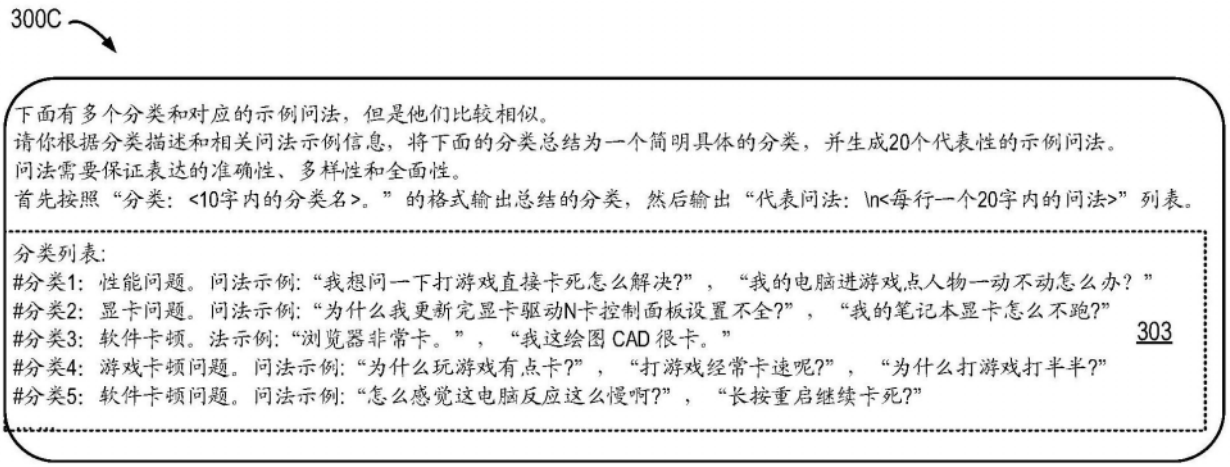

300C
下面有多个分类和对应的示例问法，但是他们比较相似。
请你根据分类描述和相关问法示例信息，将下面的分类总结为一个简明具体的分类，并生成20个代表性的示例问法。
问法需要保证表达的准确性、多样性和全面性。
首先按照“分类：<10字内的分类名>。”的格式输出总结的分类，然后输出“代表问法：\n<每行一个20字内的问法>”列表。
分类列表：
#分类1：性能问题。问法示例：“我想问一下打游戏直接卡死怎么解决？”，“我的电脑进游戏点人物一动不动怎么办？”
#分类2：显卡问题。问法示例：“为什么我更新完显卡驱动N卡控制面板设置不全？”，“我的笔记本显卡怎么不跑？”
#分类3：软件卡顿。法示例：“浏览器非常卡。”，“我这绘图 CAD 很卡。”
#分类4：游戏卡顿问题。问法示例：“为什么玩游戏有点卡？”，“打游戏经常卡速呢？”，“为什么打游戏打半半？”
#分类5：软件卡顿问题。问法示例：“怎么感觉这电脑反应这么慢啊？”，“长按重启继续卡死？”
303

图3C

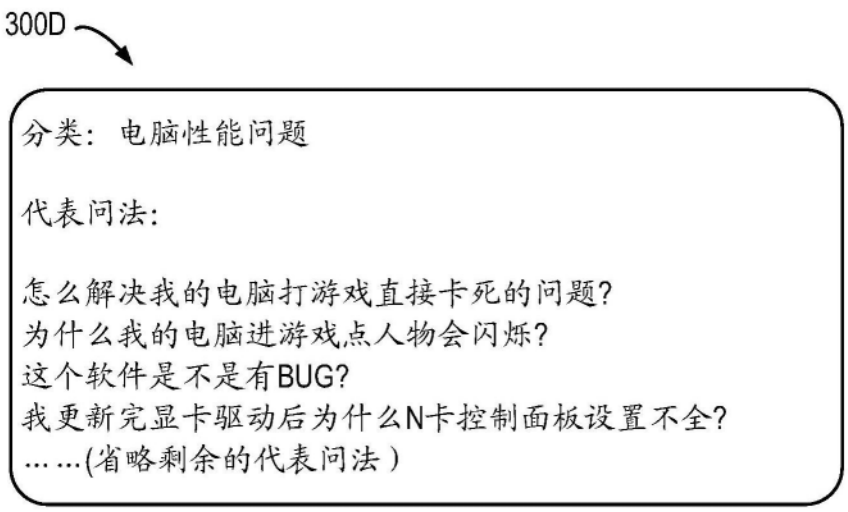

300D
分类：电脑性能问题
代表问法：
怎么解决我的电脑打游戏直接卡死的问题?
为什么我的电脑进游戏点人物会闪烁?
这个软件是不是有BUG?
我更新完显卡驱动后为什么N卡控制面板设置不全?
……(省略剩余的代表问法)

图3D

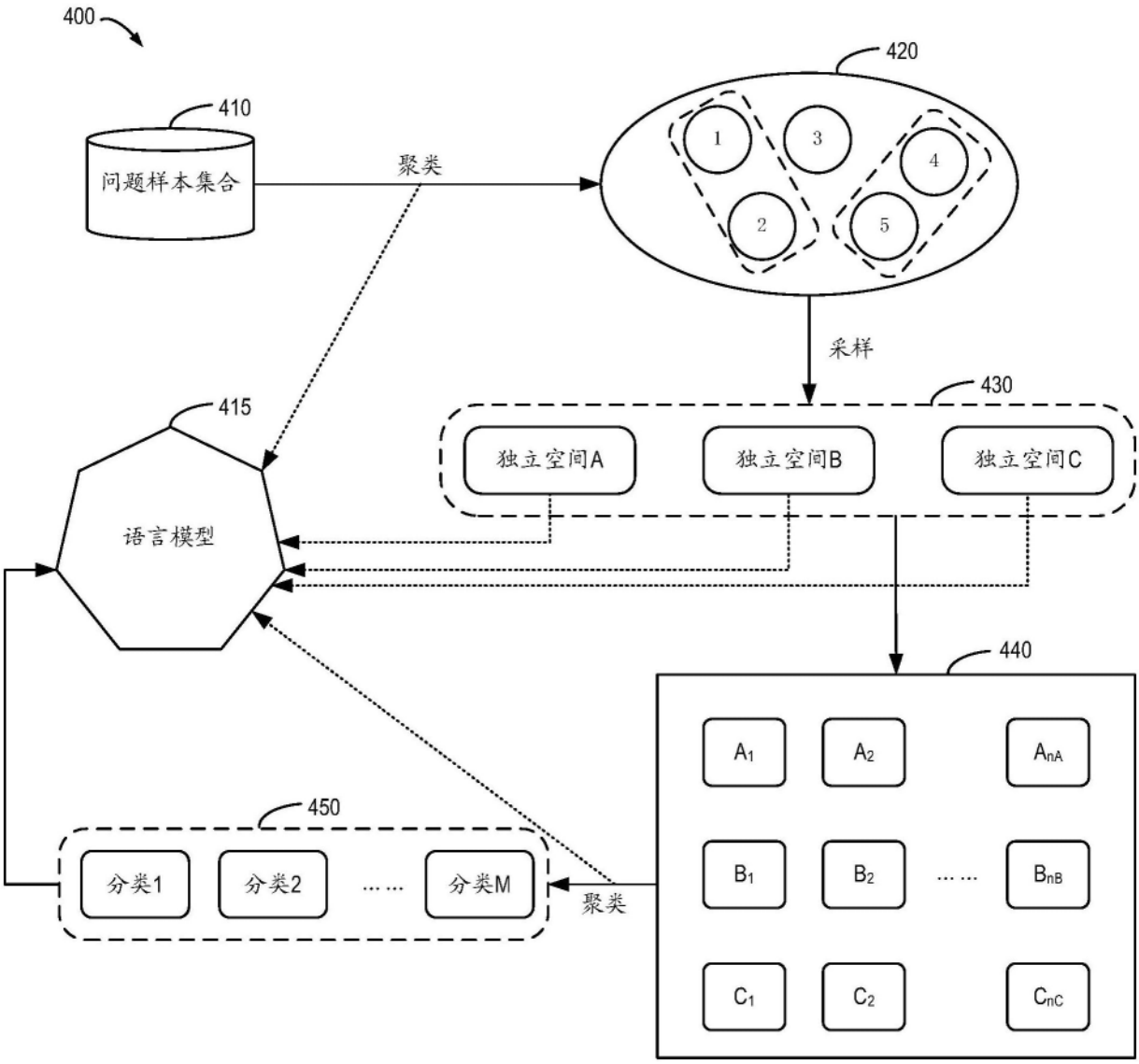

400
410
问题样本集合
聚类
420
1
3
4
2
5
采样
430
独立空间A
独立空间B
独立空间C
415
语言模型
440
A1
A2
AnA
B1
B2
……
BnB
C1
C2
CnC
450
分类1
分类2
……
分类M
聚类

图4

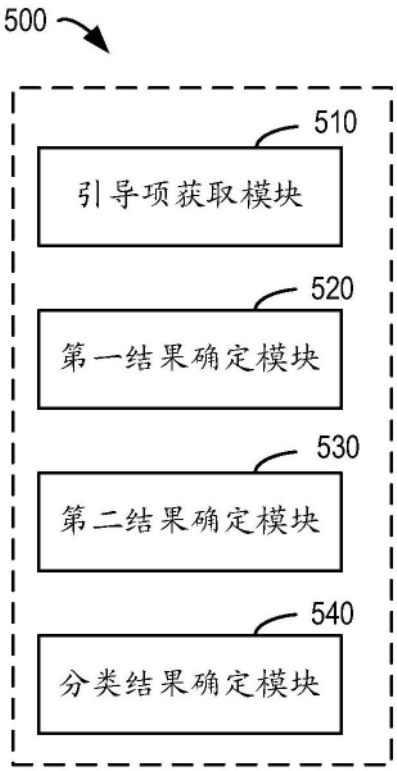

500
510
引导项获取模块
520
第一结果确定模块
530
第二结果确定模块
540
分类结果确定模块

图5

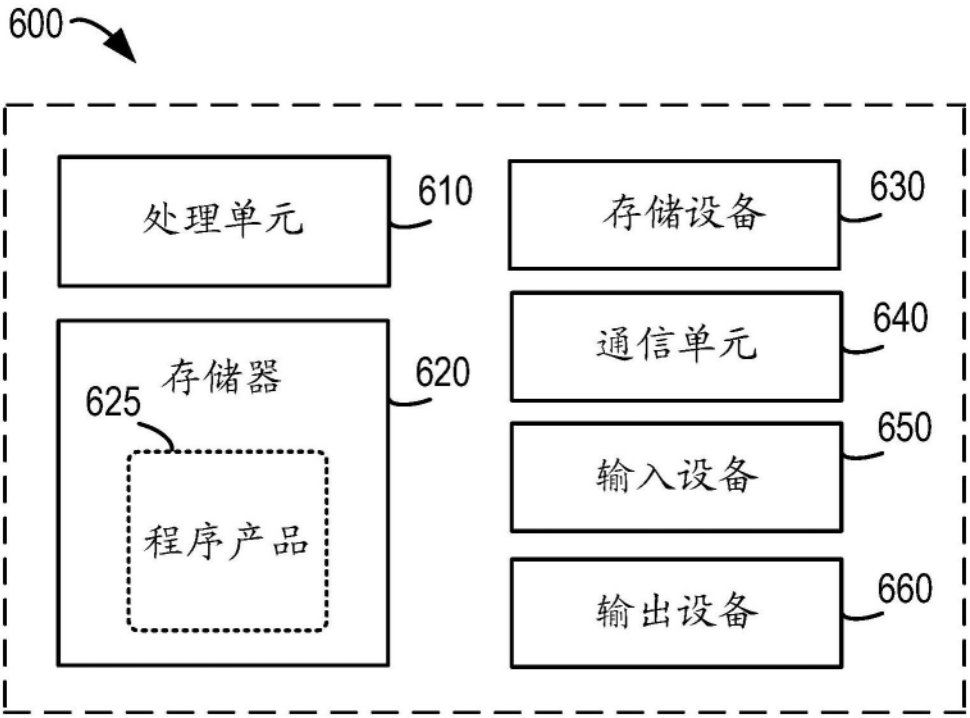

600
处理单元
610
存储设备
630
存储器
620
625
程序产品
通信单元
640
输入设备
650
输出设备
660

图6