



(12) 发明专利申请

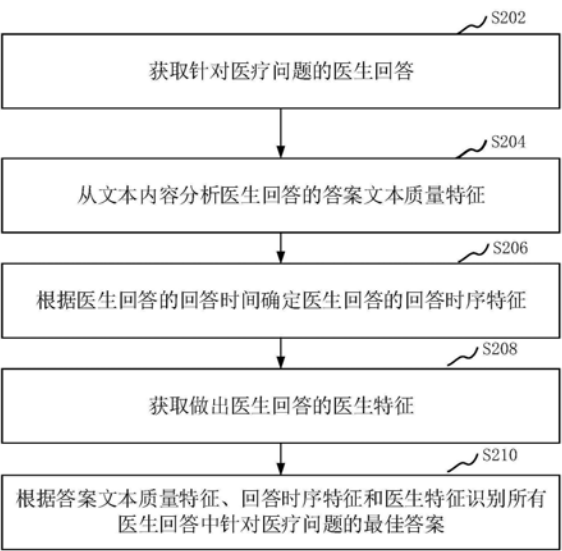
(10) 申请公布号 CN 111783473 A
(43) 申请公布日 2020. 10. 16

(21) 申请号 202010673856.0
(22) 申请日 2020.07.14
(71) 申请人 腾讯科技(深圳)有限公司
地址 518000 广东省深圳市南山区高新区
科技中一路腾讯大厦35层
(72) 发明人 康战辉
(74) 专利代理机构 广州华进联合专利商标代理
有限公司 44224
代理人 董慧 聂榕
(51) Int.Cl.
G06F 40/30 (2020.01)
G06Q 10/06 (2012.01)
G06N 3/04 (2006.01)
G06N 3/08 (2006.01)

权利要求书2页 说明书10页 附图5页

(54) 发明名称
医疗问答中最佳答案的识别方法、装置和计
算机设备

(57) 摘要
本申请涉及一种基于语义理解的医疗问答
中最佳答案的识别方法、装置、计算机设备和存
储介质。该方法包括：获取针对医疗问题的医生
回答；从文本内容分析所述医生回答的答案文本
质量特征；根据医生回答的回答时间确定所述医
生回答的回答时序特征；获取做出所述医生回答
的医生特征；根据所述答案文本质量特征、回答
时序特征和医生特征识别所有医生回答中针对
所述医疗问题的最佳答案。该方法综合考虑了答
案内容质量因素，医生因素以及回复时序因素确
定最佳答案，由于特征维度多，因而能够避免单
一特征维度识别的缺陷，大大提高了最佳答案识
别的准确性。



1. 一种医疗问答中最佳答案的识别方法,其特征在于,所述方法包括:
获取针对医疗问题的医生回答;
从文本内容分析所述医生回答的答案文本质量特征;
根据医生回答的回答时间确定所述医生回答的回答时序特征;
获取做出所述医生回答的医生特征;
根据所述答案文本质量特征、回答时序特征和医生特征识别所有医生回答中针对所述医疗问题的最佳答案。
2. 根据权利要求1所述的方法,其特征在于,根据所述答案文本质量特征、回答时序特征和医生特征识别所有医生回答中针对所述医疗问题的最佳答案,包括:
将所述答案文本质量特征、回答时序特征和医生特征输入预先根据历史最佳答案训练的识别模型,确定所有医生回答中针对所述医疗问题的最佳答案。
3. 根据权利要求2所述的方法,其特征在于,根据历史最佳答案训练的识别模型的方式包括:
将医疗问答中的已解决问题及其最佳答案作为正样本;
将医疗问答中的已解决问题及其非最佳答案或其它问题的答案作为负样本;
获取所述正样本和负样本的所述答案文本质量特征、回答时序特征及医生特征;
根据所述正样本和负样本的答案文本质量特征、回答时序特征及医生特征训练识别模型。
4. 根据权利要求1所述的方法,其特征在于,从文本内容分析所述医生回答的答案文本质量特征,包括:
获取所述医生回答的答案长度;
查找所述医生答案中非重复词个数;
基于语义从文本内容分析所述医生答案和问题的相似度特征;
对所述答案长度特征、答案非重复词个数以及答案和问题的相似度特征进行加权处理,得到所述医生回答的所述答案文本质量特征。
5. 根据权利要求1所述的方法,其特征在于,根据医生回答的回答时间确定所述医生回答的回答时序特征,包括:
确定所述医生回答的回答时间到当前时间的时长距离;
根据所述时长距离确定所有医生回答的回答时序特征,所述回答时序特征与所述时长距离成反比。
6. 根据权利要求1所述的方法,其特征在于,所述医生特征包括活跃度、贡献度或权威度中的至少一种;获取做出所述医生回答的医生特征,包括:
根据做出所述医生回答的医生在时间周期内回答医疗问题的平均次数,确定医生的活跃度;
根据在时间周期内所述医生所做出的回答被认可的数量以及医生全部回答数据,确定患者对医生的认同度;
根据医生资质信息确定医生资质的权威度。
7. 根据权利要求6所述的方法,其特征在于,根据医生资质信息确定医生资质的权威度,包括:

根据医生执业医院的等级信息和医生职称等级确定医生资质的权威度。

8. 一种医疗问答中最佳答案的识别装置, 其特征在于, 所述装置包括:

医生回答获取模块, 用于获取针对医疗问题的医生回答;

质量特征分析模块, 用于从文本内容分析所述医生回答的答案文本质量特征;

时序特征分析模块, 用于根据医生回答的回答时间确定所述医生回答的回答时序特征;

医生特征分析模块, 用于获取做出所述医生回答的医生特征;

识别模块, 用于根据所述答案文本质量特征、回答时序特征和医生特征识别所有医生回答中针对所述医疗问题的最佳答案。

9. 一种计算机设备, 包括存储器和处理器, 所述存储器存储有计算机程序, 其特征在于, 所述处理器执行所述计算机程序时实现权利要求1至7中任一项所述的方法的步骤。

10. 一种计算机可读存储介质, 存储有计算机程序, 其特征在于, 所述计算机程序被处理器执行时实现权利要求1至7中任一项所述的方法的步骤。

医疗问答中最佳答案的识别方法、装置和计算机设备

技术领域

[0001] 本申请涉及人工智能技术领域，特别是涉及一种医疗问答中最佳答案的识别方法、装置、计算机设备和存储介质。

背景技术

[0002] 随着互联网医疗的长足发展，在线医疗问答社区越来越受到患者的广泛欢迎。在线医疗问答社区是一个医患问答平台，这类问答社区由患者提出问题，若干个医生回答。患者或者其他用户可以点赞某个医生的答案。由于医疗领域具有高度的专业性，患者很难判断哪条答案的质量更高，此时如果平台方可以通过算法判定相关答案的质量，并将最高质量的答案标注为“最佳答案”并置顶，以减少其他用户在浏览该疾病答案时可以更高效看到最优质的答案，这样不论是对患者还是后续的阅读者来说将会大有裨益。

[0003] 传统的医疗问答的最佳答案是根据对答案的点赞数，尤其是患者自己点赞与否，再结合回答医生是否为“三甲”医院医生来判定。

[0004] 这种基于用户点赞行为+医生资质的最佳答案识别方法存在以下弊端：1) 在免费医患问答社区里，很多患者或者用户得到医生答案后可能并不会点赞，即此时的用户的行为数据为空；2) 单用医生资质来判定当前问题的答案的质量也又失偏颇，一个好医生产出的答案不一定每个都是高质量的。所以现有技术用于评价答案的所用特征可能存在的缺失，稀疏性以及有偏性，导致确定的最佳答案不准确。

发明内容

[0005] 基于此，有必要针对上述技术问题，提供一种能够识别的准确性的医疗问答中最佳答案的识别方法、装置、计算机设备和存储介质。

[0006] 一种医疗问答中最佳答案的识别方法，所述方法包括：

[0007] 获取针对医疗问题的医生回答；

[0008] 从文本内容分析所述医生回答的答案文本质量特征；

[0009] 根据医生回答的回答时间确定所述医生回答的回答时序特征；

[0010] 获取做出所述医生回答的医生特征；

[0011] 根据所述答案文本质量特征、回答时序特征和医生特征识别所有医生回答中针对所述医疗问题的最佳答案。

[0012] 一种医疗问答中最佳答案的识别装置，所述装置包括：

[0013] 医生回答获取模块，用于获取针对医疗问题的医生回答；

[0014] 质量特征分析模块，用于从文本内容分析所述医生回答的答案文本质量特征；

[0015] 时序特征分析模块，用于根据医生回答的回答时间确定所述医生回答的回答时序特征；

[0016] 医生特征分析模块，用于获取做出所述医生回答的医生特征；

[0017] 识别模块，用于根据所述答案文本质量特征、回答时序特征和医生特征识别所有

医生回答中针对所述医疗问题的最佳答案。

[0018] 一种计算机设备,包括存储器和处理器,所述存储器存储有计算机程序,所述处理器执行所述计算机程序时实现以下步骤:

[0019] 获取针对医疗问题的医生回答;

[0020] 从文本内容分析所述医生回答的答案文本质量特征;

[0021] 根据医生回答的回答时间确定所述医生回答的回答时序特征;

[0022] 获取做出所述医生回答的医生特征;

[0023] 根据所述答案文本质量特征、回答时序特征和医生特征识别所有医生回答中针对所述医疗问题的最佳答案。

[0024] 一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现以下步骤:

[0025] 获取针对医疗问题的医生回答;

[0026] 从文本内容分析所述医生回答的答案文本质量特征;

[0027] 根据医生回答的回答时间确定所述医生回答的回答时序特征;

[0028] 获取做出所述医生回答的医生特征;

[0029] 根据所述答案文本质量特征、回答时序特征和医生特征识别所有医生回答中针对所述医疗问题的最佳答案。

[0030] 上述医疗问答中最佳答案的识别方法、装置、计算机设备和存储介质,通过从文本内容分析医生回答的答案文本质量特征、根据医生的回答时间确定回答时序特征,及获取确定做出医生回答的医生特征,从而综合答案内容质量因素,医生因素以及回复时序因素,即综合考虑了这三个维度确定最佳答案,由于特征维度多,因而能够避免单一特征维度识别的缺陷,大大提高了最佳答案识别的准确性。

附图说明

[0031] 图1为一个实施例中医疗问答中最佳答案的识别方法的应用环境图;

[0032] 图2为一个实施例中医疗问答中最佳答案的识别方法的流程示意图;

[0033] 图3为一个实施例中在线医疗问答社区的问题列表的示意图;

[0034] 图4为一个实施例中sigmoid函数对应的曲线示意图;

[0035] 图5为一个实施例中语义模型的结构示意图;

[0036] 图6为另一个实施例中语义模型的结构示意图;

[0037] 图7为一个实施例中医疗问答中最佳答案的识别装置的结构框图;

[0038] 图8为一个实施例中计算机设备的内部结构图。

具体实施方式

[0039] 为了使本申请的目的、技术方案及优点更加清楚明白,以下结合附图及实施例,对本申请进行进一步详细说明。应当理解,此处描述的具体实施例仅仅用以解释本申请,并不用于限定本申请。

[0040] 人工智能(Artificial Intelligence, AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理

论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个综合技术,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0041] 人工智能技术是一门综合学科,涉及领域广泛,既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0042] 自然语言处理(Nature Language processing,NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此,这一领域的研究将涉及自然语言,即人们日常使用的语言,所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

[0043] 随着人工智能技术研究和进步,人工智能技术在多个领域展开研究和应用,例如常见的智能家居、智能穿戴设备、虚拟助理、智能音箱、智能营销、无人驾驶、自动驾驶、无人机、机器人、智能医疗、智能客服等,相信随着技术的发展,人工智能技术将在更多的领域得到应用,并发挥越来越重要的价值。

[0044] 本申请实施例提供的方案涉及人工智能的医疗问答中最佳答案的识别等技术,具体通过如下实施例进行说明:

[0045] 本申请提供的医疗问答中最佳答案的识别方法,可以应用于如图1所示的应用环境中。其中,患者终端102通过网络与服务器104进行通信,医生终端106通过网络与服务器104进行通信。服务器获取针对医疗问题的医生回答;从文本内容分析所述医生回答的答案文本质量特征;根据医生回答的回答时间确定所述医生回答的回答时序特征;获取做出所述医生回答的医生特征;根据所述答案文本质量特征、回答时序特征和医生特征识别所有医生回答中针对所述医疗问题的最佳答案。其中,患者终端102和医生终端106可以但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备,服务器104可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

[0046] 在一个实施例中,如图2所示,提供了一种医疗问答中最佳答案的识别方法,以该方法应用于图1中的服务器为例进行说明,包括以下步骤:

[0047] 步骤202,获取针对医疗问题的医生回答。

[0048] 随着互联网医疗的发展,出现了在线医疗问答社区。医生提供资质证明通过认证入驻在线医疗问答平台,为患者提供医疗服务,如在线看诊。患者在在线医疗问答社区可向医生提问,如在线询问等。

[0049] 医疗问题是指患者通过在线医疗问答社区所提出的与医疗相关的问题。医生回答是医生在线医疗问答社区针对患者所提出的医疗问题所做出的解答。例如,在在线医疗问答社区,某个患者提了一个关于婴儿血管瘤的问题,这个问题即是医疗问题。相关经验的医生可在在线医疗问答社区就这个问题进行回复,医生做出的回复即是医生回答。可以理解的是,针对一个医疗问题,若有多个医生回复,则会有多个医生回答。本申请中,确定某个医疗问题的最佳答案的过程中,获取的针对这个医疗问题的全部医生回答。

[0050] 步骤204,从文本内容分析医生回答的答案文本质量特征。

[0051] 文本内容包括了医疗问题的文本内容以及医生回答的文本内容。答案文本质量特征,用于分析医疗问题的文本内容和医生回答的文本内容,表征从文本内容分析角度确定的医生回答的回答质量。其中,回答质量可以从多个维度分析确定。一个实施例中,分析维度可以包括但不限于答案长度、答案非重复词个数,以及答复和问题的相似度特征。一个高质量的答案文本质量应当具有的特征包括:答案长度适中,答复中非重复词的个数少,答案和问题的相似度高。

[0052] 步骤206,根据医生回答的回答时间确定医生回答的回答时序特征。

[0053] 回答时序特征用于表征医生回答的时序。对于一个患者的提问可能历史上先后有多个医生回答,交互中患者可能会跟医生有多次追问,澄清,直至将问题本身逐步描述清楚。因此往往最新回答问题的医生具有越多的问题信息,也就能够把答案回答的越准确。因此,医生回答的时序与医生回答的质量存在负关的关系。即,医生回答的时序越晚(按回答的先后排序),医生回答的质量越高的可能性越大。

[0054] 步骤208,获取做出医生回答的医生特征。

[0055] 医生特征是指做出医生回答的该医生所具有的特征,包括但不限于医生的活跃度、贡献度和权威度。

[0056] 其中,权威度与医生资质相关,医生资质是指医生从事医疗工作所具有的资质,医生资质可以从医生的从业时长、所执业的医院级别、医生职称等维度确定。高资质与做出高质量回答的可能性成正比,即资质越高的医生,做出高质量回答的可能性也越大。

[0057] 活跃度与医生回复问题数量相关。活跃度与做出高质量回答的可能性成正比,即在在线医疗问答社区的活跃度大的医生,做出高质量回答的可能性也越大。

[0058] 医生在在线医疗问答社区的贡献度是指医生在在线医疗问答社区所做出的贡献。在该医疗问答平台回答的问题数,收获的点赞数不同,这就导致了该医生在平台上的贡献度有所不同。因此,贡献度可以从医生在医疗问答平台回答的问题数,收获的点赞数等角度评估。高贡献度的医生与做出高质量回答的可能性成正比,即在在线医疗问答社区的贡献度大的医生,做出高质量回答的可能性也越大。

[0059] 其中,医生特征还可以包括医生链接。医生链接是医生介绍的链接,通常认为具有医生链接的医生受认可的程度高,做出高质量回答的可能性也越大。

[0060] 步骤210,根据答案文本质量特征、回答时序特征和医生特征识别所有医生回答中针对医疗问题的最佳答案。

[0061] 最佳答案是指针对一个医疗问题所有的医生回答中质量最高的医生回答。最佳答案从答案文本质量特征、回答时序特征和医生特征综合确定。其中,可以利用已训练的模型,将医生回答的答案文本质量特征、回答时序特征和医生特征输入已训练好的模型,输出医疗问题的最佳答案。

[0062] 上述的医疗问答中最佳答案的识别方法,通过从文本内容分析医生回答的答案文本质量特征、根据医生的回答时间确定回答时序特征,及获取确定做出医生回答的医生特征,从而综合答案内容质量因素,医生因素以及回复时序因素,即综合考虑了这三个维度确定最佳答案,由于特征维度多,因而能够避免单一特征维度识别的缺陷,大大提高了最佳答案识别的准确性。

[0063] 在另一个实施例中,根据答案文本质量特征、回答时序特征和医生特征识别所有医生回答中针对医疗问题的最佳答案,包括:将答案文本质量特征、回答时序特征和医生特征输入预先根据历史最佳答案训练的识别模型,确定所有医生回答中针对医疗问题的最佳答案。

[0064] 具体地,识别模型是利用历史最佳答案训练得到的。训练样本为在线医疗问答社区已解决问题的最佳答案。识别模型可以为分类模型,如二分类模型,输出医生回答为医疗问题的最佳答案的概率,进而设定阈值,对于预测为正例概率>K(此处K取0.8)的答案视为最终该问题的最佳答案。

[0065] 其中,识别模型可以为逻辑回归模型,也可以为基于神经网络的分类模型,如LSTM等。

[0066] 其中,根据历史最佳答案训练的识别模型的方式包括:将医疗问答中的已解决问题及其最佳答案作为正样本;将医疗问答中的已解决问题及其非最佳答案或其它问题的答案作为负样本;获取正样本和负样本的答案文本质量特征、回答时序特征及医生特征;根据正样本和负样本的答案文本质量特征、回答时序特征及医生特征训练识别模型。

[0067] 具体地,使用爬虫从在线医疗问答社区全部问题里的“已解决问题列表”,认为已解决问题列表中的每个问题的第一条答案即为“最佳答案”,将其作为正样本。其中,在在线医疗问答社区具有问题列表,如图3所示。问题列表中包括已解决问题列表。利用爬虫从已解决问题列表中的各问题及其最佳答案。

[0068] 负样本有两种一种,一种将医疗问答中已解决的问题和该问题的非最佳答案作为负样本,一种将已解决问题和其它问题的答案作为负样本。

[0069] 正样本和负样本的答案文本质量特征、回答时序特征和医生特征的获取方式与预测过程相同,此处不再赘述。

[0070] 其中,训练模型可LR(逻辑回归)做二分类模型。逻辑回归(Logistic Regression, LR)是传统机器学习中的一种分类模型,由于算法的简单和高效,在实际中应用非常广泛。其源于将线性回归里的线性函数改为sigmoid函数,如下:

$$[0071] \quad g(x) = \frac{1}{1 + e^{-x}}$$

[0072] 该函数对应的曲线如图4所示。从图4可以看到sigmoid函数是一个s形的曲线,它的取值在[0,1]之间,在0点取值为0.5,在远离0的地方函数的值会很快接近0或是1。这个性质使能够以概率的方式来解释分类的结果。其对应的条件概率分布目标函数(即二分类目标函数)为:

$$[0073] \quad P(y = 1|x; w) = g(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

$$[0074] \quad P(y = 0|x; w) = g(w^T x) = \frac{e^{-w^T x}}{1 + e^{-w^T x}}$$

[0075] 其中xi是输入的特征维度,包括答案文本质量特征、回答时序特征和医生特征,xi=[x1i, x2i, ..., xmi], y在{0,1}中取值。(本文用{1,0}表示正例和负例)

[0076] 除了比较有可解释性的LR算法外,也可以将上述人工特征结合文本词词向量模型

(word2vec)表示打包在一起使用blstm+softmax等深度学习模型直接做分类模型训练。

[0077] 在另一个实施例中,从文本分析医生回答的答案文本质量特征,包括:获取医生回答的答案长度;查找医生答案中非重复词个数;基于语义从文本内容分析答案和问题的相似度特征;对答案长度特征、答案非重复词个数以及答案和问题的相似度特征进行加权处理,得到医生回答的答案文本质量特征。

[0078] 具体地,答案长度是指答案的字符长度,一般好的医生答案的长度适中,不会特别简短,更不会非常冗长。可通过统计字符数量,确定医生回答的答案长度。

[0079] 医生答案中非重复词个数是指医生答案中出现的出重词的个数。好的答案应该有足够的有效的信息,行文反反复复的用词应该尽可能少。医生答案中非重复词的个数可通过可首先对医生答案进行分词处理,进而遍历每个词,并查找每个词出现的次数。对于出现次数为1的词确定为非重复词,并统计非重复词的个数。

[0080] 医生答案和问题的相似度特征是指二者在语义维度的相似度,相似度越高,说明该签字越是针对该问题的,即二者的匹配度高。

[0081] 本实施例中从答案长度、非重复词个数以及答案和问题的相似度特征这三个维度表征医生回答的答案文本质量特征。在实际应用中,答案长度特征、答案非重复词个数以及答案和问题的相似度特征之和,也可以根据实际业务场景,为三个特征维度分别设置权重,例如可设置答案和问题的相似度特征具有较高的权重0.5,答案长度的权重为0.3,非重复词个数的权重为0.3,对答案长度特征、答案非重复词个数以及答案和问题的相似度特征进行加权处理,得到医生回答的答案文本质量特征。

[0082] 具体地,基于语义分析医生答案和问题的相似度特征,可采用语义模型进行分析。例如,事先用全医患社区网页对标准google bert做微调(fine-tune),然后对患者和医生答案基于调优后的bert做医疗问题和医生回答的(QA)语义相似度计算,向量相似度越高,说明该答案越是文答对题的。BERT的全称是Bidirectional Encoder Representation from Transformers,即双向Transformer的Encoder,因为decoder是不能获要预测的信息的。模型的主要创新点都在前训练(pre-train)方法上,即用了掩码语言模式(Masked LM)和下一句预测(Next Sentence Prediction)两种方法分别捕捉词语和句子级别的表征(representation)。模型结构如图5所示。

[0083] 微调(fine-tune)方式是指在已经训练好的语言模型的基础上,加入少量的特定任务限制因素(task-specific parameters),例如对于分类问题在语言模型基础上加一层softmax网络,然后在新的语料上重新训练来进行微调(fine-tune),其模型结构如图6所示。

[0084] 首先语言模型采用了Transformer Decoder的方法来进行训练,采用文本预测作为语言模型训练任务,训练完毕之后,加一层线性工程(Linear Project)来完成分类/相似度计算等自然语言处理(NLP)任务。因此总结来说,LM+Fine-Tuning的方法工作包括两步:

[0085] 构造语言模型,采用大的语料A来训练语言模型,在语言模型基础上增加少量神经网络层来完成特定任务(specific task)例如序列标注、分类等,然后采用有标记的语料B来有监督地训练模型,这个过程中语言模型的参数并不固定,依然是可训练变量(trainable variables)。

[0086] 在另一个实施例中,根据医生回答的回答时间确定医生回答的回答时序特征,包

括:确定医生回答的回答时间到当前时间的时长距离;根据时长距离确定所有医生回答的回答时序特征,回答时序特征与时长距离成反比。

[0087] 具体地,对于一个患者的提问可能历史上先后有多个医生回答,交互中患者可能会跟医生有多次追问,澄清,直至将问题本身逐步描述清楚。因此往往最新回答问题的医生具有越多的问题信息,也就能够把答案回答的越准确,全面。故此处将历史上所有医生的回答时间按到当前时间的时长距离做从小到大排序,根据时长距离确定所有医生回答的回答时序特征,回答时序特征与时长距离成反比,时长距离的数值越小,表示距离当时时间越近,答案为最佳答案的可能性越高。因此,可将回答时序特征取时长距离 T 的倒数作为表示。

[0088] 在另一个实施例中,医生特征包括活跃度、贡献度或权威度中的至少一种;获取做出医生回答的医生特征,包括:根据做出医生回答的医生在时间周期内回答医疗问题的平均次数,确定医生的活跃度;根据在时间周期内医生所做出的回答被认可的数量以及医生全部回答数据,确定患者对医生的认同度;根据医生资质信息确定医生资质的权威度。

[0089] 具体地,假设一个医生在平台上回答问题很少,频率很低,那显然他是不活跃的,一个不活跃的医生往往很难说他在该平台上是权威医生。故此处我们定义医生 i 的活跃度 $Active_Score$ 计算方法为其特定时间段内(如一个月)回答医疗问题的平均次数,公式表达如下:

[0090] $Active_Score(i) = Answer_Count(i) / Period$

[0091] 此处 $Period$ 为时间周期常数,取30天。 $Answer_Count$ 为时间周期内的该医生的全部回答次数。

[0092] 如果医生对患者的问题回答让患者满意,患者会通过点赞来对该医生表示认同。所以此处医生的受认同度 $Sat_Score(i)$ 公式可以表示为:

[0093] $Sat_Score(i) = Star_Num / (Total_Question_Num + 1)$

[0094] 其中,认同可通过点赞行为来确定,其中 $Star_Num$ 表示该医生在该特定时间段内(也取30天)被点赞的问题回答数, $Total_Question_Num$ 则指的是总回答问题数。分母同时加1的原因是防止除零操作。

[0095] 医生资质的权威度,即传统上医生的学术素养分,包含了医生所在医院等级和职称信息综合计算。根据医生个人信息页里登记的所在医院和职称等级信息,结合互联网上事先公开可得的医院级别资格表数据可知,此处根据1~3级医院+甲,乙两等,共计6个等级,分别从高到底记医院等级分 $Hospital_Rank$ 为6~1分。另外职称至少分为助理医师,医师,副主任医师,主任医师等级别,分别记为职称等级分 Pro_Rank 为2,4,7,8。即权威度 $Specialty_Score(i)$ 则为:

[0096] $Specialty_Score(i) = Hospital_Rank * Pro_Rank$

[0097] 本实施例中,根据活跃度、贡献度和权威度确定医生特征,医生特征 $Doctor_Score$ 为:

[0098] $Doctor_Score = Active_Score(i) * Sat_Score(i) * \log(Specialty_Score(i))$

[0099] 本申请的医疗问答中最佳答案的识别方法,结合答案文本质量特征、回答时序特征和医生特征,利用分类模型来机器自动识别最佳答案。综合考虑了这三个维度确定最佳答案,由于特征维度多,因而能够避免单一特征维度识别的缺陷,大大提高了最佳答案识别的准确性。利用基于人工智能的分类模型,提高了最佳答案识别的智能度和效率。

[0100] 应该理解的是,虽然图2的流程图中的各个步骤按照箭头的指示依次显示,但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,这些步骤可以以其它的顺序执行。而且,图2中的至少一部分步骤可以包括多个步骤或者多个阶段,这些步骤或者阶段并不必然是在同一时刻执行完成,而是可以在不同的时刻执行,这些步骤或者阶段的执行顺序也不必然是依次进行,而是可以与其它步骤或者其它步骤中的步骤或者阶段的至少一部分轮流或者交替地执行。

[0101] 在一个实施例中,如图7所示,提供了一种医疗问答中最佳答案的识别装置,该装置可以采用软件模块或硬件模块,或者是二者的结合成为计算机设备的一部分,该装置具体包括:医生回答获取模块、质量特征分析模块、时序特征分析模块、医生特征分析模块和识别模块,其中:

[0102] 医生回答获取模块701,用于获取针对医疗问题的医生回答。

[0103] 质量特征分析模块702,用于从文本内容分析所述医生回答的答案文本质量特征。

[0104] 时序特征分析模块703,用于根据医生回答的回答时间确定所述医生回答的回答时序特征。

[0105] 医生特征分析模块704,用于获取做出所述医生回答的医生特征。

[0106] 识别模块705,用于根据所述答案文本质量特征、回答时序特征和医生特征识别所有医生回答中针对所述医疗问题的最佳答案。

[0107] 上述医疗问答中最佳答案的识别装置,通过从文本内容分析医生回答的答案文本质量特征、根据医生的回答时间确定回答时序特征,及获取确定做出医生回答的医生特征,从而综合答案内容质量因素,医生因素以及回复时序因素,即综合考虑了这三个维度确定最佳答案,由于特征维度多,因而能够避免单一特征维度识别的缺陷,大大提高了最佳答案识别的准确性。

[0108] 在另一个实施例中,识别模块,用于将所述答案文本质量特征、回答时序特征和医生特征输入预先根据历史最佳答案训练的识别模型,确定所有医生回答中针对所述医疗问题的最佳答案。

[0109] 在另一个实施例中,医疗问答中最佳答案的识别装置,还包括:

[0110] 正样本获取模块,用于将医疗问答中的已解决问题及其最佳答案作为正样本。

[0111] 负样本获取模块,用于将医疗问答中的已解决问题及其非最佳答案或其它问题的答案作为负样本。

[0112] 特征获取模块,用于获取所述正样本和负样本的答案文本质量特征、回答时序特征和医生特征。

[0113] 训练模块,用于根据所述正样本和负样本的答案文本质量特征、回答时序特征和医生特征训练识别模型。

[0114] 在另一个实施例中,质量特征分析模块702,用于获取所述医生回答的答案长度;查找所述医生答案中非重复词个数;基于语义从文本内容分析所述医生答案和问题的相似度特征;对所述答案长度特征、答案非重复词个数以及答案和问题的相似度特征进行加权处理,得到所述医生回答的所述答案文本质量特征。

[0115] 在另一个实施例中,时序特征分析模块703,用于确定所述医生回答的回答时间到当前时间的时长距离;根据所述时长距离确定所有医生回答的回答时序特征,所述回答时

序特征与所述时长距离成反比。

[0116] 在另一个实施例中,所述医生特征包括活跃度、贡献度或权威度中的至少一种;医生特征分析模块,包括:

[0117] 活跃度分析模块,用于根据做出所述医生回答的医生在时间周期内回答医疗问题的平均次数,确定医生的活跃度;

[0118] 认同度分析模块,用于根据在时间周期内所述医生所做出的回答被认可的数量以及医生全部回答数据,确定患者对医生的认同度;

[0119] 权威度分析模块,用于根据医生资质信息确定医生资质的权威度。

[0120] 其中,权威度分析模块,用于根据医生执业医院的等级信息和医生职称等级确定医生资质的权威度。

[0121] 关于医疗问答中最佳答案的识别装置的具体限定可以参见上文中对于医疗问答中最佳答案的识别方法的限定,在此不再赘述。上述医疗问答中最佳答案的识别装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中,也可以以软件形式存储于计算机设备中的存储器中,以便于处理器调用执行以上各个模块对应的操作。

[0122] 在一个实施例中,提供了一种计算机设备,该计算机设备可以是服务器,其内部结构图可以如图8所示。该计算机设备包括通过系统总线连接的处理器、存储器和网络接口。其中,该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于存储问题及医生答案数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种医疗问答中最佳答案的识别方法。

[0123] 本领域技术人员可以理解,图8中示出的结构,仅仅是与本申请方案相关的部分结构的框图,并不构成对本申请方案所应用于其上的计算机设备的限定,具体的计算机设备可以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

[0124] 在一个实施例中,还提供了一种计算机设备,包括存储器和处理器,存储器中存储有计算机程序,该处理器执行计算机程序时实现上述各方法实施例中的步骤。

[0125] 在一个实施例中,提供了一种计算机可读存储介质,存储有计算机程序,该计算机程序被处理器执行时实现上述各方法实施例中的步骤。

[0126] 在一个实施例中,提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得该计算机设备执行上述各方法实施例中的步骤。

[0127] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的计算机程序可存储于一非易失性计算机可读存储介质中,该计算机程序在执行时,可包括如上述各方法的实施例的流程。其中,本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和易失性存储器中的至少一种。非易失性存储器可包括只读存储器(Read-

Only Memory,ROM)、磁带、软盘、闪存或光存储器等。易失性存储器可包括随机存取存储器(Random Access Memory,RAM)或外部高速缓冲存储器。作为说明而非局限,RAM可以是多种形式,比如静态随机存取存储器(Static Random Access Memory,SRAM)或动态随机存取存储器(Dynamic Random Access Memory,DRAM)等。

[0128] 以上实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0129] 以上所述实施例仅表达了本申请的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对发明专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本申请构思的前提下,还可以做出若干变形和改进,这些都属于本申请的保护范围。因此,本申请专利的保护范围应以所附权利要求为准。

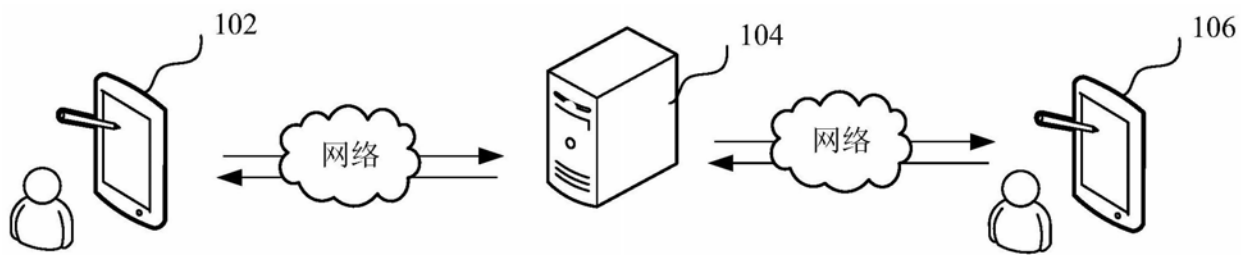


图1

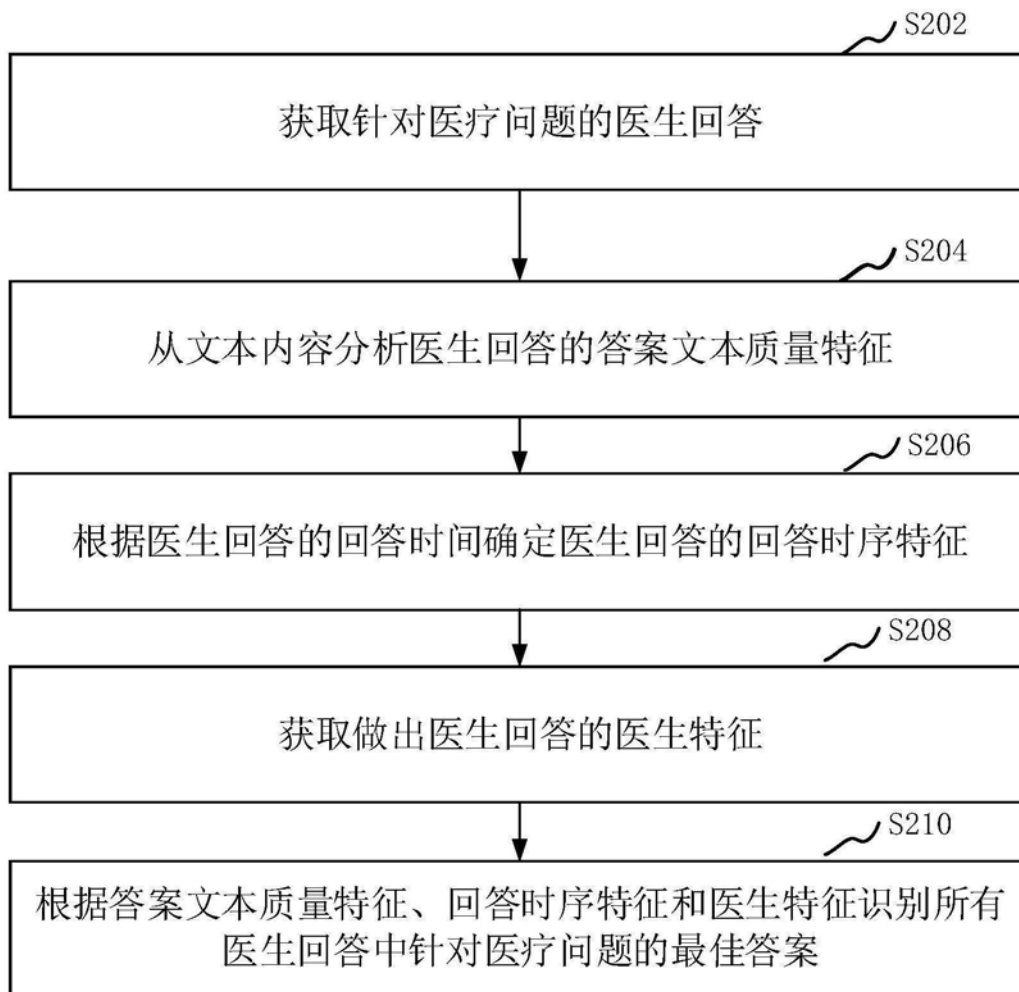


图2

全部问题 已解决问题 悬赏问题 热门问题 求助问题 更新问题				
提问标题	回答数	状态	提问	
[骨科] 肌肉拉伤用红花油好还是云南白药	1	✓ 已解决	2	
[外科其它] 手腕扭伤用红花油还是云南白药好	1	✓ 已解决	2	
[男科] 肾宝片饭前吃还是饭后吃	1	✓ 已解决	2	
[乳腺外科] 乳腺4a良性还是恶性多	1	✓ 已解决	2	
[脑积水] 脑积水引起脑萎缩还是脑萎缩引起脑积水?	1	✓ 已解决	2	
[骨科] 脚后跟长骨刺用白醋泡还是陈醋泡	1	✓ 已解决	2	
[骨科] 消炎药饭前吃还是饭后吃	1	✓ 已解决	2	
[普外科] 乳腺脓肿让脓液自动破好还是切开	1	✓ 已解决	2	
[泌尿外科] 男性尿路感染吃头孢还是左氧氟沙星男	1	✓ 已解决	2	
[骨科] 尾椎骨骨折后多走动好还是尽量不要动女	1	✓ 已解决	2	

图3

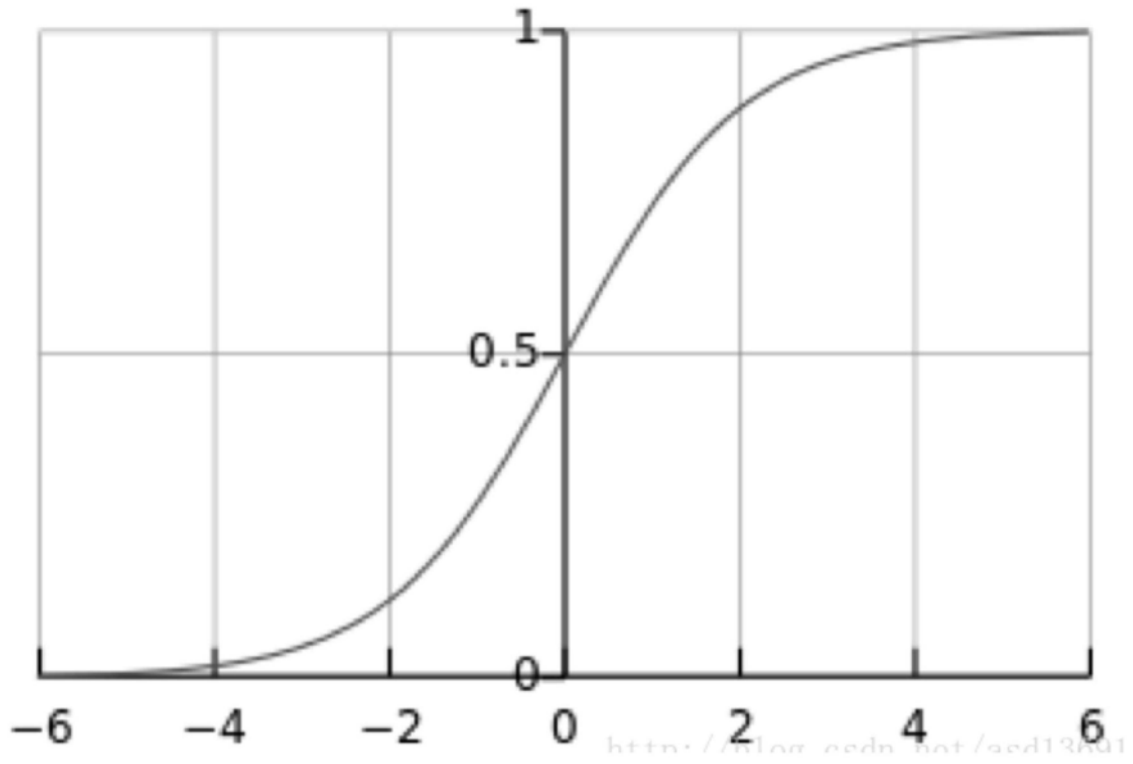


图4

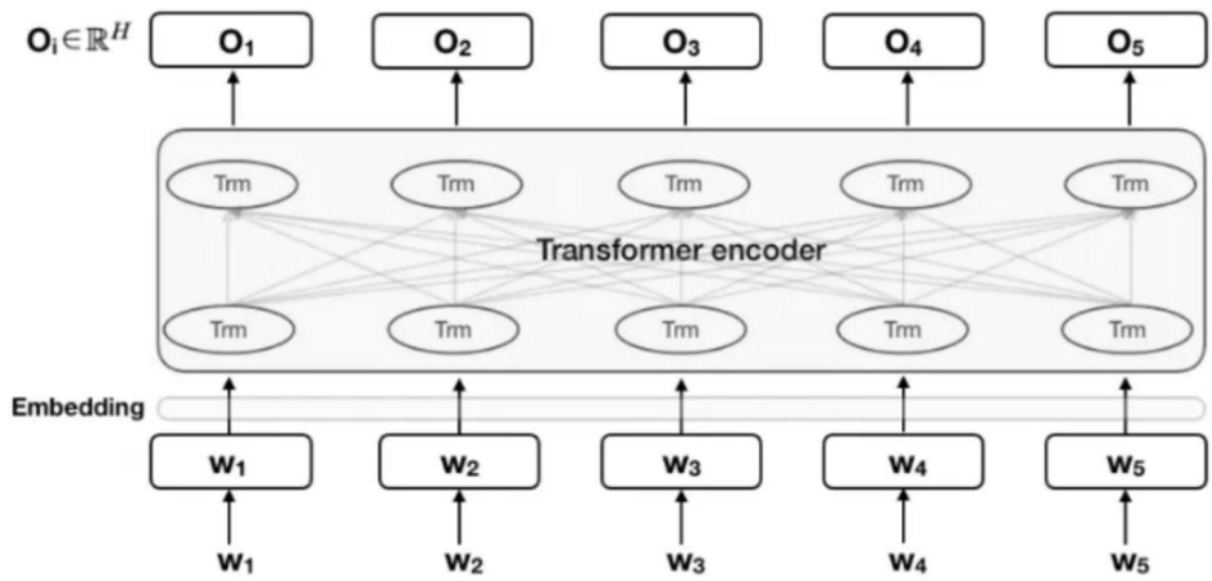


图5

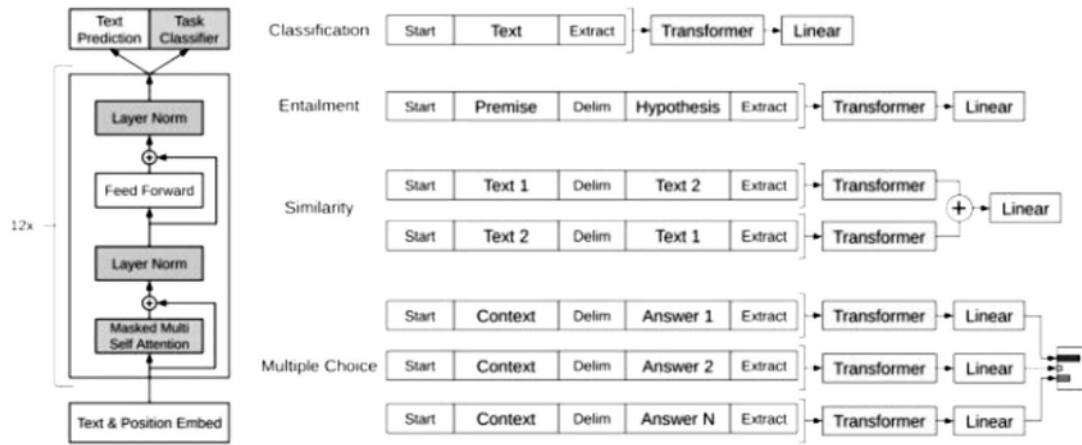


图6

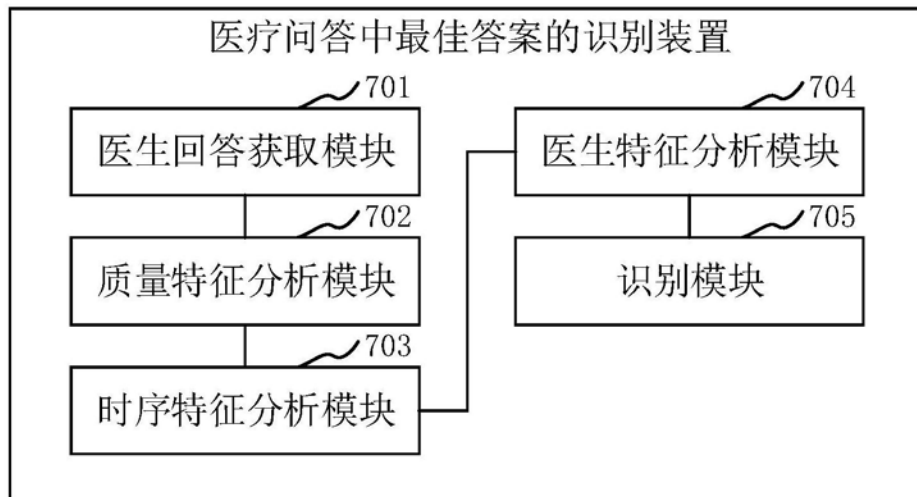


图7

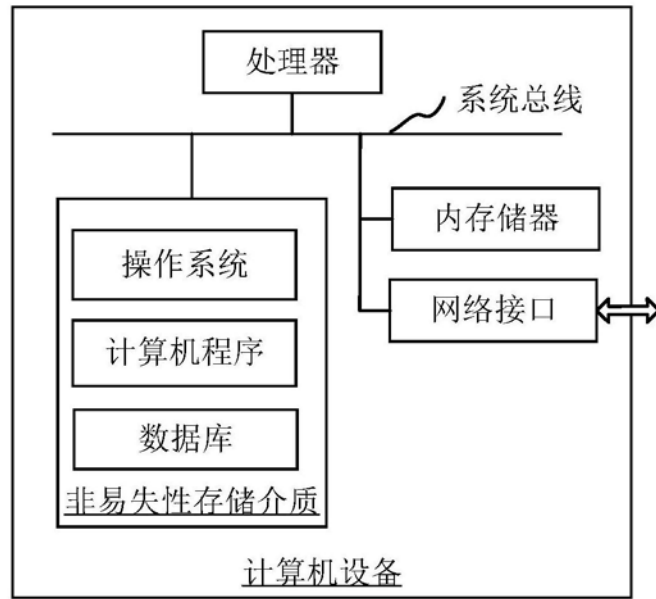


图8