(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110472027 B
(45)授权公告日 2024.05.14

(21)申请号 201910653241.9

(22)申请日 2019.07.18

(65)同一申请的已公布的文献号
申请公布号 CN 110472027 A

(43)申请公布日 2019.11.19

(73)专利权人 平安科技(深圳)有限公司
地址 518000 广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼

(72)发明人 石志娟 徐小方

(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 郝传鑫 熊永强

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06F 16/338(2019.01)
G06F 40/289(2020.01)
G06F 40/30(2020.01)

(56)对比文件
CN 107784123 A,2018.03.09
KR 20100036486 A,2010.04.08
CN 106682192 A,2017.05.17
CN 109815492 A,2019.05.28
CN 106599278 A,2017.04.26
CN 109815314 A,2019.05.28
US 2012084291 A1,2012.04.05

审查员 徐军

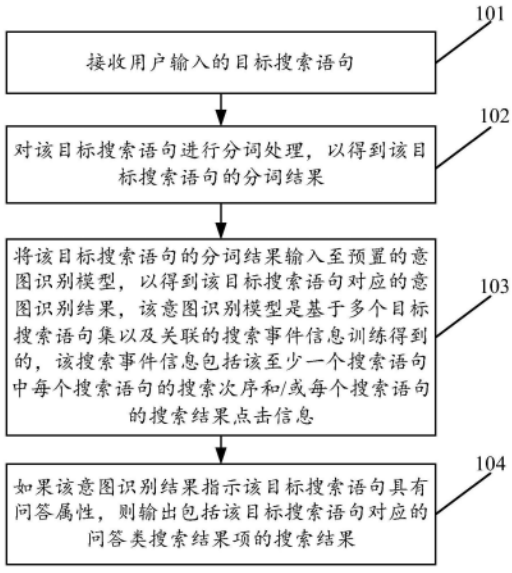
权利要求书3页 说明书16页 附图3页

(54)发明名称

意图识别方法、设备及计算机可读存储介质

(57)摘要

本申请公开了一种意图识别方法、设备及计算机可读存储介质，应用于人工智能技术领域。其中，该方法包括：接收用户输入的目标搜索语句；对所述目标搜索语句进行分词处理，以得到所述目标搜索语句的分词结果；将所述目标搜索语句的分词结果输入至预置的意图识别模型，以得到所述目标搜索语句对应的意图识别结果，所述意图识别结果用于指示所述目标搜索语句是否具有问答属性；如果所述意图识别结果指示所述目标搜索语句具有问答属性，则输出包括所述目标搜索语句对应的问答类搜索结果项的搜索结果。采用本申请，有助于提升意图识别的准确性。



CN 110472027 B

1.一种意图识别方法,其特征在于,包括:

接收用户输入的目标搜索语句;

对所述目标搜索语句进行分词处理,以得到所述目标搜索语句的分词结果,所述目标搜索语句的分词结果包括组成所述目标搜索语句的多个分词;

将所述目标搜索语句的分词结果输入至预置的意图识别模型,以得到所述目标搜索语句对应的意图识别结果,所述意图识别模型是基于多个目标搜索语句集以及所述多个目标搜索语句集中每个目标搜索语句集关联的搜索事件信息、搜索语句的加权系数、搜索语句中预设关键词的加权系数训练得到的,通过所述搜索语句的加权系数、搜索语句中预设关键词的加权系数对所述搜索语句的搜索事件信息进行加权;针对每个目标搜索语句集中的多个搜索语句,该多个搜索语句中任两个搜索语句之间的搜索时间间隔不超过预设时间阈值,且该多个搜索语句中任两个搜索语句之间的关键词的重叠率高于预设重叠率阈值;所述每个目标搜索语句集包括至少一个搜索语句,所述搜索事件信息包括所述至少一个搜索语句中每个搜索语句的搜索次序和所述每个搜索语句的搜索结果点击信息,所述意图识别结果用于指示所述目标搜索语句是否具有问答属性;所述至少一个搜索语句中搜索次序大的搜索语句的加权系数高于搜索次序小的搜索语句的加权系数,以及,所述至少一个搜索语句中搜索结果点击信息包括的搜索结果点击项中存在特定问答网站的展示结果或搜索结果中存在特定问答网站的展示结果的搜索语句的加权系数高于不存在该特定问答网站的展示结果的搜索语句的加权系数;所述搜索结果点击信息包括点击的各搜索结果项的浏览时长,且浏览时长小于预设时长阈值的搜索结果项被过滤掉;

如果所述意图识别结果指示所述目标搜索语句具有问答属性,则确定所述目标搜索语句的问答属性所属的问答类别,并根据所述问答类别输出包括所述目标搜索语句对应的问答类搜索结果项的搜索结果;问答类别分为包括疑问词的显式问答搜索语句类别,以及,不包括疑问词的隐式问答搜索语句类别,且不同问答类别对应的展示内容或页面不同;

其中,多个目标搜索语句集是从搜索语句数据库中选取出的,选取方式为:

确定待训练的意图识别模型的应用领域信息;

根据所述应用领域信息从所述搜索语句数据库包括的多个子数据库中确定出目标子数据库,所述子数据库与应用领域一一对应,每个子数据库包括对应的应用领域下的多个搜索语句集以及每个搜索语句集关联的搜索事件信息,所述目标子数据库对应的应用领域与所述应用领域信息指示的应用领域相同;

从所述目标子数据库选取所述多个目标搜索语句集;目标搜索语句集的搜索语句中预设关键词设置有加权系数,所述预设关键词为所述应用领域信息对应应用领域中特有词语或该应用领域中出现频率高的词语。

2.根据权利要求1所述的方法,其特征在于,在所述将所述目标搜索语句的分词结果输入至预置的意图识别模型,以得到所述目标搜索语句对应的意图识别结果之前,所述方法还包括:

从搜索语句数据库中选取多个目标搜索语句集;其中,所述搜索语句数据库中记录了多个搜索语句集以及每个搜索语句集关联的搜索事件信息,所述每个搜索语句集包括至少一个搜索语句,所述搜索事件信息包括所述至少一个搜索语句中每个搜索语句的搜索次序和/或所述每个搜索语句的搜索结果点击信息;

分别对所述多个目标搜索语句集中每个目标搜索语句集包括的搜索语句进行分词处理,以得到所述每个目标搜索语句集的分词结果,所述每个目标搜索语句集的分词结果包括组成该目标搜索语句集的搜索语句的多个分词;

根据所述每个目标搜索语句集关联的搜索事件信息,确定所述每个目标搜索语句集包括的搜索语句是否具有问答属性;

将所述多个目标搜索语句集中具有问答属性的目标搜索语句集的分词结果作为正样本,以及将所述多个目标搜索语句集中不具有问答属性的目标搜索语句集的分词结果作为负样本,并利用所述多个目标搜索语句集对应的正样本和负样本训练得到意图识别模型;其中,所述意图识别模型用于识别输入的搜索语句是否具有问答属性。

3.根据权利要求2所述的方法,其特征在于,所述根据所述目标搜索语句集关联的搜索事件信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性,包括:

根据所述目标搜索语句集关联的搜索事件信息所包括的每个搜索语句的搜索次序,确定所述目标搜索语句集包括的所述至少一个搜索语句中最大搜索次序对应的搜索语句;

根据所述最大搜索次序对应的搜索语句的搜索结果点击信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性。

4.根据权利要求3所述的方法,其特征在于,所述搜索结果点击信息包括搜索结果项的点击总数量和问答类的搜索结果项的点击数量;所述根据所述最大搜索次序对应的搜索语句的搜索结果点击信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性,包括:

计算所述最大搜索次序对应的搜索语句的搜索结果点击信息包括的问答类的搜索结果项的点击数量与搜索结果项的点击总数量之间的第一比值;

如果所述搜索结果项的点击总数量大于预设的第一数目阈值,且所述第一比值大于预设的第一比例阈值,确定所述目标搜索语句集包括的搜索语句具有问答属性。

5.根据权利要求2所述的方法,其特征在于,所述根据所述目标搜索语句集关联的搜索事件信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性,包括:

确定所述目标搜索语句集包括的所述至少一个搜索语句中每个搜索语句对应的加权系数,所述至少一个搜索语句中搜索次序大的搜索语句的加权系数高于搜索次序小的搜索语句的加权系数;

根据每个搜索语句对应的加权系数和所述目标搜索语句集关联的搜索事件信息中所述每个搜索语句的搜索结果点击信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性。

6.根据权利要求2-5任一项所述的方法,其特征在于,在所述利用所述多个目标搜索语句集对应的正样本和负样本训练得到意图识别模型之前,所述方法还包括:

计算所述正样本对应的搜索语句的数量与所述负样本对应的搜索语句的数量之间的差值的绝对值;

判断所述绝对值是否超过预设的第三数目阈值;

如果所述绝对值超过所述第三数目阈值,按照预设的样本平衡规则对所述正样本和/或所述负样本进行处理,以得到处理后的正样本和负样本;

所述利用所述多个目标搜索语句集对应的正样本和负样本训练得到意图识别模型,包

括：

利用处理后的正样本和负样本训练得到所述意图识别模型。

7.一种意图识别装置，其特征在于，包括用于执行如权利要求1-6任一项所述的方法的单元。

8.一种意图识别设备，其特征在于，包括处理器和存储器，所述处理器和存储器相互连接，其中，所述存储器用于存储计算机程序，所述计算机程序包括程序指令，所述处理器被配置用于调用所述程序指令，执行如权利要求1-6任一项所述的方法。

9.一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程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行时使所述处理器执行如权利要求1-6任一项所述的方法。

意图识别方法、设备及计算机可读存储介质

技术领域

[0001] 本申请涉及人工智能技术领域，尤其涉及一种意图识别方法、设备及计算机可读存储介质。

背景技术

[0002] 目前，搜索引擎可以基于用户输入的搜索语句，识别出该搜索语句的意图，以基于识别出的意图为用户提供搜索结果。一般的搜索语句包括具有问答意图的搜索语句和不具有问答意图的搜索语句，如果识别出某一搜索语句具有问答意图，则该搜索语句的搜索结果中可以提供多条问答数据以供用户查看，以便于尽快解决用户的问题，增强用户体验。目前判断搜索语句是否具有问答意图一般是通过判断该搜索语句是否包括疑问词，如果包括疑问词则确定该搜索语句具有问答意图，否则确定该搜索语句不具有问答意图。然而，实际上某些具有问答意图的搜索语句也可能不包括疑问词，这就导致该基于疑问词的问答意图识别方式不可靠，意图识别的准确性较差。

发明内容

[0003] 本申请实施例提供一种意图识别方法、设备及计算机可读存储介质，能够根据搜索语句集关联的搜索事件信息训练得到意图识别模型以进行问答意图识别，有助于提升意图识别的准确性。

[0004] 第一方面，本申请实施例提供了一种意图识别方法，包括：

[0005] 接收用户输入的目标搜索语句；

[0006] 对所述目标搜索语句进行分词处理，以得到所述目标搜索语句的分词结果，所述目标搜索语句的分词结果包括组成所述目标搜索语句的多个分词；

[0007] 将所述目标搜索语句的分词结果输入至预置的意图识别模型，以得到所述目标搜索语句对应的意图识别结果，所述意图识别模型是基于多个目标搜索语句集以及所述多个目标搜索语句集中每个目标搜索语句集关联的搜索事件信息训练得到的，所述每个目标搜索语句集包括至少一个搜索语句，所述搜索事件信息包括所述至少一个搜索语句中每个搜索语句的搜索次序和/或所述每个搜索语句的搜索结果点击信息，所述意图识别结果用于指示所述目标搜索语句是否具有问答属性；

[0008] 如果所述意图识别结果指示所述目标搜索语句具有问答属性，则输出包括所述目标搜索语句对应的问答类搜索结果项的搜索结果。

[0009] 可选的，在所述将所述目标搜索语句的分词结果输入至预置的意图识别模型，以得到所述目标搜索语句对应的意图识别结果之前，所述方法还包括：

[0010] 从搜索语句数据库中选取多个目标搜索语句集；其中，所述搜索语句数据库中记录了多个搜索语句集以及每个搜索语句集关联的搜索事件信息，所述每个搜索语句集包括至少一个搜索语句，所述搜索事件信息包括所述至少一个搜索语句中每个搜索语句的搜索次序和/或所述每个搜索语句的搜索结果点击信息；

[0011] 分别对所述多个目标搜索语句集中每个目标搜索语句集包括的搜索语句进行分词处理,以得到所述每个目标搜索语句集的分词结果,所述每个目标搜索语句集的分词结果包括组成该目标搜索语句集的搜索语句的多个分词;

[0012] 根据所述每个目标搜索语句集关联的搜索事件信息,确定所述每个目标搜索语句集包括的搜索语句是否具有问答属性;

[0013] 将所述多个目标搜索语句集中具有问答属性的目标搜索语句集的分词结果作为正样本,以及将所述多个目标搜索语句集中不具有问答属性的目标搜索语句集的分词结果作为负样本,并利用所述多个目标搜索语句集对应的正样本和负样本训练得到意图识别模型;其中,所述意图识别模型用于识别输入的搜索语句是否具有问答属性。

[0014] 可选的,所述根据所述目标搜索语句集关联的搜索事件信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性,包括:

[0015] 根据所述目标搜索语句集关联的搜索事件信息所包括的每个搜索语句的搜索次序,确定所述目标搜索语句集包括的所述至少一个搜索语句中最大搜索次序对应的搜索语句;

[0016] 根据所述最大搜索次序对应的搜索语句的搜索结果点击信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性。

[0017] 可选的,所述搜索结果点击信息包括搜索结果项的点击总数量和问答类的搜索结果项的点击数量;所述根据所述最大搜索次序对应的搜索语句的搜索结果点击信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性,包括:

[0018] 计算所述最大搜索次序对应的搜索语句的搜索结果点击信息包括的问答类的搜索结果项的点击数量与搜索结果项的点击总数量之间的第一比值;

[0019] 如果所述搜索结果项的点击总数量大于预设的第一数目阈值,且所述第一比值大于预设的第一比例阈值,确定所述目标搜索语句集包括的搜索语句具有问答属性。

[0020] 可选的,所述根据所述目标搜索语句集关联的搜索事件信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性,包括:

[0021] 确定所述目标搜索语句集包括的所述至少一个搜索语句中每个搜索语句对应的加权系数,所述至少一个搜索语句中搜索次序大的搜索语句的加权系数高于搜索次序小的搜索语句的加权系数;

[0022] 根据每个搜索语句对应的加权系数和所述目标搜索语句集关联的搜索事件信息中的搜索结果点击信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性。

[0023] 可选的,所述从搜索语句数据库中选取多个目标搜索语句集,包括:

[0024] 从搜索语句数据库中确定出现次数大于预设的第二数目阈值的搜索语句集,并将确定出的所述出现次数大于所述第二数目阈值的搜索语句集作为所述多个目标搜索语句集;或者,

[0025] 从搜索语句数据库中确定出现次数与所述搜索语句数据库中搜索语句总数量之间的第二比值大于预设的第二比例阈值的搜索语句集,并将确定出的所述第二比值大于所述第二比例阈值的搜索语句集作为所述多个目标搜索语句集;

[0026] 其中,搜索语句集的出现次数为该搜索语句集包括的搜索语句的出现次数之和,或者,搜索语句集的出现次数为该搜索语句集包括的搜索语句的出现次数的平均值。

[0027] 可选的,所述从搜索语句数据库中选取多个目标搜索语句集,包括:

[0028] 确定待训练的意图识别模型的应用领域信息;

[0029] 根据所述应用领域信息从所述搜索语句数据库包括的多个子数据库中确定出目标子数据库,所述子数据库与应用领域一一对应,每个子数据库包括对应的应用领域下的多个搜索语句集以及每个搜索语句集关联的搜索事件信息,所述目标子数据库对应的应用领域与所述应用领域信息指示的应用领域相同;

[0030] 从所述目标子数据库选取所述多个目标搜索语句集。

[0031] 可选的,在所述利用所述多个目标搜索语句集对应的正样本和负样本训练得到意图识别模型之前,所述方法还包括:

[0032] 计算所述正样本对应的搜索语句的数量与所述负样本对应的搜索语句的数量之间的差值的绝对值;

[0033] 判断所述绝对值是否超过预设的第三数目阈值;

[0034] 如果所述绝对值超过所述第三数目阈值,按照预设的样本平衡规则对所述正样本和/或所述负样本进行处理,以得到处理后的正样本和负样本;

[0035] 所述利用所述多个目标搜索语句集对应的正样本和负样本训练得到意图识别模型,包括:

[0036] 利用处理后的正样本和负样本训练得到所述意图识别模型。

[0037] 第二方面,本申请实施例提供了一种意图识别设备,该意图识别设备包括用于执行上述第一方面的方法的单元。

[0038] 第三方面,本申请实施例提供了另一种意图识别设备,包括处理器和存储器,所述处理器和存储器相互连接,其中,所述存储器用于存储支持意图识别设备执行上述方法的计算机程序,所述计算机程序包括程序指令,所述处理器被配置用于调用所述程序指令,执行上述第一方面的方法。可选的,该意图识别设备还可包括用户接口和/或通信接口。

[0039] 第四方面,本申请实施例提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,所述计算机程序包括程序指令,所述程序指令当被处理器执行时使所述处理器执行上述第一方面的方法。

[0040] 本申请实施例能够在获取到用户输入的目标搜索语句时,通过对该目标搜索语句进行分词处理以得到分词结果,将该分词结果输入至基于多个搜索语句集及其关联的搜索事件信息训练得到的的意图识别模型,以获取得到该目标搜索语句是否具有问答属性,进而在该目标搜索语句具有问答属性时,输出包括问答类搜索结果项的搜索结果,本申请能够根据搜索语句集关联的搜索事件信息训练得到意图识别模型以进行问答意图识别,使得提升了意图识别的准确性,问答意图识别的可靠性较高。

附图说明

[0041] 为了更清楚地说明本申请实施例技术方案,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,下面描述中的附图是本申请的一些实施例,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。

[0042] 图1是本申请实施例提供的一种意图识别方法的流程示意图;

[0043] 图2是本申请实施例提供的另一种意图识别方法的流程示意图;

[0044] 图3是本申请实施例提供的一种意图识别设备的结构示意图；

[0045] 图4是本申请实施例提供的另一种意图识别设备的结构示意图。

具体实施方式

[0046] 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，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，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，都属于本申请保护的范围。

[0047] 本申请的技术方案可应用于意图识别设备中，该意图识别设备可包括服务器、终端、机器人或其他识别设备，用于训练意图识别模型、对用户搜索语句的意图进行识别等等。本申请涉及的终端可以是手机、电脑、平板、个人计算机、智能手表等，本申请不做限定。

[0048] 具体的，本申请可通过将待进行意图识别的目标搜索语句输入至基于多个搜索语句集及其关联的搜索事件信息训练得到的意图识别模型，以获取得到该目标搜索语句的意图识别结果，确定该目标搜索语句是否具有问答属性，进而在该目标搜索语句具有问答属性时输出问答类的搜索结果，即能够根据搜索语句集关联的搜索事件信息训练得到意图识别模型以进行问答意图识别，这就提升了意图识别的准确性，使得问答意图识别的可靠性较高。以下分别详细说明。

[0049] 请参见图1，图1是本申请实施例提供的一种意图识别方法的流程示意图。具体的，本实施例的技术方案可应用于上述的意图识别设备中。如图1所示，该意图识别方法可以包括以下步骤：

[0050] 101、接收用户输入的目标搜索语句。

[0051] 其中，该目标搜索语句为待进行意图识别的搜索语句。可以理解，在其他实施例中，该目标搜索语句还可以通过其他方式获取得到，比如从搜索队列获取；该目标搜索语句可以是通过文本方式输入的，也可以是通过语音方式输入的，等等，对于该目标搜索语句的获取方式或者输入方式，本申请不做限定。

[0052] 102、对该目标搜索语句进行分词处理，以得到该目标搜索语句的分词结果。

[0053] 其中，该目标搜索语句的分词结果可包括组成该目标搜索语句的多个分词(还可称为词、词语、词条等等)。可选的，该多个分词可以是指该目标搜索语句的所有分词；或者，该多个分词可以是指该所有分词中的部分分词，比如为该所有分词中去掉无意义的分词(如去掉停用词或其他无意义的分词)后的分词，例如，可预置一个过滤列表，该过滤列表可包括各种停用词或其他无意义的词，如“啊”、“哦”、“的”等等，从而在对目标搜索语句进行分词后，能够通过与该过滤列表中的词进行匹配对比的方式确定出查询语句中的停用词等无意义的词，并去掉这些词，以减小后续确定搜索语句是否具有问答属性的检测开销；等等，此处不一一列举。

[0054] 可选的，该分词处理对应的分词方法可以为结巴分词或斯坦福分词法或其他分词方法，本申请不做限定。

[0055] 103、将该目标搜索语句的分词结果输入至预置的意图识别模型，以得到该目标搜索语句对应的意图识别结果。其中，该意图识别模型可用于识别搜索语句是否具有问答属性。该意图识别模型可以是基于多个目标搜索语句集以及该多个目标搜索语句集中每个目

标搜索语句集关联的搜索事件信息训练得到的,每个目标搜索语句集可包括至少一个搜索语句,该搜索事件信息可包括该至少一个搜索语句中每个搜索语句的搜索次序和/或每个搜索语句的搜索结果点击信息,该意图识别结果可用于指示该目标搜索语句是否具有问答属性。

[0056] 在该将该目标搜索语句的分词结果输入至预置的意图识别模型,以得到该目标搜索语句对应的意图识别结果之前,可预先训练得到该意图识别模型。具体的,可通过获取多个目标搜索语句集及其关联的搜索事件信息,比如从预置的搜索语句数据库获取,并分别根据各个目标搜索语句集中每个搜索语句的搜索次序和/或每个搜索语句的搜索结果点击信息确定出各个目标搜索语句集对应的意图,比如确定该各个目标搜索语句集包括的搜索语句是否具有问答属性,进而根据各个目标搜索语句集包括的搜索语句及其是否具有问答属性的确定结果训练得到该意图识别模型。

[0057] 其中,该搜索次序可用于指示各搜索语句的搜索先后顺序,该搜索结果点击信息可用于指示用户点击的搜索结果项的信息。可选的,该搜索次序可以是文本信息,或者为标识信息(如1、2、3…),或者该搜索事件信息中还可包括每个搜索语句的搜索时间,该每个搜索语句的搜索次序可以是通过搜索时间来指示的,等等,本申请不做限定。该搜索结果点击信息可包括搜索结果项的点击总数量、问答类的搜索结果项的点击数量、非问答类的搜索结果的点击数量(比如通过设置标签来指示为问答类还是非问答类)和/或点击的各搜索结果项的浏览时长等等。

[0058] 可选的,该多个目标搜索语句集可以为该搜索语句数据库中出现次数大于第一数目阈值的搜索语句集;或者,可以为该搜索语句数据库中所占比例大于预设比例值的搜索语句集;或者,该搜索语句数据库还可记录各搜索语句的搜索时间,选取的多个该目标搜索语句集可以为历史时间窗内如前一个月内的搜索语句集;或者,选取的多个目标搜索语句集还可以是结合待训练的意图识别模型的应用领域确定出的,或者,该选取的多个目标搜索语句集还可以是通过上述任两个或多个选取方式结合选取的,等等,此处不一一列举。从而有助于提升选取的模型训练数据的可靠性。

[0059] 其中,确定该每个目标搜索语句集包括的搜索语句是否具有问答属性,也可称为确定该每个目标搜索语句集是否具有问答属性。可选的,在确定一目标搜索语句集包括的搜索语句是否具有问答属性时,可以根据该目标搜索语句集关键的搜索事件信息中部分搜索语句的搜索结果点击信息来确定,比如根据该目标搜索语句集中搜索次序前M的搜索语句(即搜索时间最近/最晚的M个搜索语句)的搜索结果点击信息来确定,M为大于或等于1的整数;或者,可以根据该目标搜索语句集中所有搜索语句的搜索结果点击信息来确定;或者,可以根据该目标搜索语句集中各搜索语句的加权系数和各搜索语句的搜索结果点击信息来确定,等等,此处不一一列举。

[0060] 104、如果该意图识别结果指示该目标搜索语句具有问答属性,则输出包括该目标搜索语句对应的问答类搜索结果项的搜索结果。

[0061] 在确定出该目标搜索语句具有问答属性之后,即可获取问答数据,即该目标搜索结果对应的问答类的搜索结果项,并进行展示,以提供给用户。比如可将问答类的搜索结果项按照生成时间或与该目标搜索语句的相关度显示在输出界面的前面,将非问答类的搜索结果项显示在所有问答类搜索结果项之后;又如可以从问答类的搜索结果项选取部分搜索

结果项,如生成时间最新的前N项或与该目标搜索语句的相关度最高的前M项显示在输出界面,并在该N项或M项之后显示该目标搜索语句对应的非问答类的搜索结果项(如仍显示生成时间最新的前E项或与该目标搜索语句的相关度最高的前F项),其中,该N、M、E和F均为大于0的整数;又如,该输出界面可仅显示该目标搜索结果对应的问答类的搜索结果项,等等,此处不一一列举。

[0062] 在本实施例中,意图识别设备能够在获取到用户输入的目标搜索语句时,通过对该目标搜索语句进行分词处理以得到分词结果,将该分词结果输入至基于多个搜索语句集及其关联的搜索事件信息训练得到的的意图识别模型,以获取得到该目标搜索语句是否具有问答属性,进而在该目标搜索语句具有问答属性时,输出包括问答类搜索结果项的搜索结果,使得能够根据搜索语句集关联的搜索事件信息训练得到意图识别模型以进行问答意图识别,由此提升了意图识别的准确性,问答意图识别的可靠性较高。

[0063] 请参见图2,图2是本申请实施例提供的另一种意图识别方法的流程示意图。具体的,如图2所示,该意图识别方法可以包括以下步骤:

[0064] 201、从搜索语句数据库中选取多个目标搜索语句集,该搜索语句数据库中记录了多个搜索语句集以及每个搜索语句集关联的搜索事件信息。

[0065] 其中,每个搜索语句集包括一个或多个搜索语句,即包括至少一个搜索语句,该搜索事件信息包括该至少一个搜索语句中每个搜索语句的搜索次序和/或该每个搜索语句的搜索结果点击信息,此处不赘述。

[0066] 可选的,如果一搜索语句集包括多个搜索语句,该多个搜索语句中任两个搜索语句之间的搜索时间间隔不超过预设时间阈值,且所述多个搜索语句中任两个搜索语句之间的关键词(如去除无意义的词之后的其他分词)的重叠率高于预设重叠率阈值。也就是说,该搜索语句集包括的搜索语句可以是指预设时间范围内(如间隔首次搜索2分钟内)的类似搜索语句,即关键词(如对搜索语句中除去语气词和停用词的分词作为关键词)重叠率高于预设重叠率阈值的搜索语句。例如,预设的时间阈值为2min,重叠率阈值为70%,两个搜索语句的搜索时间间隔为30s,即不超过该预设时间阈值,两个搜索语句的关键词分别为5个和6个,两个搜索语句重叠的(相同的)关键词有4个,每个关键词的权值相同,未存在加权系数,即重叠率为4/5=80%(即可以取两个关键词中较小的关键词数目,在其他实施例中,还可以取较大的关键词数目,或者还可以取两者的平均值,等等,此处不一一列举),大于重叠率阈值,则可将这两个搜索语句放入同一搜索语句集。因用户在第一次搜索结果不理想的情况下,可能变换搜索语句句型或结构来进行搜索。进一步可选的,还可预先为预设关键词(如领域特有词语或出现频率较高的词语)设置加权系数,各预设关键词对应的加权系数可以相同也可以不同;进而在基于关键词确定类似搜索语句以确定搜索语句集时,可以通过匹配搜索语句中是否存在该特定关键词,如果存在该特定关键词,则可按照该特定关键词的加权系数对该特定关键词进行加权处理,或者说对该关键词重叠率进行加权处理,即增加该关键词重叠率后再进行类似搜索语句的判断,根据搜索时间和该加权处理后的搜索语句的重叠率来确定该搜索语句集。从而有助于提升搜索语句集确定的可靠性。

[0067] 进一步可选的,可选的,该多个目标搜索语句集可以为该搜索语句数据库中出现次数大于第一数目阈值的搜索语句集;或者,可以为该搜索语句数据库中所占比例大于预设比例值的搜索语句集;或者,该搜索语句数据库还可记录各搜索语句的搜索时间,选取的

多个该目标搜索语句集可以为历史时间窗内如前一个月内的搜索语句集;或者,选取的多个目标搜索语句集还可以是结合待训练的意图识别模型的应用领域确定出的,或者,该选取的多个目标搜索语句集还可以是通过上述任两个或多个选取方式结合选取的,等等,此处不一一列举。从而有助于提升选取的模型训练数据的可靠性。

[0068] 例如,在一种可能的实施方式中,在选取该多个目标搜索语句集时,意图识别设备可以从搜索语句数据库中确定出现次数大于预设的第二数目阈值的搜索语句集,并将确定出的该出现次数大于该第二数目阈值的搜索语句集作为该多个目标搜索语句集;或者,可以从搜索语句数据库中确定出现次数与该搜索语句数据库中搜索语句总数量之间的第二比值大于预设的第二比例阈值的搜索语句集,并将确定出的该第二比值大于该第二比例阈值的搜索语句集作为该多个目标搜索语句集;或者,还可以从搜索语句数据库中确定出现次数大于预设的第二数目阈值,且第二比值大于预设的第二比例阈值的搜索语句集,并将确定出的该出现次数大于该第二数目阈值且第二比值大于预设的第二比例阈值的搜索语句集作为该多个目标搜索语句集,等等,此处不一一列举。其中,搜索语句集的出现次数可以为该搜索语句集包括的搜索语句的出现次数之和,或者,搜索语句集的出现次数可以为该搜索语句集包括的搜索语句的出现次数的平均值,或者,搜索语句集的出现次数可以为该搜索语句集包括的搜索语句的最高出现次数,等等,此处不一一列举。搜索语句出现次数可以是指该搜索数据库中该搜索语句的数目或该搜索数据库中与该搜索语句的相似度高于阈值的搜索语句的数目等等,本申请不做限定。

[0069] 又如,在一种可能的实施方式中,在选取该多个目标搜索语句集时,意图识别设备可以确定待训练的意图识别模型的应用领域信息,并根据该应用领域信息从该搜索语句数据库包括的多个子数据库中确定出目标子数据库,进而从该目标子数据库选取所述多个目标搜索语句集。其中,该子数据库与应用领域一一对应,每个子数据库包括对应的应用领域下的多个搜索语句集(其数量大于选取的目标搜索语句集的数量)以及每个搜索语句集关联的搜索事件信息,该目标子数据库对应的应用领域与该应用领域信息指示的应用领域相同。也就是说,该搜索语句数据库可包括各应用领域下的子数据库,每个子数据库包括一应用领域下的搜索语句集以及每个搜索语句集关联的每个搜索语句的搜索次序和搜索结果点击信息等等,从而在选取目标搜索语句集时,可以通过确定待训练的意图识别模型的应用领域信息(如领域标签)来确定子数据库(如携带有该领域标签的子数据库),并从中选取目标搜索语句集。从而能够进一步提升选取的模型训练数据的可靠性,进而提升训练效果。

[0070] 202、分别对该多个目标搜索语句集中每个目标搜索语句集包括的搜索语句进行分词处理,以得到该每个目标搜索语句集的分词结果。

[0071] 其中,该每个目标搜索语句集的分词结果包括组成该目标搜索语句集的搜索语句的多个分词,该多个分词可以是指该目标搜索语句集的搜索语句的所有分词,或者可以是指该所有分词中的部分分词,比如为该所有分词中去掉无意义的分词(如去掉停用词或其他无意义的分词)后的分词,例如,可预置一个过滤列表,该过滤列表可包括各种停用词或其他无意义的词,如“啊”、“哦”、“的”等等,从而在对目标搜索语句集的搜索语句进行分词后,能够通过与该过滤列表中的词进行匹配对比的方式确定出查询语句中的停用词等无意义的词,并去掉这些词,以减小后续确定搜索语句是否具有问答属性的检测开销;或者,该多个分词可以是指该目标搜索语句集中搜索次序最大(即最近一次搜索)的搜索语句的分

词(可以是该搜索次序最大的搜索语句的所有分词或部分分词,此处不赘述),等等,此处不赘述。

[0072] 203、根据该每个目标搜索语句集关联的搜索事件信息,确定该每个目标搜索语句集包括的搜索语句是否具有问答属性。

[0073] 可选的,该搜索结果点击信息可包括搜索结果项的点击总数量和问答类的搜索结果项的点击数量;在根据搜索语句的搜索结果点击信息确定搜索语句是否具有问答属性时,可以通过将该搜索语句的搜索结果点击信息包括的搜索结果项的点击总数量与预设的第一数目阈值进行比较,以及计算该搜索语句的搜索结果点击信息包括的问答类的搜索结果项的点击数量与搜索结果项的点击总数量之间的第一比值,将该第一比值与预设的第一比例值进行比较;如果该搜索结果项的点击总数量大于预设的第一数目阈值,且该第一比值大于预设的第一比例阈值,则可确定该搜索语句具有问答属性;否则,可表明不具有问答属性(或者可结合其余方式进一步判断)。或者,可选的,该搜索结果点击信息可包括搜索结果项的点击总数量、问答类的搜索结果项的点击数量和点击的各搜索结果项的浏览时长;在根据搜索语句的搜索结果点击信息确定搜索语句是否具有问答属性时,可以过滤掉浏览时长小于预设时长阈值的搜索结果项,并确定过滤该搜索结果项之后剩余的搜索结果项的点击总数量(即搜索结果点击信息包括的搜索结果项的点击总数量减去浏览时长小于预设时长阈值的搜索结果项的数量),确定过滤该搜索结果项之后剩余的问答类的搜索结果项的点击数量(即搜索结果点击信息包括的问答类的搜索结果项的点击数量减去浏览时长小于预设时长阈值的搜索结果项的数量),以及计算该剩余的问答类的搜索结果项的点击数量与该剩余的搜索结果项的点击总数量之间的第一比值;如果该剩余的搜索结果项的点击总数量大于预设的第一数目阈值,且该第一比值大于预设的第一比例阈值,则可确定该搜索语句具有问答属性。或者,可选的,在根据搜索语句的搜索结果点击信息确定搜索语句是否具有问答属性时,还可以通过将该搜索语句的搜索结果点击信息包括的问答类的搜索结果项的点击数量与预设的另一数目阈值进行比较;如果该问答类的搜索结果项的点击数量大于该另一数目阈值,则可确定该搜索语句具有问答属性,等等,此处不一一列举。

[0074] 例如,在一种可能的实施方式中,在确定该目标搜索语句集包括的搜索语句是否具有问答属性时,可以根据该目标搜索语句集关联的搜索事件信息所包括的每个搜索语句的搜索次序,确定该目标搜索语句集包括的该至少一个搜索语句中最大搜索次序对应的搜索语句;根据该最大搜索次序对应的搜索语句的搜索结果点击信息,确定该目标搜索语句集包括的搜索语句是否具有问答属性。根据该最大搜索次序对应的搜索语句的搜索结果点击信息,确定该目标搜索语句集包括的搜索语句是否具有问答属性的方式,可参照上述根据搜索语句的搜索结果点击信息确定搜索语句是否具有问答属性的方式,此处不赘述。如果该最大搜索次序对应的搜索语句具有问答属性,即可确定该目标搜索语句集包括的搜索语句具有问答属性。也就是说,在确定一目标搜索语句是否具有问答属性时,可以从该关联的搜索事件中最近的一次搜索事件,即最大搜索次数对应的搜索语句的搜索事件,根据该最大搜索次数下用户点击的搜索结果信息确定该目标搜索语句是否具有问答属性,因前面几次搜索得到的搜索结果可能并不是用户想要的,则可以以后续的点击为准,以提升判断效率,并确保判断准确性。

[0075] 又如,在一种可能的实施方式中,在确定该目标搜索语句集包括的搜索语句是否

具有问答属性时,可以根据该目标搜索语句集中所有搜索语句的搜索结果点击信息来确定,具体可参照上述根据搜索语句的搜索结果点击信息确定搜索语句是否具有问答属性的方式,比如统计该所有搜索语句的搜索结果点击信息中问答类的搜索结果项的点击数量之和,并判断该点击数量之和是否超过预设数目阈值,如果超过,则可表明该目标搜索语句集包括的搜索语句具有问答属性,等等,此处不赘述。

[0076] 又如,在一种可能的实施方式中,可以预先设置得到各搜索语句的加权系数,比如包括疑问词的搜索语句的加权系数高于未包括疑问词的搜索语句的加权系数,和/或,搜索次序大的搜索语句的加权系数高于搜索次序小的搜索语句的加权系数(即搜索次序越大的搜索语句,其加权系数越高),和/或,对于搜索结果点击信息包括的搜索结果点击项中存在特定问答网站的展示结果或搜索结果中存在特定问答网站的展示结果的搜索语句,其加权系数高于不存在该特定问答网站的展示结果的搜索语句的加权系数,等等。在该目标搜索语句集包括的搜索语句是否具有问答属性时,可以确定该目标搜索语句集包括的该至少一个搜索语句中每个搜索语句对应的加权系数;根据每个搜索语句对应的加权系数和该目标搜索语句集关联的搜索事件信息中的搜索结果点击信息,确定该目标搜索语句集包括的搜索语句是否具有问答属性。根据该加权系数和搜索结果点击信息可以是指通过该加权系数对搜索结果点击信息的参数如问答类的搜索结果点击项的数量、浏览时长等等进行加权,加权后,其问答属性确定方式具体可参照上述根据搜索语句的搜索结果点击信息确定搜索语句是否具有问答属性的方式。例如,可以通过各搜索语句的加权系数对各搜索语句对应的问答类搜索结果项的点击数量进行加权处理(如每一搜索语句对应的问答类搜索结果项的点击数量为2,加权系数为1.5,加权后的问答类搜索结果项的点击数量则为2*1.5=3);如果该目标搜索语句集中各搜索语句的搜索结果项的点击总数量大于预设的第一数目阈值,且各搜索语对应的问答类的搜索结果项的点击数量(加权后)之和与各搜索语句的搜索结果项的点击总数量之间的第一比值大于预设的第一比例阈值,则可确定该目标搜索语句集包括的搜索语句具有问答属性。也就是说,在确定一目标搜索语句集是否具有问答属性时,可以根据每次搜索次数下用户点击的搜索结果信息以及每次搜索结果的权重来确定该目标搜索语句是否具有问答属性。从而有助于提升确定出的搜索语句集的问答属性的可靠性。

[0077] 可选的,在本申请中,该步骤202和203的执行顺序不受限定,比如还可先执行步骤203,再执行步骤202,或者,该步骤202和203可同时执行,本申请不做限定。

[0078] 204、将该多个目标搜索语句集中具有问答属性的目标搜索语句集的分词结果作为正样本,以及将该多个目标搜索语句集中不具有问答属性的目标搜索语句集的分词结果作为负样本。

[0079] 可选的,该多个目标搜索语句集中具有问答属性的目标搜索语句集的分词结果可以包括该多个目标搜索语句集中具有问答属性的搜索语句的分词,其可以是某一个或多个目标搜索语句集的所有分词,也可以是其中的部分分词。也即,该正样本可包括具有问答属性的搜索语句的分词,该负样本可包括不具有问答属性的搜索语句的分词。比如确定某一目标搜索语句集的搜索语句具有问答属性时,可以将该目标搜索语句集的搜索语句的所有分词(可去除无意义的分词)作为正样本。

[0080] 在一些实施例中,可以将具有问答属性的搜索语句的分词作为正训练样本,将不

具有问答属性的搜索语句的分词作为负训练样本,以基于该正训练样本和负训练样本训练得到意图识别模型,使得后续能够通过该意图识别模型快速识别输入的搜索语句是否具有问答属性。进而能够根据识别出的该输入的搜索语句是否具有问答属性的识别结果,来为用户返回信息。例如,该具有问答属性的搜索语句可表明具有问答需求,不具有问答属性的搜索语句可表明不具有问答需求,由此可根据搜索语句是否具有问答需求向用户返回不同的页面(界面),提供不同的需求内容。

[0081] 可选的,该步骤201-204的描述和上述图1所示实施例的相关描述可相互参照,此处不赘述。

[0082] 205、计算该正样本对应的搜索语句的数量与该负样本对应的搜索语句的数量之间的差值的绝对值。

[0083] 206、判断该绝对值是否超过预设的数目阈值。

[0084] 207、如果该绝对值超过该数目阈值,按照预设的样本平衡规则对该正样本和/或该负样本进行处理,以得到处理后的正样本和负样本。

[0085] 可选的,在确定该多个目标搜索语句集对应的正样本和负样本之后,还可进一步确定正负样本的数量是否平衡,比如判断该正样本对应的搜索语句的数量与该负样本对应的搜索语句的数量之间的差值的绝对值是否超过预设的数目阈值如预设的第三数目阈值,如果超过,则可表明该正负样本的数量不平衡。因在训练模型时,很多时候正负样本不平衡,导致训练出的模型识别准确性较差,因容易对比例大的样本造成过拟合,也就是说预测容易偏向样本数较多的分类,这就大大降低了模型的范化能力,导致其识别结果不可靠。因此,在训练之前,可分别统计正负样本的数量,在两者差距过大如相差超过预设的第三数目阈值时,可以按照预设的样本平衡规则平衡正负样本的数量之后再进行训练。该预设的样本平衡规则可以为多种,具体可根据正负样本的数量进行选择,或者根据训练场景进行选择。例如,针对正样本较少的情况,可以采用增加正样本的方式来平衡正负样本;又如,针对负样本较少的情况,可以采用增加负样本来平衡正负样本;又如,针对需要大量样本进行训练的场景(如场景标签为多样本),可采用合成样本的方式平衡正负样本;又如,针对可靠性要求较高的场景(如场景标签为高可靠性),可采用改变样本权重的方式平衡正负样本,具体可预先设置得到该各样本平衡规则以及选择使用的场景等。可选的,平衡正负样本的方式可以如下:

[0086] 1)上采样:增加样本数较少的样本,其方式是直接复制原来的样本。比如可以在样本较少时采用。

[0087] 2)下采样:减少样本数较多的样本,其方式是丢弃这些多余的样本。比如可以在样本较多时采用。如可按照点击总数量对目标搜索语句进行排序,丢弃点击总数量较少的目标搜索语句对应的样本。

[0088] 3)合成样本:增加样本数目较少的那一类的样本,合成指的是通过组合已有的样本的各个特征以产生新的样本。具体的,该产生新样本的方式可以是从各个特征中随机选出一些特征或者通过一些方式选出某些特定的特征(如出现次数高于阈值的特征,或者样本相似度高于阈值如欧氏距离小于阈值的样本之间的特征等等)之后,然后拼接成一个新的样本,从而增加了样本数目较少的类别的样本数。不同于上采样是单纯的复制样本,而这里则是拼接得到新的样本,使得能够进一步提升模型训练的可靠性。

[0089] 4)改变样本权重:增大关键分词的权重,假如对于正样本,对于具有明显问答属性的分词可以乘上一个权重,以提升判断可靠性。

[0090] 在得到正负样本之后,即可训练得到意图识别模型,使得后续能够根据该意图识别模型快速识别出输入的搜索语句是否具有问答意图。

[0091] 208、利用处理后的正样本和负样本训练得到该意图识别模型。

[0092] 其中,该意图识别模型可用于识别输入的搜索语句是否具有问答属性。该模型可以是基于二叉树的模型,也可以是基于多叉树的模型,也可以是神经网络模型,等等,本申请不做限定。

[0093] 可选的,在确定出具有问答意图的搜索语句(集)之后,还可进一步确定出该搜索语句的问答类别,比如是属于包括疑问词的显式问答搜索语句,还是属于不包括疑问词的隐式问答搜索语句。并可基于上述的正负样本和各正样本所属的问答类别(如类别标签)训练得到该意图识别模型。使得后续在利用该意图识别模型识别搜索语句的问答属性时,不仅能够识别出问答类的搜索语句,即具有问答意图的搜索语句,还可确定出该问答意图所属的问答类别。进一步可选的,可预先设置得到各问答类别和展示内容/页面(的关键词或内容标题格式)等的对应的关系,进而可根据问答类别区别为用户展示内容,提升了页面展示的灵活性。

[0094] 209、接收用户输入的目标搜索语句。

[0095] 210、对该目标搜索语句进行分词处理,以得到该目标搜索语句的分词结果。

[0096] 211、将该目标搜索语句的分词结果输入至预置的意图识别模型,以得到该目标搜索语句对应的意图识别结果。

[0097] 212、如果该意图识别结果指示该目标搜索语句具有问答属性,则输出包括该目标搜索语句对应的问答类搜索结果项的搜索结果。

[0098] 可选的,该步骤209-212的描述和上述图1所示实施例中步骤101-104的相关描述可相互参照,此处不赘述。

[0099] 在本实施例中,意图识别设备能够通过对选取的多个目标搜索语句集包括的搜索语句进行分词处理,以得到每个目标搜索语句集的分词结果,并根据每个目标搜索语句集关联的搜索事件信息确定每个目标搜索语句集包括的搜索语句是否具有问答属性,进而能够将该多个目标搜索语句集中具有问答属性的搜索语句的分词结果作为正样本以及将不具有问答属性的搜索语句的分词结果作为负样本,并按照预设的样本平衡规则对正负样本进行平衡后,基于该平衡后的正负样本训练得到意图识别模型以进行问答意图识别,使得能够通过将获取的搜索语句输入至该意图识别模型以识别该目标搜索语句是否具有问答属性,进而在该目标搜索语句具有问答属性时,输出包括问答类搜索结果项的搜索结果,这就提升了意图识别的准确性,使得问答意图识别的可靠性和召回率较高。

[0100] 上述方法实施例都是对本申请的意图识别方法的举例说明,对各个实施例的描述都各有侧重,某个实施例中没有详述的部分,可以参见其他实施例的相关描述。

[0101] 请参见图3,图3是本申请实施例提供的一种意图识别设备的结构示意图。本申请实施例的意图识别设备包括用于执行上述意图识别方法的单元。具体的,本实施例的意图识别设备300可包括:获取单元301和处理单元302。其中,

[0102] 获取单元301,用于接收用户输入的目标搜索语句;

[0103] 处理单元302,用于对所述目标搜索语句进行分词处理,以得到所述目标搜索语句的分词结果,所述目标搜索语句的分词结果包括组成所述目标搜索语句的多个分词;

[0104] 处理单元302,还用于将所述目标搜索语句的分词结果输入至预置的意图识别模型,以得到所述目标搜索语句对应的意图识别结果,所述意图识别模型是基于多个目标搜索语句集以及所述多个目标搜索语句集中每个目标搜索语句集关联的搜索事件信息训练得到的,所述每个目标搜索语句集包括至少一个搜索语句,所述搜索事件信息包括所述至少一个搜索语句中每个搜索语句的搜索次序和/或所述每个搜索语句的搜索结果点击信息,所述意图识别结果用于指示所述目标搜索语句是否具有问答属性;

[0105] 处理单元302,还用于如果所述意图识别结果指示所述目标搜索语句具有问答属性,则输出包括所述目标搜索语句对应的问答类搜索结果项的搜索结果。

[0106] 可选的,获取单元301,还用于从搜索语句数据库中选取多个目标搜索语句集;其中,所述搜索语句数据库中记录了多个搜索语句集以及每个搜索语句集关联的搜索事件信息,所述每个搜索语句集包括至少一个搜索语句,所述搜索事件信息包括所述至少一个搜索语句中每个搜索语句的搜索次序和/或所述每个搜索语句的搜索结果点击信息;

[0107] 处理单元302,还用于分别对所述多个目标搜索语句集中每个目标搜索语句集包括的搜索语句进行分词处理,以得到所述每个目标搜索语句集的分词结果,所述每个目标搜索语句集的分词结果包括组成该目标搜索语句集的搜索语句的多个分词;

[0108] 处理单元302,还可用于根据所述每个目标搜索语句集关联的搜索事件信息,确定所述每个目标搜索语句集包括的搜索语句是否具有问答属性;将所述多个目标搜索语句集中具有问答属性的目标搜索语句集的分词结果作为正样本,以及将所述多个目标搜索语句集中不具有问答属性的目标搜索语句集的分词结果作为负样本,并利用所述多个目标搜索语句集对应的正样本和负样本训练得到意图识别模型;其中,所述意图识别模型用于识别输入的搜索语句是否具有问答属性。

[0109] 可选的,处理单元302,可具体用于根据所述目标搜索语句集关联的搜索事件信息所包括的每个搜索语句的搜索次序,确定所述目标搜索语句集包括的所述至少一个搜索语句中最大搜索次序对应的搜索语句;根据所述最大搜索次序对应的搜索语句的搜索结果点击信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性。

[0110] 可选的,所述搜索结果点击信息包括搜索结果项的点击总数量和问答类的搜索结果项的点击数量;

[0111] 处理单元302,在所述根据所述最大搜索次序对应的搜索语句的搜索结果点击信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性时,可具体用于:计算所述最大搜索次序对应的搜索语句的搜索结果点击信息包括的问答类的搜索结果项的点击数量与搜索结果项的点击总数量之间的第一比值;如果所述搜索结果项的点击总数量大于预设的第一数目阈值,且所述第一比值大于预设的第一比例阈值,确定所述目标搜索语句集包括的搜索语句具有问答属性。

[0112] 可选的,处理单元302,可具体用于确定所述目标搜索语句集包括的所述至少一个搜索语句中每个搜索语句对应的加权系数,所述至少一个搜索语句中搜索次序大的搜索语句的加权系数高于搜索次序小的搜索语句的加权系数;根据每个搜索语句对应的加权系数和所述目标搜索语句集关联的搜索事件信息中的搜索结果点击信息,确定所述目标搜索语

句集包括的搜索语句是否具有问答属性。

[0113] 可选的,获取单元301,可具体用于:从搜索语句数据库中确定出现次数大于预设的第二数目阈值的搜索语句集,并将确定出的所述出现次数大于所述第二数目阈值的搜索语句集作为所述多个目标搜索语句集;或者,从搜索语句数据库中确定出现次数与所述搜索语句数据库中搜索语句总数量之间的第二比值大于预设的第二比例阈值的搜索语句集,并将确定出的所述第二比值大于所述第二比例阈值的搜索语句集作为所述多个目标搜索语句集。

[0114] 其中,搜索语句集的出现次数为该搜索语句集包括的搜索语句的出现次数之和,或者,搜索语句集的出现次数为该搜索语句集包括的搜索语句的出现次数的平均值。

[0115] 可选的,获取单元301,可具体用于:确定待训练的意图识别模型的应用领域信息;根据所述应用领域信息从所述搜索语句数据库包括的多个子数据库中确定出目标子数据库;从所述目标子数据库选取所述多个目标搜索语句集。

[0116] 其中,所述子数据库与应用领域一一对应,每个子数据库包括对应的应用领域下的多个搜索语句集以及每个搜索语句集关联的搜索事件信息,所述目标子数据库对应的应用领域与所述应用领域信息指示的应用领域相同。

[0117] 可选的,处理单元302,还可用于在所述利用所述多个目标搜索语句集对应的正样本和负样本训练得到意图识别模型之前,计算所述正样本对应的搜索语句的数量与所述负样本对应的搜索语句的数量之间的差值的绝对值;判断所述绝对值是否超过预设的第三数目阈值;如果所述绝对值超过所述第三数目阈值,按照预设的样本平衡规则对所述正样本和/或所述负样本进行处理,以得到处理后的正样本和负样本;

[0118] 处理单元302,可具体用于利用处理后的正样本和负样本训练得到所述意图识别模型。

[0119] 具体的,该意图识别设备可通过上述单元实现上述图1至图2所示实施例中的意图识别方法中的部分或全部步骤。应理解,本申请实施例是对应方法实施例的装置实施例,对方法实施例的描述,也适用于本申请实施例,此处不赘述。

[0120] 请参见图4,图4是本申请实施例提供的另一种意图识别设备的结构示意图。该意图识别设备用于执行上述的方法。如图4所示,本实施例中的意图识别设备400可以包括:一个或多个处理器401和存储器402。可选的,该意图识别设备还可包括一个或多个用户接口403,和/或,一个或多个通信接口404。上述处理器401、用户接口403、通信接口404和存储器402可通过总线405连接,或者可以通过其他方式连接,图4中以总线方式进行示例说明。其中,存储器402用于存储计算机程序,所述计算机程序包括程序指令,处理器401用于执行存储器402存储的程序指令。

[0121] 其中,处理器401可用于调用所述程序指令执行以下步骤:调用用户接口403接收用户输入的目标搜索语句;对所述目标搜索语句进行分词处理,以得到所述目标搜索语句的分词结果,所述目标搜索语句的分词结果包括组成所述目标搜索语句的多个分词;将所述目标搜索语句的分词结果输入至预置的意图识别模型,以得到所述目标搜索语句对应的意图识别结果,所述意图识别模型是基于多个目标搜索语句集以及所述多个目标搜索语句集中每个目标搜索语句集关联的搜索事件信息训练得到的,所述每个目标搜索语句集包括至少一个搜索语句,所述搜索事件信息包括所述至少一个搜索语句中每个搜索语句的搜索

次序和/或所述每个搜索语句的搜索结果点击信息,所述意图识别结果用于指示所述目标搜索语句是否具有问答属性;如果所述意图识别结果指示所述目标搜索语句具有问答属性,则调用用户接口403输出包括所述目标搜索语句对应的问答类搜索结果项的搜索结果。

[0122] 可选的,处理器401在执行所述将所述目标搜索语句的分词结果输入至预置的意图识别模型,以得到所述目标搜索语句对应的意图识别结果之前,还可执行以下步骤:从搜索语句数据库中选取多个目标搜索语句集;其中,所述搜索语句数据库中记录了多个搜索语句集以及每个搜索语句集关联的搜索事件信息,所述每个搜索语句集包括至少一个搜索语句,所述搜索事件信息包括所述至少一个搜索语句中每个搜索语句的搜索次序和/或所述每个搜索语句的搜索结果点击信息;分别对所述多个目标搜索语句集中每个目标搜索语句集包括的搜索语句进行分词处理,以得到所述每个目标搜索语句集的分词结果,所述每个目标搜索语句集的分词结果包括组成该目标搜索语句集的搜索语句的多个分词;根据所述每个目标搜索语句集关联的搜索事件信息,确定所述每个目标搜索语句集包括的搜索语句是否具有问答属性;将所述多个目标搜索语句集中具有问答属性的搜索语句的分词结果作为正样本,以及将所述多个目标搜索语句集中不具有问答属性的搜索语句的分词结果作为负样本,并利用所述多个目标搜索语句集对应的正样本和负样本训练得到意图识别模型;其中,所述意图识别模型用于识别输入的搜索语句是否具有问答属性。

[0123] 可选的,处理器401在执行所述根据所述目标搜索语句集关联的搜索事件信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性时,可具体执行以下步骤:根据所述目标搜索语句集关联的搜索事件信息所包括的每个搜索语句的搜索次序,确定所述目标搜索语句集包括的所述至少一个搜索语句中最大搜索次序对应的搜索语句;根据所述最大搜索次序对应的搜索语句的搜索结果点击信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性。

[0124] 可选的,处理器401在执行所述根据所述目标搜索语句集关联的搜索事件信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性时,可具体执行以下步骤:确定所述目标搜索语句集包括的所述至少一个搜索语句中每个搜索语句对应的加权系数;根据每个搜索语句对应的加权系数和所述目标搜索语句集关联的搜索事件信息中的搜索结果点击信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性。

[0125] 进一步可选的,所述至少一个搜索语句中搜索次序大的搜索语句的加权系数高于搜索次序小的搜索语句的加权系数,和/或,所述至少一个搜索语句中包括疑问词的搜索语句的加权系数高于未包括疑问词的搜索语句的加权系数,等等,此处不赘述。

[0126] 可选的,所述搜索结果点击信息包括搜索结果项的点击总数量和问答类的搜索结果项的点击数量;

[0127] 处理器401在执行所述根据所述最大搜索次序对应的搜索语句的搜索结果点击信息,确定所述目标搜索语句集包括的搜索语句是否具有问答属性时,可具体执行以下步骤:计算所述最大搜索次序对应的搜索语句的搜索结果点击信息包括的问答类的搜索结果项的点击数量与搜索结果项的点击总数量之间的第一比值;如果所述搜索结果项的点击总数量大于预设的第一数目阈值,且所述第一比值大于预设的第一比例阈值,确定所述目标搜索语句集包括的搜索语句具有问答属性。

[0128] 可选的,处理器401在执行所述从搜索语句数据库中选取多个目标搜索语句集时,

可具体执行以下步骤:从搜索语句数据库中确定出现次数大于预设的第二数目阈值的搜索语句集,并将确定出的所述出现次数大于所述第二数目阈值的搜索语句集作为所述多个目标搜索语句集;或者,从搜索语句数据库中确定出现次数与所述搜索语句数据库中搜索语句总数量之间的第二比值大于预设的第二比例阈值的搜索语句集,并将确定出的所述第二比值大于所述第二比例阈值的搜索语句集作为所述多个目标搜索语句集;

[0129] 其中,搜索语句集的出现次数为该搜索语句集包括的搜索语句的出现次数之和,或者,搜索语句集的出现次数为该搜索语句集包括的搜索语句的出现次数的平均值。

[0130] 可选的,处理器401在执行所述从搜索语句数据库中选取多个目标搜索语句集时,可具体执行以下步骤:确定待训练的意图识别模型的应用领域信息;根据所述应用领域信息从所述搜索语句数据库包括的多个子数据库中确定出目标子数据库;从所述目标子数据库选取所述多个目标搜索语句集。

[0131] 其中,所述子数据库与应用领域一一对应,每个子数据库包括对应的应用领域下的多个搜索语句集以及每个搜索语句集关联的搜索事件信息,所述目标子数据库对应的应用领域与所述应用领域信息指示的应用领域相同。

[0132] 可选的,处理器401在执行所述利用所述多个目标搜索语句集对应的正样本和负样本训练得到意图识别模型之前,还可执行以下步骤:计算所述正样本对应的搜索语句的数量与所述负样本对应的搜索语句的数量之间的差值的绝对值;判断所述绝对值是否超过预设的第三数目阈值;如果所述绝对值超过所述第三数目阈值,按照预设的样本平衡规则对所述正样本和/或所述负样本进行处理,以得到处理后的正样本和负样本;

[0133] 处理器401在执行所述利用所述多个目标搜索语句集对应的正样本和负样本训练得到意图识别模型时,可具体执行以下步骤:利用处理后的正样本和负样本训练得到所述意图识别模型。

[0134] 其中,所述处理器401可以是中央处理单元(Central Processing Unit,CPU),该处理器还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0135] 用户接口403可包括输入设备和输出设备,输入设备可以包括触控板、麦克风等,输出设备可以包括显示器(LCD等)、扬声器等。

[0136] 通信接口404可包括接收器和发射器,用于与其他设备进行通信。

[0137] 存储器402可以包括只读存储器和随机存取存储器,并向处理器401提供指令和数据。存储器402的一部分还可以包括非易失性随机存取存储器。例如,存储器402还可以存储上述的多个搜索语句集、每个搜索语句集关联的搜索事件信息等等。

[0138] 具体实现中,本申请实施例中所描述的处理器401等可执行上述图1至图2所示的方法实施例中所描述的实现方式,也可执行本申请实施例图3所描述的各单元的实现方式,此处不赘述。

[0139] 本申请实施例还提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,所述计算机程序被处理器执行时可实现图1至图2所对应实施例中描述的意

图识别方法中的部分或全部步骤,也可实现本申请图3或图4所示实施例的意图识别设备的功能,此处不赘述。

[0140] 本申请实施例还提供了一种包含指令的计算机程序产品,当其在计算机上运行时,使得计算机执行上述方法中的部分或全部步骤。

[0141] 所述计算机可读存储介质可以是前述任一实施例所述的意图识别设备的内部存储单元,例如意图识别设备的硬盘或内存。所述计算机可读存储介质也可以是所述意图识别设备的外部存储设备,例如所述意图识别设备上配备的插接式硬盘,智能存储卡(Smart Media Card,SMC),安全数字(Secure Digital,SD)卡,闪存卡(Flash Card)等。

[0142] 在本申请中,术语“和/或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系。

[0143] 在本申请的各种实施例中,上述各过程的序号的大小并不意味着执行顺序的先后,各过程的执行顺序应以其功能和内在逻辑确定,而不应对本申请实施例的实施过程构成任何限定。

[0144] 以上所述,仅为本申请的部分实施方式,但本申请的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到各种等效的修改或替换,这些修改或替换都应涵盖在本申请的保护范围之内。

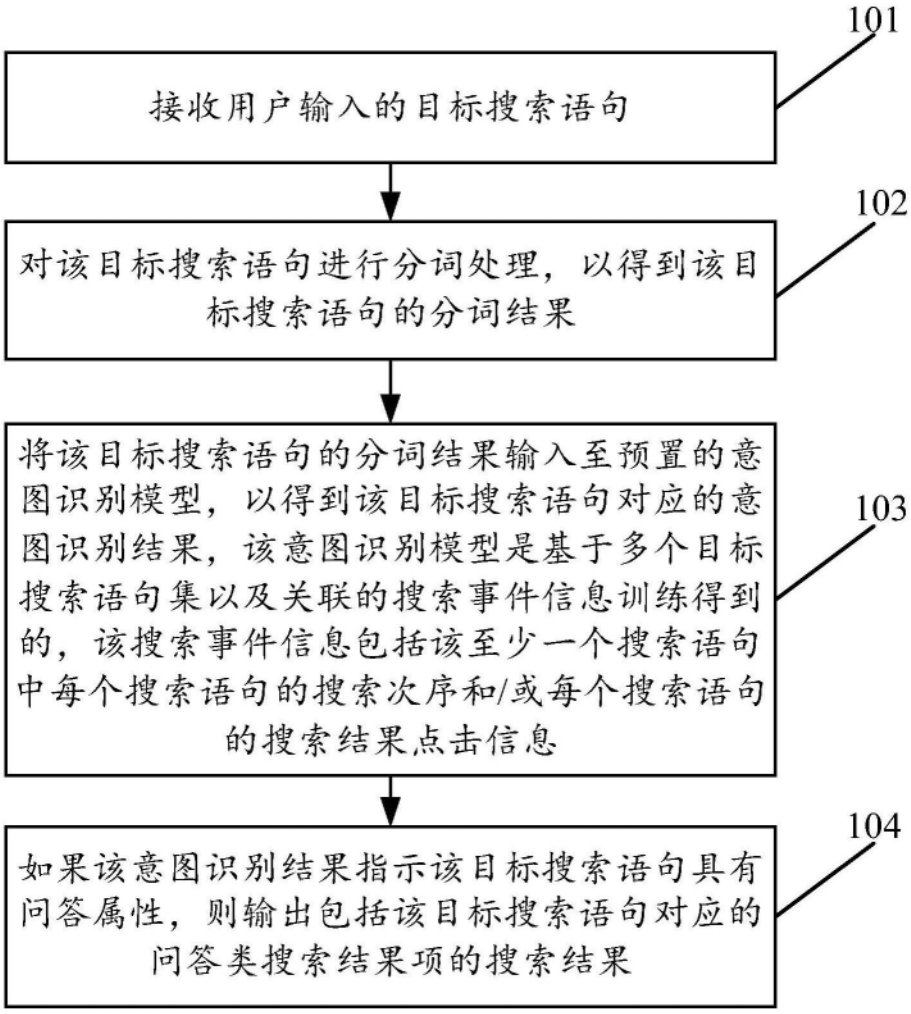

101
接收用户输入的目标搜索语句
102
对该目标搜索语句进行分词处理，以得到该目标搜索语句的分词结果
103
将该目标搜索语句的分词结果输入至预置的意图识别模型，以得到该目标搜索语句对应的意图识别结果，该意图识别模型是基于多个目标搜索语句集以及关联的搜索事件信息训练得到的，该搜索事件信息包括该至少一个搜索语句中每个搜索语句的搜索次序和/或每个搜索语句的搜索结果点击信息
104
如果该意图识别结果指示该目标搜索语句具有问答属性，则输出包括该目标搜索语句对应的问答类搜索结果项的搜索结果

图1

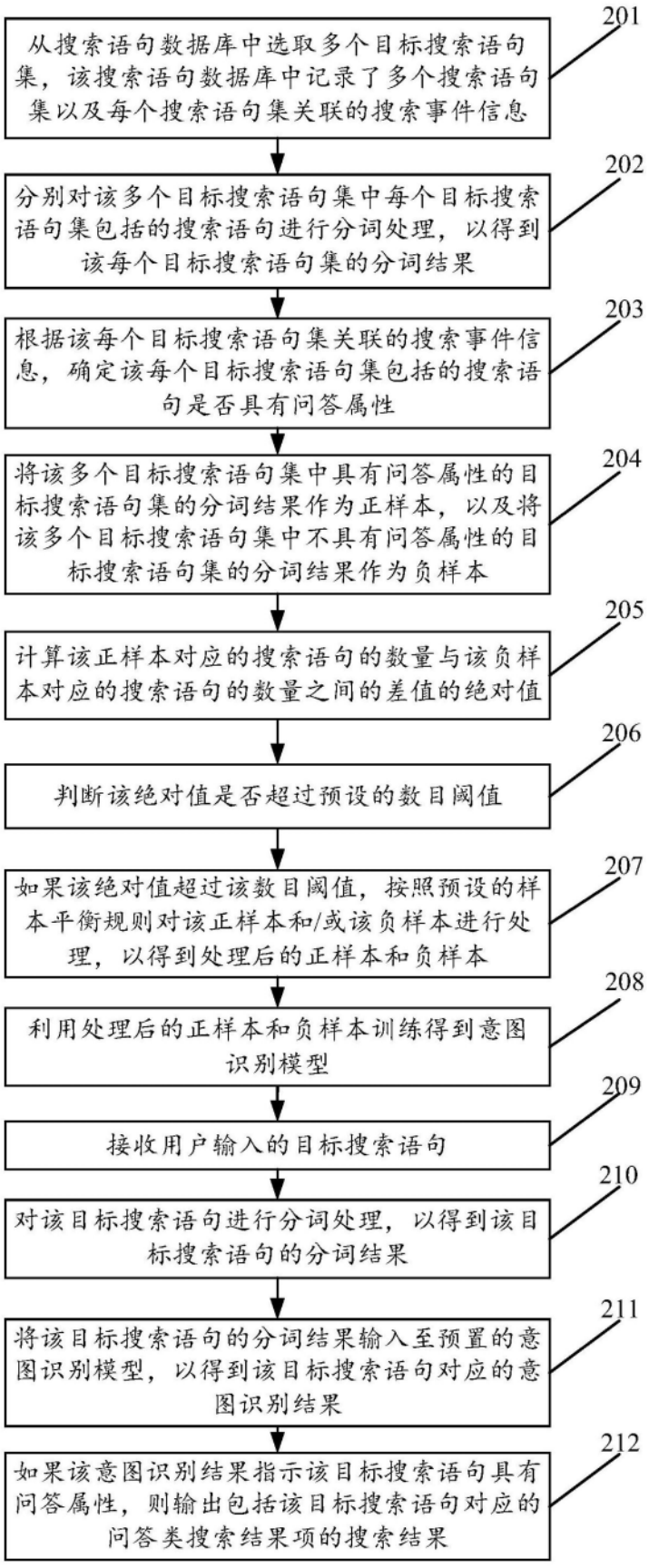

201
从搜索语句数据库中选取多个目标搜索语句集，该搜索语句数据库中记录了多个搜索语句集以及每个搜索语句集关联的搜索事件信息
202
分别对该多个目标搜索语句集中每个目标搜索语句集包括的搜索语句进行分词处理，以得到该每个目标搜索语句集的分词结果
203
根据该每个目标搜索语句集关联的搜索事件信息，确定该每个目标搜索语句集包括的搜索语句是否具有问答属性
204
将该多个目标搜索语句集中具有问答属性的目标搜索语句集的分词结果作为正样本，以及将该多个目标搜索语句集中不具有问答属性的目标搜索语句集的分词结果作为负样本
205
计算该正样本对应的搜索语句的数量与该负样本对应的搜索语句的数量之间的差值的绝对值
206
判断该绝对值是否超过预设的数目阈值
207
如果该绝对值超过该数目阈值，按照预设的样本平衡规则对该正样本和/或该负样本进行处理，以得到处理后的正样本和负样本
208
利用处理后的正样本和负样本训练得到意图识别模型
209
接收用户输入的目标搜索语句
210
对该目标搜索语句进行分词处理，以得到该目标搜索语句的分词结果
211
将该目标搜索语句的分词结果输入至预置的意图识别模型，以得到该目标搜索语句对应的意图识别结果
212
如果该意图识别结果指示该目标搜索语句具有问答属性，则输出包括该目标搜索语句对应的问答类搜索结果项的搜索结果

图2

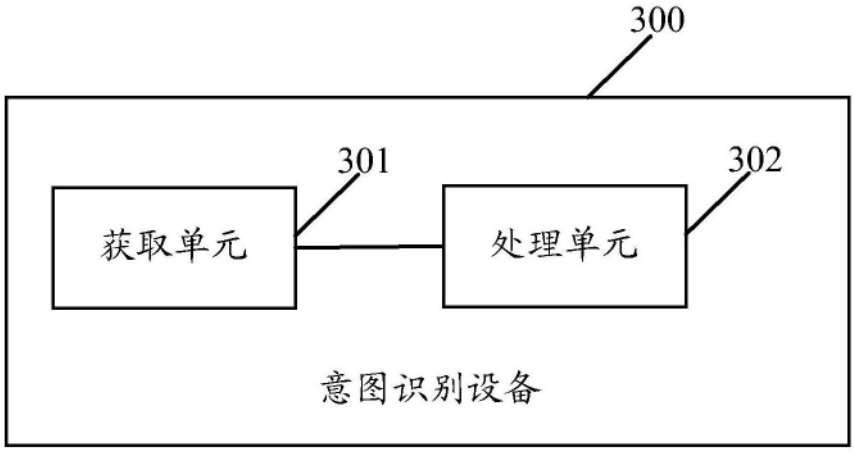

300
301
302
获取单元
处理单元
意图识别设备

图3

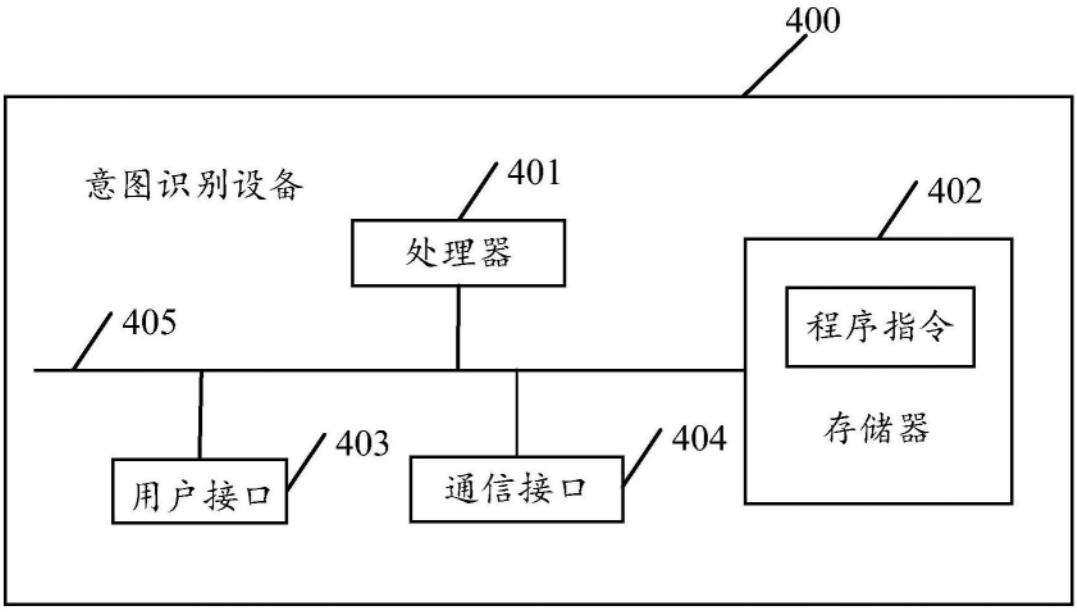

400
意图识别设备
401
处理器
402
程序指令
405
存储器
403
404
用户接口
通信接口

图4