(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111079854 B
(45)授权公告日 2024.04.23

(21)申请号 201911381956.X
(22)申请日 2019.12.27
(65)同一申请的已公布的文献号
申请公布号 CN 111079854 A
(43)申请公布日 2020.04.28
(73)专利权人 联想(北京)有限公司
地址 100085 北京市海淀区上地西路6号2幢2层201-H2-6
(72)发明人 杨沛 杨帆 葛羽辰 张成松
(74)专利代理机构 北京派特恩知识产权代理有限公司 11270
专利代理师 徐升升 张颖玲
(51)Int.Cl.
G06F 18/2415(2023.01)
G06N 3/0455(2023.01)
G06N 3/044(2023.01)
G06N 3/0464(2023.01)

(56)对比文件
CN 108304387 A,2018.07.20
CN 109388795 A,2019.02.26
CN 110147551 A,2019.08.20
CN 110210024 A,2019.09.06
CN 110276075 A,2019.09.24
CN 110298019 A,2019.10.01
US 2019087490 A1,2019.03.21

审查员 宋娜娜

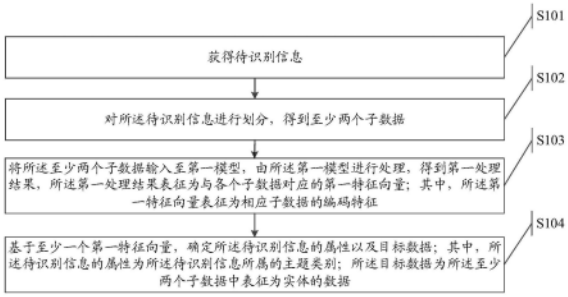
权利要求书2页 说明书11页 附图6页

(54)发明名称

信息识别方法、设备及存储介质

(57)摘要

本申请实施例公开了一种信息识别方法、设备及存储介质,其中,所述方法包括:获得待识别信息;对所述待识别信息进行划分,得到至少两个子数据;将所述至少两个子数据输入至第一模型,由所述第一模型进行处理得到第一处理结果,所述第一处理结果表征为与各个子数据对应的第一特征向量;其中,所述第一特征向量表征为相应子数据的编码特征;基于至少一个第一特征向量,确定所述待识别信息的属性以及目标数据;其中,所述待识别信息的属性为所述待识别信息所属的主题类别;所述目标数据为所述至少两个子数据中表征为实体的数据。



CN 111079854 B

1.一种信息识别方法,所述方法包括:

获得待识别信息;

对所述待识别信息进行划分,得到至少两个子数据;

将所述至少两个子数据输入至第一模型,由所述第一模型进行处理,得到第一处理结果,所述第一处理结果表征为与各个子数据对应的第一特征向量;其中,所述第一特征向量表征为相应子数据的编码特征;

基于至少一个第一特征向量,确定所述待识别信息的属性以及目标数据;其中,所述待识别信息的属性为所述待识别信息所属的主题类别;所述目标数据为所述至少两个子数据中表征为实体的数据;

所述基于至少一个第一特征向量,确定所述目标数据,包括:

获得第二特征向量,所述第二特征向量表征为所述待识别信息的解码特征;所述解码特征在不同时刻的内容不同;

依据所述第二特征向量和所述至少一个第一特征向量,对各个子数据进行编码;

对编码后的各个子数据进行解码,得到所述目标数据。

2.根据权利要求1所述的方法,其特征在于,所述基于所述第一特征向量中的至少一个第一特征向量,确定所述待识别信息的属性,包括:

将所述至少一个第一特征向量输入至第二模型;

由所述第二模型基于所输入的第一特征向量,对所述待识别信息属于各个预定主题类别的概率进行计算;

依据计算出的概率,确定所述待识别信息所属的主题类别。

3.根据权利要求1所述的方法,其特征在于,所述依据所述第二特征向量和所述至少一个第一特征向量,对各个子数据进行编码,包括:

将所述各个第一特征向量与所述第二特征向量分别进行相乘再相加运算,得到运算结果;

将所述运算结果和所述各个第一特征向量进行相乘运算,得到待识别信息的编码数据。

4.根据权利要求1或3所述的方法,其特征在于,所述对编码后的各个子数据进行解码,得到所述目标数据,包括:

将编码后的待识别信息输入至第三模型,得到各个子数据的特征信息;

根据各个子数据的特征信息,计算各个子数据表征为实体数据的概率;

根据各子数据表征为实体数据的概率,确定各个子数据中表征为实体的数据。

5.根据权利要求3所述的方法,其特征在于,在得到运算结果之后,所述方法还包括:

将所述运算结果进行归一化操作;

所述将所述运算结果和所述各个第一特征向量进行相乘运算,得到各个子数据的编码数据,包括:

将归一化的所述运算结果与所述各个第一特征向量进行相乘运算,得到所述编码数据。

6.根据权利要求1或2所述的方法,其特征在于,在确定出目标数据的情况下,

对目标数据进行实体类别的划分,确定目标数据所属的实体类别。

7.一种信息识别设备,所述设备包括:

获得单元,用于获得待识别信息;

划分单元,用于对所述待识别信息进行划分,得到至少两个子数据;

处理单元,用于对所述至少两个子数据进行处理,得到第一处理结果,所述第一处理结果表征为与各个子数据对应的第一特征向量;其中,所述第一特征向量表征为相应子数据的编码特征;

确定单元,用于基于至少一个第一特征向量,确定所述待识别信息的属性以及目标数据;其中,所述待识别信息的属性为所述待识别信息所属的主题类别;所述目标数据为所述至少两个子数据中表征为实体的数据;

确定单元,还用于获得第二特征向量,所述第二特征向量表征为所述待识别信息的解码特征;所述解码特征在不同时刻的内容不同;依据所述第二特征向量和所述至少一个第一特征向量,对各个子数据进行编码;对编码后的各个子数据进行解码,得到所述目标数据。

8.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现权利要求1至6任一所述方法的步骤。

9.一种信息识别设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,其特征在于,所述处理器执行所述程序时实现权利要求1至6任一所述方法的步骤。

信息识别方法、设备及存储介质

技术领域

[0001] 本申请涉及识别技术,具体涉及一种信息识别方法、设备及存储介质。

背景技术

[0002] 相关技术中可对一段文本数据所涉及的主题类别进行识别,如识别为该段文本数据属于科技类文章、体育类文章、或情感类文章。由于我国自然语言通常带有一定的语气和语调,相同的文本数据可能表达不同的含义,这就导致了识别准确性不足。除此之外,一段文本数据中通常会包括有诸如人名、地名、机构名等表示为实体的数据(如地名这一实体数据表示为城市、县等)。某个表示为实体的数据其所属的实体类别指的是该数据为人名、地名或机构名。在实际应用中,表示为实体的数据通常可为推荐或搜索提供了一定的帮助如对某城市的旅游路线的搜索。可见,亟需一种既能够准确识别出主题类别又能够识别出实体数据的方案。

发明内容

[0003] 为解决现有存在的技术问题,本申请实施例提供一种信息识别方法、设备及存储介质。

[0004] 本申请实施例的技术方案是这样实现的:

[0005] 本申请实施例提供一种信息识别方法,所述方法包括:

[0006] 获得待识别信息;

[0007] 对所述待识别信息进行划分,得到至少两个子数据;

[0008] 将所述至少两个子数据输入至第一模型,由所述第一模型进行处理,得到第一处理结果,所述第一处理结果表征为与各个子数据对应的第一特征向量;其中,所述第一特征向量表征为相应子数据的编码特征;

[0009] 基于至少一个第一特征向量,确定所述待识别信息的属性以及目标数据;其中,所述待识别信息的属性为所述待识别信息所属的主题类别;所述目标数据为所述至少两个子数据中表征为实体的数据。

[0010] 上述方案中,所述基于所述第一特征向量中的至少一个第一特征向量,确定所述待识别信息的属性,包括:

[0011] 将所述至少一个第一特征向量输入至第二模型;

[0012] 由所述第二模型基于所输入的第一特征向量,对所述待识别信息属于各个预定主题类别的概率进行计算;

[0013] 依据计算出的概率,确定所述待识别信息所属的主题类别。

[0014] 上述方案中,所述基于至少一个第一特征向量,确定目标数据,包括:

[0015] 获得第二特征向量,所述第二特征向量表征为所述待识别信息的解码特征;

[0016] 依据所述第二特征向量和所述至少一个第一特征向量,对待识别信息进行编码;

[0017] 对编码后的待识别信息进行解码,得到所述目标数据。

[0018] 上述方案中,所述依据所述第二特征向量和所述至少一个第一特征向量,对各个子数据进行编码,包括:

[0019] 将所述各个第一特征向量与所述第二特征向量分别进行相乘再相加运算,得到运算结果;

[0020] 将所述运算结果和所述各个第一特征向量进行相乘运算,得到待识别信息的编码数据。

[0021] 上述方案中,所述对编码后的各个子数据进行解码,得到所述目标数据,包括:

[0022] 将编码后的待识别信息输入至第三模型,得到各个子数据的特征信息;

[0023] 根据各个子数据的特征信息,计算各个子数据表征为实体数据的概率;

[0024] 根据各子数据表征为实体数据的概率,确定各个子数据中表征为实体的数据。

[0025] 上述方案中,在得到运算结果之后,所述方法还包括:

[0026] 将所述运算结果进行归一化操作;

[0027] 所述将所述运算结果和所述各个第一特征向量进行相乘运算,得到各个子数据的编码数据,包括:

[0028] 将归一化的所述运算结果与所述各个第一特征向量进行相乘运算,得到所述编码数据。

[0029] 上述方案中,在确定出目标数据的情况下,

[0030] 对目标数据进行实体类别的划分,确定目标数据所属的实体类别。

[0031] 本申请实施例提供一种信息识别设备,所述设备包括:

[0032] 获得单元,用于获得待识别信息;

[0033] 划分单元,用于对所述待识别信息进行划分,得到至少两个子数据;

[0034] 处理单元,用于对所述至少两个子数据进行处理,得到第一处理结果,所述第一处理结果表征为与各个子数据对应的第一特征向量;其中,所述第一特征向量表征为相应子数据的编码特征;

[0035] 确定单元,用于基于至少一个第一特征向量,确定所述待识别信息的属性以及目标数据;其中,所述待识别信息的属性为所述待识别信息所属的主题类别;所述目标数据为所述至少两个子数据中表征为实体的数据。

[0036] 本申请实施例提供一种计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现前述方法的步骤。

[0037] 本申请实施例提供一种信息识别设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,所述处理器执行所述程序时实现前述方法的步骤。

[0038] 本申请实施例提供一种信息识别方法、设备及存储介质,其中,所述方法包括:获得待识别信息;对所述待识别信息进行划分,得到至少两个子数据;将所述至少两个子数据输入至第一模型,由所述第一模型进行处理得到第一处理结果,所述第一处理结果表征为与各个子数据对应的第一特征向量;其中,所述第一特征向量表征为相应子数据的编码特征;基于至少一个第一特征向量,确定所述待识别信息的属性以及目标数据;其中,所述待识别信息的属性为所述待识别信息所属的主题类别;所述目标数据为所述至少两个子数据中表征为实体的数据。

[0039] 本申请实施例中,既能够识别出主体类别又能够识别出实体数据,与相关技术相

比,同时进行主体类别和实体数据的识别,省时又省力。而且利用第一模型进行识别,第一模型具有很强的鲁棒性,可大大提高识别准确度。

附图说明

[0040] 为了更清楚地说明本申请实施例或现有技术中的技术方案,下面将对实施例或现有技术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申请的实施例,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,还可以根据提供的附图获得其他的附图。

[0041] 图1为本申请提供信息识别方法的实施例的实现流程示意图一;

[0042] 图2为本申请提供信息识别方法的实施例的实现流程示意图二;

[0043] 图3为本申请提供信息识别方法的实施例的实现流程示意图三;

[0044] 图4为本申请提供信息识别方法的实施例的实现流程示意图四;

[0045] 图5为本申请提供信息识别方法的实施例的实现流程示意图五;

[0046] 图6为本申请提供的识别原理示意图;

[0047] 图7为本申请提供信息识别设备的组成结构示意图;

[0048] 图8为本申请提供信息识别设备的硬件构成示意图。

具体实施方式

[0049] 为使本申请的目的、技术方案和优点更加清楚明白,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清楚、完整地描述,显然,所描述的实施例仅仅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。在不冲突的情况下,本申请中的实施例及实施例中的特征可以相互任意组合。在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行。并且,虽然在流程图中示出了逻辑顺序,但是在某些情况下,可以以不同于此处的顺序执行所示出或描述的步骤。

[0050] 本申请提供信息识别方法的实施例,如图1所示,所述方法包括:

[0051] 步骤(S)101:获得待识别信息;

[0052] S102:对所述待识别信息进行划分,得到至少两个子数据;

[0053] 在S101～S102中,待识别信息可以是任何多媒体数据如文本数据、音频数据、视频数据等。优选为文本数据。在待识别信息为文本数据的情况下,将文本数据按照字、词等进行划分,得到至少两个子数据。在待识别数据为非文本数据如音频数据或视频数据的情况下,将非文本数据进行非文本数据到文本数据的转换,得到文本数据,再将文本数据进行划分。

[0054] S103:将所述至少两个子数据输入至第一模型,由所述第一模型进行处理,得到第一处理结果,所述第一处理结果表征为与各个子数据对应的第一特征向量;其中,所述第一特征向量表征为相应子数据的编码特征;

[0055] S104:基于至少一个第一特征向量,确定所述待识别信息的属性以及目标数据;其中,所述待识别信息的属性为所述待识别信息所属的主题类别;所述目标数据为所述至少

两个子数据中表征为实体的数据。

[0056] 前述方案中,将待识别信息划分的子数据输入至第一模型,通过第一模型对子数据的处理得到与各个子数据对应的(第一)特征向量,基于特征向量进行待识别信息所属的主题类别的识别以及识别其中的表征为实体的数据。可见,本申请实施例中既能够识别出主题类别又能够识别出实体数据,与相关技术相比,同时进行主体类别和实体数据的识别,省时又省力。而且利用第一模型进行识别,第一模型具有很强的鲁棒性,可大大提高识别准确度。

[0057] 在一个可选的实施例中,如图2所示,S104中所述基于所述第一特征向量中的至少一个第一特征向量,确定所述待识别信息的属性,包括:

[0058] S201:将所述至少一个第一特征向量输入至第二模型;

[0059] S202:由所述第二模型基于所输入的第一特征向量,对所述待识别信息属于各个预定主题类别的概率进行计算;

[0060] S203:依据计算出的概率,确定所述待识别信息所属的主题类别。

[0061] 前述方案为对待识别信息所属的主题类别进行识别的方案。通过第二模型对待识别信息属于各个预定主题类别的概率进行计算,依据计算出的概率来确定待识别信息所属的主题类别。其中,第二模型具有强稳定性,可保证识别准确率和准确度。

[0062] 在一个可选的实施例中,如图3所示,S104中所述基于至少一个第一特征向量,确定目标数据,包括:

[0063] S301:获得第二特征向量,所述第二特征向量表征为所述待识别信息的解码特征;

[0064] S302:依据所述第二特征向量和所述至少一个第一特征向量,对待识别信息进行编码;

[0065] S303:对编码后的待识别信息进行解码,得到所述目标数据。

[0066] 前述方案为对待识别信息中表征为实体数据进行识别的方案。通过两个特征向量:第一特征向量和第二特征向量对待识别信息的各个子数据进行编码,并对编码后的各个子数据进行解码,进而识别出待识别信息中表征为实体的数据。也即本方案中通过先编码再解码的方案得到待识别信息中表征为实体的数据,编码操作的出现可使得待识别信息中表征为实体的数据的特征更为突显,更利于对表征为实体的数据进行识别。

[0067] 在一个可选的实施例中,如图4所示,S302所述依据所述第二特征向量和所述至少一个第一特征向量,对各个子数据进行编码,包括:

[0068] S3021:将所述各个第一特征向量与所述第二特征向量分别进行相乘再相加运算,得到运算结果;

[0069] S3022:将所述运算结果和所述各个第一特征向量进行相乘运算,得到待识别信息的编码数据。

[0070] 在S3021和S3022中,编码过程是:先将两个特征向量分别相乘再相加,然后将分别相乘再相加的结果与各第一特征向量进行相乘,进而实现对待识别信息中的编码。其中,第一特征向量表征为编码特征,第二特征向量表征为解码特征,通过编解码特征的如上运算进行待识别信息的编码,可将待识别信息中表征为实体的数据的特征更为突显,如此便可方便实现对实体数据的识别,也可保证识别准确性。

[0071] 可选的,在S3021之后也即在得到运算结果之后,所述方法还包括:

[0072] 将所述运算结果进行归一化操作；

[0073] 相应的，所述S3022将所述运算结果和所述各个第一特征向量进行相乘运算，得到待识别信息的编码数据，包括：将归一化的所述运算结果与所述各个第一特征向量进行相乘运算，得到所述编码数据。

[0074] 此处，为保证数据运算的统一性，在S3021得到运算结果之后，先对运算结果进行归一化操作，以将运算结果统一到相同的空间中去如均为将各运算结果统一为小于1的小数或分数，如此便有利于编码的实现。

[0075] 在一个可选的实施例中，如图5所示，所述S303对编码后的各个子数据进行解码，得到所述目标数据，进一步可通过如下方式实现：

[0076] S3031：将编码后的待识别信息输入至第三模型，得到各个子数据的特征信息；

[0077] S3032：根据各个子数据的特征信息，计算各个子数据表征为实体数据的概率；

[0078] S3033：根据各子数据表征为实体数据的概率，确定各个子数据中表征为实体的数据。

[0079] 前述方案中，根据待识别信息划分的各个子数据的特征信息进行各个子数据为实体数据的概率，并通过概率确定各个子数据中表征为实体的数据。这种从子数据的特征角度出发进行实体数据的识别，可保证实体数据的识别准确性。

[0080] 本领域技术人员应该而知，在实际应用中，实体数据的类别有多种如表征为地名的实体数据、表征为人名的实体数据、表征为城市名的实体数据，本申请实施例中，在识别出待识别信息中的表征为实体的数据之后，还需要对该表征为实体的数据的类别进行进一步识别。在技术实现上，对目标数据进行实体类别的划分，确定目标数据所属的实体类别。具体的，可通过将识别出的待识别信息中的实体数据与设定的各表征为对应实体类别的数据库中进行匹配，如果识别出的待识别信息中的实体数据在哪个类别实体数据的数据库中出现。

[0081] 如果识别出的待识别信息中的实体数据出现在表征为城市名的实体数据的数据库中，则识别出的待识别信息中的实体数据为一城市名称。如果识别出的待识别信息中的实体数据出现在表征为人名的实体数据的数据库中，则识别出的待识别信息中的实体数据为人的名字。也即不仅实现对待识别信息中表征为实体的数据进行识别，还可进一步识别其属于哪种类别的实体，可在一定程度上满足实际应用需求。

[0082] 下面结合附图6对本申请实施例作进一步详细的说明。

[0083] 可以理解，本申请实施例的信息识别方法可应用于信息识别设备中，该设备可以是任何合理的设备、装置、系统等，如为服务器、虚拟机等。该信息识别设备可对用户日常的行为数据如阅读数据、收听音频、观看视频进行采集，并通过其采集的数据识别该用户所阅读的数据、收听音频、或观看的视频所属的主题类别、以及这些数据中表征为实体的数据。

[0084] 本应用场景中，用户使用移动终端如手机进行文章的阅读，信息识别设备采集用户阅读的数据，如该用户阅读一段文章(文本数据)、该文章存在有这样一句话：诸葛亮在荆州，将该句作为待识别信息，识别该句中的表征为实体的数据如“诸葛亮”(人名)和“荆州”(地名)，以及识别该用户阅读的该文章的类别如为小说类、科技类还是体育类。可以理解，如果信息识别设备采集到用户通过音频或视频的方式进行观看的数据，则将这些非文本数据转换为文本数据，将转换后的文本数据作为待识别信息。识别过程如下所述：

[0085] 如图6所示,本应用场景中信息识别设备包括编码端(Encoder)和解码端(Decoder),用于识别待识别信息-文本数据中的实体数据。也即文本数据中的实体数据通过编码端和解码端的配合完成识别。文本数据所属的主题类别的识别过程在编码端完成。

[0086] 在具体实现上,编码端至少包括第一模型T1和第二模型T2,解码端包括第三模型。其中,第一模型T1和第三模型可以为神经网络模型、深度网络学习模型,进一步的可以如循环神经网络(RNN,Recurrent Neural Network)模型和卷积神经网络(CNN)模型。本应用场景中以第一模型T1为RNN、第三模型为全连接神经网络为例,第二模型T2为分类器T2为例。

[0087] 先对待识别信息属于哪种主题类别进行说明。

[0088] 在具体实现上,对待识别信息“诸葛亮在荆州”按照词汇进行子数据的划分,得到三个子数据v1～v3,其中v1＝诸葛亮,v2＝在,v3＝荆州。三个子数据依次送入到第一模型T1-RNN网络中。本领域技术人员应该理解,RNN网络中包括有多个神经网络,这些神经网络用于计算RNN的输入数据中的特征向量。本应用场景中,在输入v1至RNN网络的情况下,RNN网络中的神经网络为v1计算出特征向量h1。在输入v1+v2至RNN网络的情况下,RNN网络中的神经网络为v1+v2计算出特征向量h2。在输入v1+v2+v3至RNN网络的情况下,RNN网络中的神经网络为v1+v2+v3计算出特征向量h3。特征向量能够表示对应于各输入的子数据的文本特征如文本表达的含义和/或感情色彩。在待识别信息的全部子数据输入至RNN网络的情况下,RNN网络中的神经网络为v1+v2+v3计算出特征向量h3,将特征向量h3作为识别待识别信息的主题需要的特征向量,将特征向量h3输入至分类器T2,该分类器T2对特征向量h3表示的待识别信息的文本特征进行分析,进一步的对待识别信息属于各个预定主题类别的概率进行计算。如对待识别信息属于小说类、科技类、体育类的概率进行计算,从计算出的几个概率中挑选出取值最大的概率,取得最大概率值时使用的类别即为待识别信息所属的主题类别。在本应用场景中,分类器T2计算待识别信息属于小说类的概率最大,则可以确认待识别信息-诸葛亮在荆州属于小说类。可以理解,由于特征向量h1～h3由编码端的RNN网络计算而得,所以将其视为编码特征即为前述的第一特征向量。由于RNN模型具有很好的健壮性和鲁棒性,不易受外界环境的干扰,可提高主题类别的识别准确性和准确率。且本应用场景中利用特征向量h3和分类器T2即实现对主题类别的识别,实现难度不高,在工程上易于实现,容易推广使用。

[0089] 接下来对待识别信息中的实体数据进行识别的方案进行说明。

[0090] 信息识别设备的解码由全连接神经网络实现。该网络包括N1网络和分类器N2。其中,N1网络用于提供解码特征,由于其在不同时刻提供的解码特征c_i不同,所以待识别信息中对实体数据的识别需要按照不同时刻进行各个子数据是否为实体数据的逐一识别。

[0091] 假定在第i＝1时刻对第1子数据即v1＝诸葛亮是否为实体数据进行识别。先来看第i＝1时刻对待识别信息进行编码的过程:在第i＝1时刻,N1网络输出初始化特征向量c_{i-1}＝c_0。在第1时刻下将编码端的RNN网络计算出的特征向量h1～h3分别与c_0相乘再相加,得到$\alpha_1^{i=0}=h1*c_0=f(h1,c_0)$、$\alpha_2^{i=0}=h2*c_0=f(h2,c_0)$和$\alpha_3^{i=0}=h3*c_0$。再对$\alpha_1^{i=0}$、$\alpha_2^{i=0}$、$\alpha_3^{i=0}$进行归一化处理,得到$\alpha_1^{i=0'}=\alpha_1^{i=0}\Big/\left\|\alpha_1^{i=0}+\alpha_2^{i=0}+\alpha_3^{i=0}\right\|$、

$\alpha_2^{i=0'}=\alpha_2^{i=0}\Big/\left\|\alpha_1^{i=0}+\alpha_2^{i=0}+\alpha_3^{i=0}\right\|$、$\alpha_3^{i=0'}=\alpha_3^{i=0}\Big/\left\|\alpha_1^{i=0}+\alpha_2^{i=0}+\alpha_3^{i=0}\right\|$。其中，|| ||表示模值。然后，将$\alpha_1^{i=0'}$和h1相乘、$\alpha_2^{i=0'}$和h2相乘、$\alpha_3^{i=0'}$和h3相乘，这三个相乘结果再相加得到$E_{i=1}=\alpha_1^{i=0'}*h1+\alpha_2^{i=0'}*h2+\alpha_3^{i=0'}*h3$。$E_{i=1}$即可视为在第1时刻被编码后的待识别信息，至此编码端对待识别信息的编码过程执行完毕。前述方案中，对$\alpha_1^{i=0}$、$\alpha_2^{i=0}$、$\alpha_3^{i=0}$的归一化处理使得数据统一到相同的空间，如此便方便后续的运算，有利于编码的实现。此外，编码操作的出现可使得待识别信息中表征为实体的数据的特征更为突显，更利于对表征为实体的数据进行识别。而且如上的编码操作可保证待识别信息从编码端到解码端的传输安全性。可以理解，由于特征向量c_0由解码端的网络计算而得，所以将其视为解码特征即为前述的第二特征向量。

[0092] 编码端传输编码后的待识别信息即$E_{i=1}$至解码端。解码端的N1网络从$E_{i=1}$中解析出待识别信息，并获得待识别信息中的第1子数据v1的特征信息如文本特征，该文本特征可以是v1属于实体类别的信息或不属于实体类别的特征，将该特征信息输入至分类器，分类器根据v1属于实体类别的信息或不属于实体类别的信息，对v1为实体数据的概率或不为实体数据的概率进行计算。如果经计算得出v1为实体数据的概率为0.8大于第一阈值如0.7，或者得出v1不为实体数据的概率为0.3小于第二阈值如0.2，则可以认为待识别信息中的第1子数据v1为实体数据(表征为实体的数据)。至此，由编码端和解码端的配合完成了对待识别信息中的第1子数据v1是否为实体数据的识别。该识别方法采用了具有很强鲁棒性和稳定性的第一模型至第三模型，可大大保证识别准确性。

[0093] 假定在第i＝2时刻对第2子数据即v2＝在是否为实体数据进行识别。先来看第i＝2时刻对待识别信息进行编码的过程：在第i＝2时刻，N1网络输出第2时刻需要使用的特征向量$c_i=c_1$。从前述第1时刻的处理过程可以看出c_1是由编码端的N1网络对$E_{i=1}$进行解析而得到的。在第2时刻下将编码端的RNN网络计算出的特征向量h1～h3分别与c_1相乘再相加，得到$\alpha_1^{i=1}=h1*c_1=f(h1,c_1)$、$\alpha_2^{i=1}=h2*c_1=f(h2,c_1)$和$\alpha_3^{i=1}=h3*c_1$。再对$\alpha_1^{i=1}$、$\alpha_2^{i=1}$、$\alpha_3^{i=1}$进行归一化处理，得到$\alpha_1^{i=1'}=\alpha_1^{i=1}\Big/\left\|\alpha_1^{i=1}+\alpha_2^{i=1}+\alpha_3^{i=1}\right\|$、$\alpha_2^{i=1'}=\alpha_2^{i=1}\Big/\left\|\alpha_1^{i=1}+\alpha_2^{i=1}+\alpha_3^{i=1}\right\|$、$\alpha_3^{i=1'}=\alpha_3^{i=1}\Big/\left\|\alpha_1^{i=1}+\alpha_2^{i=1}+\alpha_3^{i=1}\right\|$。其中，|| ||表示模值。然后，将$\alpha_1^{i=1'}$和h1相乘、$\alpha_2^{i=1'}$和h2相乘、$\alpha_3^{i=1'}$和h3相乘，这三个相乘结果再相加得到$E_{i=2}=\alpha_1^{i=1'}*h1+\alpha_2^{i=1'}*h2+\alpha_3^{i=1'}*h3$。$E_{i=2}$即可视为在第2时刻被编码后的待识别信息，至此编码端对待识别信息的编码过程执行完毕。前述方案中，对$\alpha_1^{i=1}$、$\alpha_2^{i=1}$、$\alpha_3^{i=1}$的归一化处理使得数据统一到相同的空间，如此便方便后续的运算，有利于编码的实现。此外，编码操作的出现可使得待识别信息中表征为实体的数据的特征更为突显，更利于对表征为实体的数据进行识别。而且如上的编码操作可保证待识别信息从编码端到解码端的传输安全性。可以理解，由于特征向量c_1由解码端的网络计算而得，所以将其视为解码特征即为前述的第二特征向量。

[0094] 编码端传输编码后的待识别信息即$E_{i=2}$至解码端。解码端的N1网络从$E_{i=2}$中解析出待识别信息，并获得待识别信息中的第2子数据v2的特征信息如文本特征，该文本特征可

以是v2属于实体类别的信息或不属于实体类别的特征,将该特征信息输入至分类器,分类器根据v2属于实体类别的信息或不属于实体类别的信息,对v2为实体数据的概率或不为实体数据的概率进行计算。如果经计算得出v2为实体数据的概率为0.78大于第一阈值如0.7,或者得出v2不为实体数据的概率为0.28小于第二阈值如0.2,则可以认为待识别信息中的第2子数据v2为实体数据(表征为实体的数据)。至此,由编码端和解码端的配合完成了对待识别信息中的第2子数据v2是否为实体数据的识别。该识别方法采用了具有很强鲁棒性和稳定性的第一模型至第三模型,可大大保证识别准确性。

[0095] 假定在第i=3时刻对第3子数据即v2=荆州是否为实体数据进行识别。可以理解,该识别过程使用的解码特征为c_2。从前述第2时刻的处理过程可以看出c_2是由编码端的N1网络对$E_{i=2}$进行解析而得到的。具体识别过程请参见前述的对第1和/或第2子数据是否为实体数据的识别方案进行说明,重复之处不再赘述。

[0096] 本应用场景中,经过如上的识别过程可以得知,在“诸葛亮在荆州”这句话中可识别出第1子数据和第3子数据均为实体数据。进一步的,各自为哪种实体类别的数据,还需要与预先设定的几种实体数据库进行匹配。本应用场景中第1子数据会出现在表征为人名的实体数据的数据库中,则识别出的第1子数据为人的名字且该名字为“诸葛亮”。第3子数据会出现在表征为地名的实体数据的数据库中,则识别出的第3子数据为地名且该地名为“荆州”。可以理解,表征为人名的实体数据的数据库中记载有任何合理的人名如著名人物的名字。表征为地名的实体数据的数据库中记载有任何合理的地名如县级市的名称、地级市名称和省会的名称等。由此可见,本应用场景中不仅实现对待识别信息中表征为实体的数据进行识别,还可进一步识别其属于哪种类别的实体,可在一定程度上满足实际应用需求。

[0097] 从前述方案可知,通过编码端和解码端实现了对待识别信息中的主题类别和实体数据的同时识别。编码端和解码端均利用了具有强健性和稳定性的模型,可保证识别准确性。且编码端提供的编码方案可更为突出待识别信息中的各子数据的特征,进而更有利于对各子数据是否为实体数据进行识别。在实际应用中,对主题类别和实体数据进行同时识别之后,可以针对不同用户喜欢爱的文章或视频进行针对性的推荐,一方面可保证推送的准确性和针对性;另一方面,对用户来说其可自动接收到自身喜欢观看的视频或文章,可大大提升用户的使用体验。

[0098] 可以理解,前述是以待识别信息为“诸葛亮在荆州”为例进行的说明,此外,任何文本数据或经转换后得到的文本数据均可采用如上方案进行某段话或某句话或某个文章所属的主题类别的识别和实体数据的识别。

[0099] 本申请实施例还提供一种信息识别设备,如图7所示,所述设备包括:获得单元701、划分单元702、处理单元703和确定单元704;其中,

[0100] 获得单元701,用于获得待识别信息;

[0101] 划分单元702,用于对所述待识别信息进行划分,得到至少两个子数据;

[0102] 处理单元703,用于对所述至少两个子数据进行处理,得到第一处理结果,所述第一处理结果表征为与各个子数据对应的第一特征向量;其中,所述第一特征向量表征为相应子数据的编码特征;

[0103] 确定单元704,用于基于至少一个第一特征向量,确定所述待识别信息的属性以及目标数据;其中,所述待识别信息的属性为所述待识别信息所属的主题类别;所述目标数据

为所述至少两个子数据中表征为实体的数据。

[0104] 在一个可选的实施例中,确定单元704,用于将所述至少一个第一特征向量输入至第二模型;由所述第二模型基于所输入的第一特征向量,对所述待识别信息属于各个预定主题类别的概率进行计算;依据计算出的概率,确定所述待识别信息所属的主题类别。

[0105] 在一个可选的实施例中,确定单元704,用于获得第二特征向量,所述第二特征向量表征为所述待识别信息的解码特征;依据所述第二特征向量和所述至少一个第一特征向量,对待识别信息进行编码;对编码后的待识别信息进行解码,得到所述目标数据。

[0106] 进一步的,确定单元704,还用于将所述各个第一特征向量与所述第二特征向量分别进行相乘再相加运算,得到运算结果;将所述运算结果和所述各个第一特征向量进行相乘运算,得到待识别信息的编码数据。

[0107] 在一个可选的实施例中,确定单元704,用于将编码后的待识别信息输入至第三模型,得到各个子数据的特征信息;根据各个子数据的特征信息,计算各个子数据表征为实体数据的概率;根据各子数据表征为实体数据的概率,确定各个子数据中表征为实体的数据。

[0108] 在一个可选的实施例中,确定单元704,用于在得到运算结果之后,将所述运算结果进行归一化操作;将归一化的所述运算结果与所述各个第一特征向量进行相乘运算,得到所述编码数据。

[0109] 在一个可选的实施例中,所述确定单元704,还用于在确定出目标数据的情况下,对目标数据进行实体类别的划分,确定目标数据所属的实体类别。

[0110] 可以理解,所述信息识别设备中的获得单元701、划分单元702、处理单元703和确定单元704在实际应用中均可由识别设备的中央处理器(CPU,Central Processing Unit)、数字信号处理器(DSP,Digital Signal Processor)、微控制单元(MCU,Microcontroller Unit)或可编程门阵列(FPGA,Field—Programmable Gate Array)实现。

[0111] 需要说明的是,本申请实施例的信息识别设备,由于该信息识别设备解决问题的原理与前述的信息识别方法相似,因此,信息识别设备的实施过程及实施原理均可以参见前述信息识别方法的实施过程及实施原理描述,重复之处不再赘述。

[0112] 本申请实施例还提供一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时至少用于执行图1至图6任一所示方法的步骤。所述计算机可读存储介质具体可以为存储器。所述存储器可以为如图8所示的存储器62。

[0113] 本申请实施例还提供了一种终端。图8为本申请实施例的信息识别设备的硬件结构示意图,如图8所示,信息识别设备包括:用于进行数据传输的通信组件63、至少一个处理器61和用于存储能够在处理器61上运行的计算机程序的存储器62。终端中的各个组件通过总线系统64耦合在一起。可理解,总线系统64用于实现这些组件之间的连接通信。总线系统64除包括数据总线之外,还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见,在图8中将各种总线都标为总线系统64。

[0114] 其中,所述处理器61执行所述计算机程序时至少执行图1至图6任一所示方法的步骤。

[0115] 可以理解,存储器62可以是易失性存储器或非易失性存储器,也可包括易失性和非易失性存储器两者。其中,非易失性存储器可以是只读存储器(ROM,Read Only Memory)、可编程只读存储器(PROM,Programmable Read-Only Memory)、可擦除可编程只读存储器

(EPROM,Erasable Programmable Read-Only Memory)、电可擦除可编程只读存储器(EEPROM,Electrically Erasable Programmable Read-Only Memory)、磁性随机存取存储器(FRAM,ferromagnetic random access memory)、快闪存储器(Flash Memory)、磁表面存储器、光盘、或只读光盘(CD-ROM,Compact Disc Read-Only Memory);磁表面存储器可以是磁盘存储器或磁带存储器。易失性存储器可以是随机存取存储器(RAM,Random Access Memory),其用作外部高速缓存。通过示例性但不是限制性说明,许多形式的RAM可用,例如静态随机存取存储器(SRAM,Static Random Access Memory)、同步静态随机存取存储器(SSRAM,Synchronous Static Random Access Memory)、动态随机存取存储器(DRAM,Dynamic Random Access Memory)、同步动态随机存取存储器(SDRAM,Synchronous Dynamic Random Access Memory)、双倍数据速率同步动态随机存取存储器(DDRSDRAM,Double Data Rate Synchronous Dynamic Random Access Memory)、增强型同步动态随机存取存储器(ESDRAM,Enhanced Synchronous Dynamic Random Access Memory)、同步连接动态随机存取存储器(SLDRAM,SyncLink Dynamic Random Access Memory)、直接内存总线随机存取存储器(DRRAM,Direct Rambus Random Access Memory)。本申请实施例描述的存储器62旨在包括但不限于这些和任意其它适合类型的存储器。

[0116] 上述本申请实施例揭示的方法可以应用于处理器61中,或者由处理器61实现。处理器61可能是一种集成电路芯片,具有信号的处理能力。在实现过程中,上述方法的各步骤可以通过处理器61中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器61可以是通用处理器、DSP,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。处理器61可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者任何常规的处理器等。结合本申请实施例所公开的方法的步骤,可以直接体现为硬件译码处理器执行完成,或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于存储介质中,该存储介质位于存储器62,处理器61读取存储器62中的信息,结合其硬件完成前述方法的步骤。

[0117] 在示例性实施例中,信息识别设备可以被一个或多个应用专用集成电路(ASIC,Application Specific Integrated Circuit)、DSP、可编程逻辑器件(PLD,Programmable Logic Device)、复杂可编程逻辑器件(CPLD,Complex Programmable Logic Device)、FPGA、通用处理器、控制器、MCU、微处理器(Microprocessor)、或其他电子元件实现,用于执行前述的信息识别设备。

[0118] 在本申请所提供的几个实施例中,应该理解到,所揭露的设备和方法,可以通过其它的方式实现。以上所描述的设备实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,如:多个单元或组件可以结合,或可以集成到另一个系统,或一些特征可以忽略,或不执行。另外,所显示或讨论的各组成部分相互之间的耦合、或直接耦合、或通信连接可以是通过一些接口,设备或单元的间接耦合或通信连接,可以是电性的、机械的或其它形式的。

[0119] 上述作为分离部件说明的单元可以是、或也可以不是物理上分开的,作为单元显示的部件可以是、或也可以不是物理单元,即可以位于一个地方,也可以分布到多个网络单元上;可以根据实际的需要选择其中的部分或全部单元来实现本实施例方案的目的。

[0120] 另外,在本申请各实施例中的各功能单元可以全部集成在一个处理单元中,也可

以是各单元分别单独作为一个单元,也可以两个或两个以上单元集成在一个单元中;上述集成的单元既可以采用硬件的形式实现,也可以采用硬件加软件功能单元的形式实现。

[0121] 本领域普通技术人员可以理解:实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成,前述的程序可以存储于一计算机可读取存储介质中,该程序在执行时,执行包括上述方法实施例的步骤;而前述的存储介质包括:移动存储设备、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0122] 或者,本申请上述集成的单元如果以软件功能模块的形式实现并作为独立的产品销售或使用时,也可以存储在一个计算机可读取存储介质中。基于这样的理解,本申请实施例的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机、服务器、或者网络设备等)执行本申请各个实施例所述方法的全部或部分。而前述的存储介质包括:移动存储设备、ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0123] 本申请所提供的几个方法实施例中所揭露的方法,在不冲突的情况下可以任意组合,得到新的方法实施例。

[0124] 本申请所提供的几个产品实施例中所揭露的特征,在不冲突的情况下可以任意组合,得到新的产品实施例。

[0125] 本申请所提供的几个方法或设备实施例中所揭露的特征,在不冲突的情况下可以任意组合,得到新的方法实施例或设备实施例。

[0126] 以上所述,仅为本申请的具体实施方式,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,可轻易想到变化或替换,都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应以所述权利要求的保护范围为准。

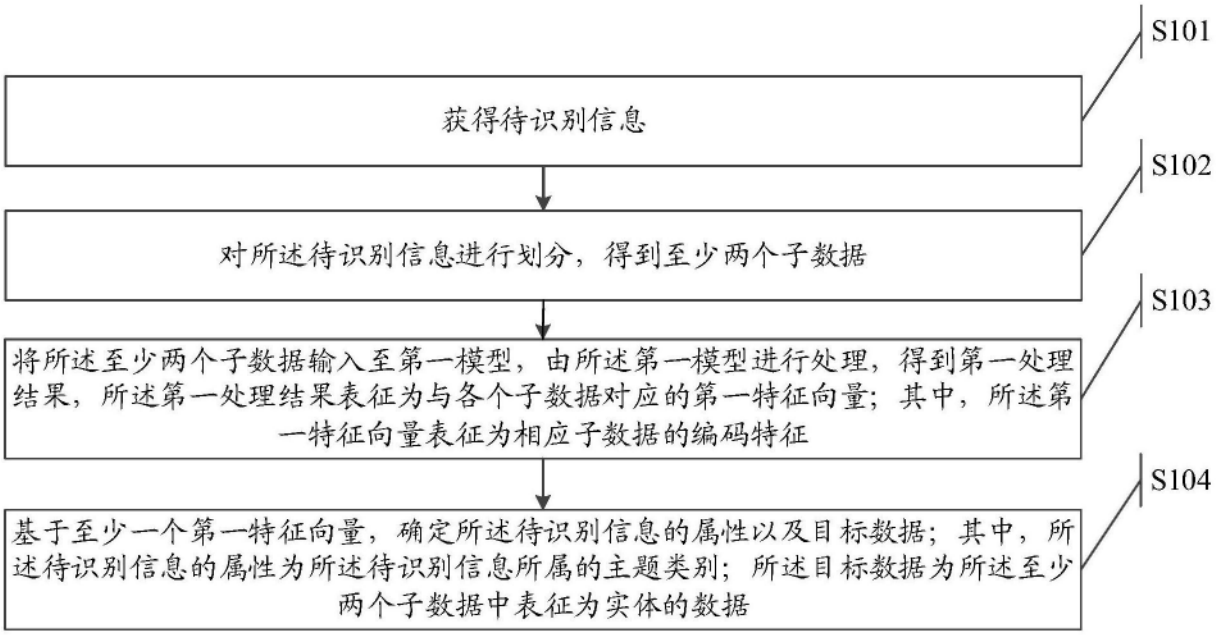

S101
获得待识别信息
S102
对所述待识别信息进行划分，得到至少两个子数据
S103
将所述至少两个子数据输入至第一模型，由所述第一模型进行处理，得到第一处理结果，所述第一处理结果表征为与各个子数据对应的第一特征向量；其中，所述第一特征向量表征为相应子数据的编码特征
S104
基于至少一个第一特征向量，确定所述待识别信息的属性以及目标数据；其中，所述待识别信息的属性为所述待识别信息所属的主题类别；所述目标数据为所述至少两个子数据中表征为实体的数据

图1

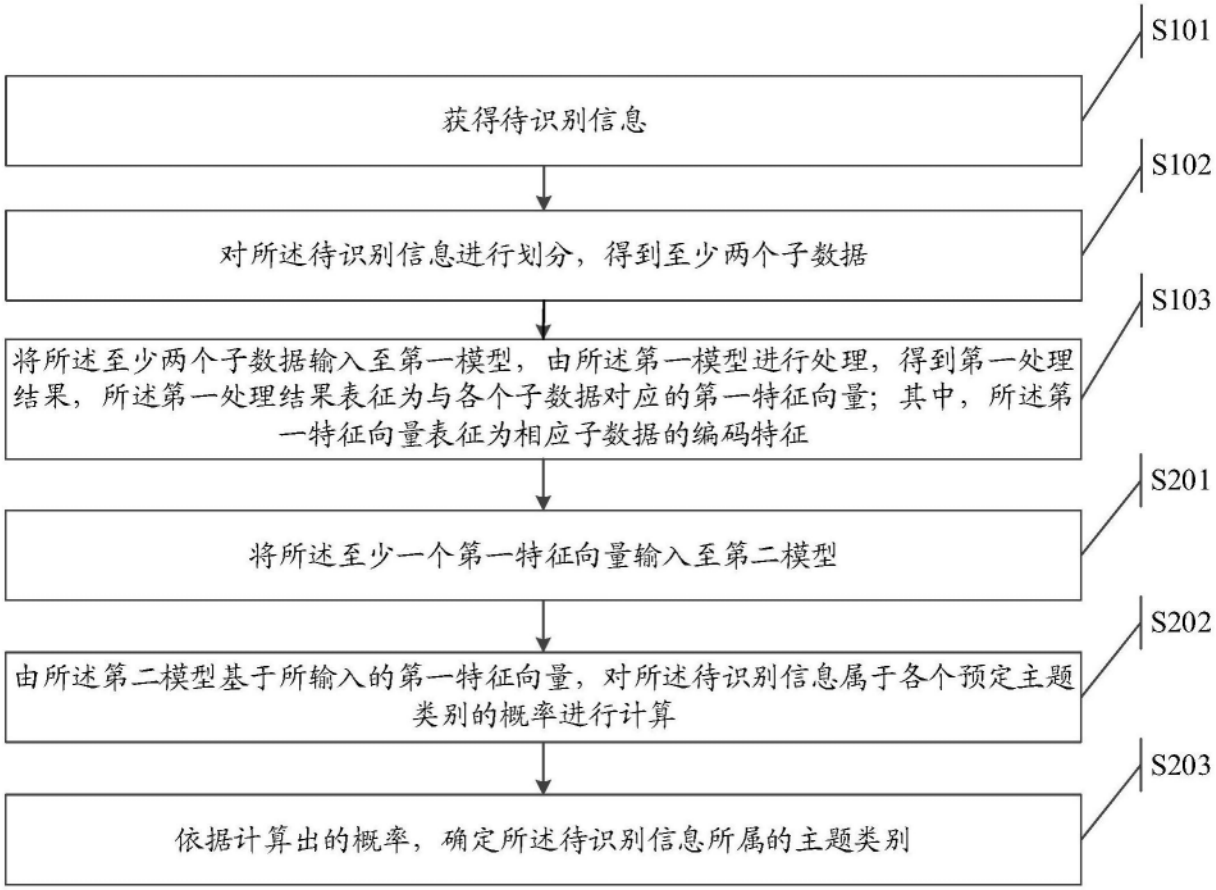

S101
获得待识别信息
S102
对所述待识别信息进行划分，得到至少两个子数据
S103
将所述至少两个子数据输入至第一模型，由所述第一模型进行处理，得到第一处理结果，所述第一处理结果表征为与各个子数据对应的第一特征向量；其中，所述第一特征向量表征为相应子数据的编码特征
S201
将所述至少一个第一特征向量输入至第二模型
S202
由所述第二模型基于所输入的第一特征向量，对所述待识别信息属于各个预定主题类别的概率进行计算
S203
依据计算出的概率，确定所述待识别信息所属的主题类别

图2

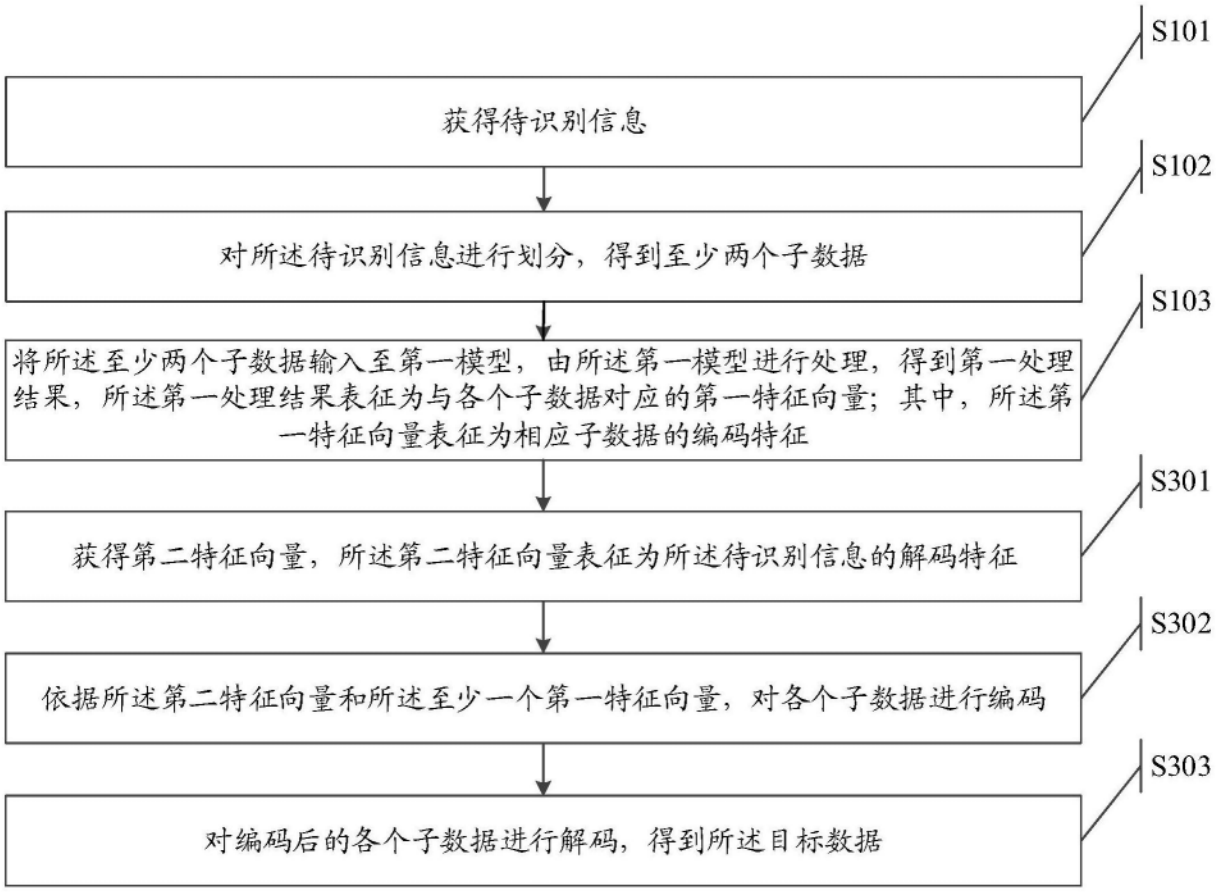

S101
获得待识别信息
S102
对所述待识别信息进行划分，得到至少两个子数据
S103
将所述至少两个子数据输入至第一模型，由所述第一模型进行处理，得到第一处理结果，所述第一处理结果表征为与各个子数据对应的第一特征向量；其中，所述第一特征向量表征为相应子数据的编码特征
S301
获得第二特征向量，所述第二特征向量表征为所述待识别信息的解码特征
S302
依据所述第二特征向量和所述至少一个第一特征向量，对各个子数据进行编码
S303
对编码后的各个子数据进行解码，得到所述目标数据

图3

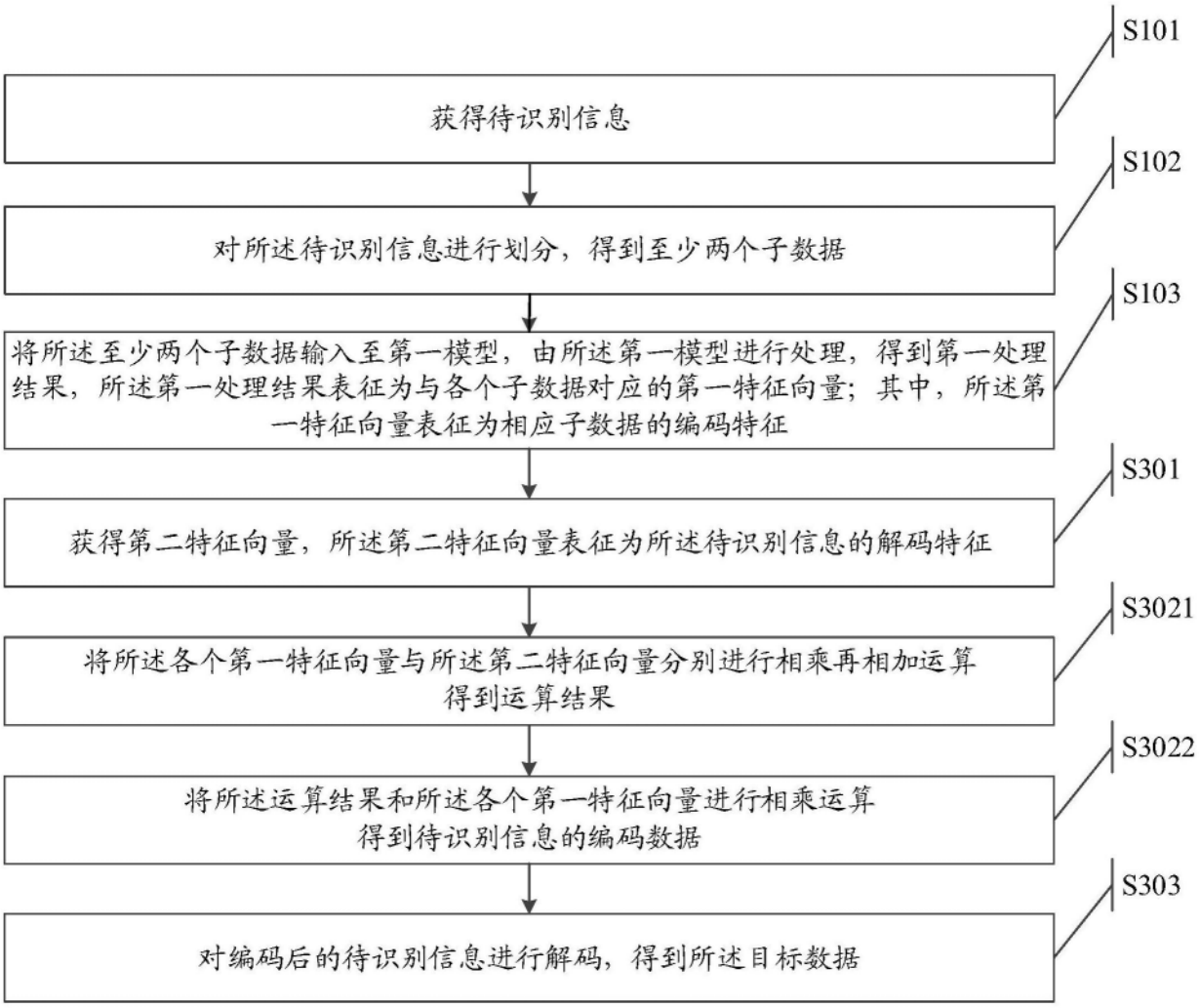

S101
获得待识别信息
S102
对所述待识别信息进行划分，得到至少两个子数据
S103
将所述至少两个子数据输入至第一模型，由所述第一模型进行处理，得到第一处理结果，所述第一处理结果表征为与各个子数据对应的第一特征向量；其中，所述第一特征向量表征为相应子数据的编码特征
S301
获得第二特征向量，所述第二特征向量表征为所述待识别信息的解码特征
S3021
将所述各个第一特征向量与所述第二特征向量分别进行相乘再相加运算得到运算结果
S3022
将所述运算结果和所述各个第一特征向量进行相乘运算得到待识别信息的编码数据
S303
对编码后的待识别信息进行解码，得到所述目标数据

图4


S101
获得待识别信息
S102
对所述待识别信息进行划分，得到至少两个子数据
S103
将所述至少两个子数据输入至第一模型，由所述第一模型进行处理，得到第一处理结果，所述第一处理结果表征为与各个子数据对应的第一特征向量；其中，所述第一特征向量表征为相应子数据的编码特征
S301
获得第二特征向量，所述第二特征向量表征为所述待识别信息的解码特征
S302
依据所述第二特征向量和所述至少一个第一特征向量，对各个子数据进行编码
S3031
将编码后的待识别信息输入至第三模型，得到各个子数据的特征信息；
S3032
根据各个子数据的特征信息，计算各个子数据表征为实体数据的概率
S3033
根据各子数据表征为实体数据的概率，确定各个子数据中表征为实体的数据

图5

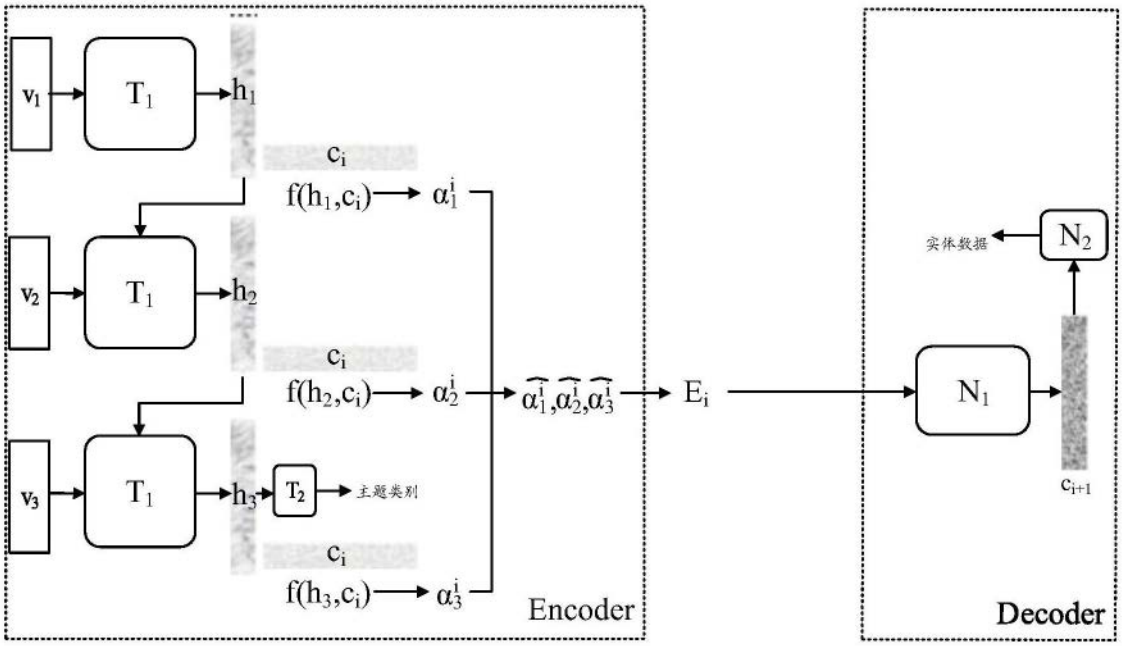

v1
T1
h1
ci
f(h1,ci)
α1i
v2
T1
h2
ci
f(h2,ci)
α2i
α1i,α2i,α3i
Ei
v3
T1
h3
T2
主题类别
ci
f(h3,ci)
α3i
Encoder
实体数据
N2
N1
ci+1
Decoder

图6

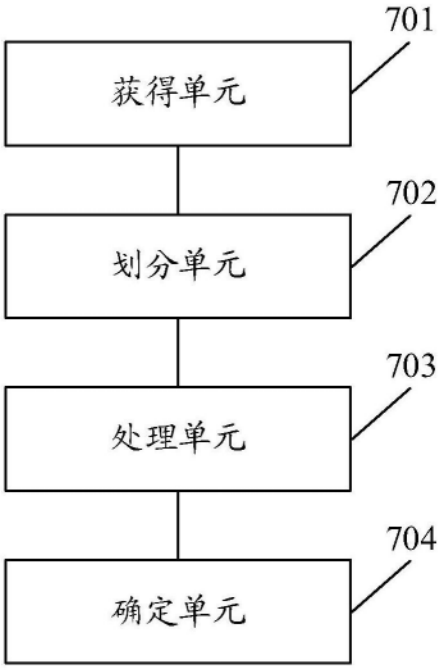

701
获得单元
702
划分单元
703
处理单元
704
确定单元

图7

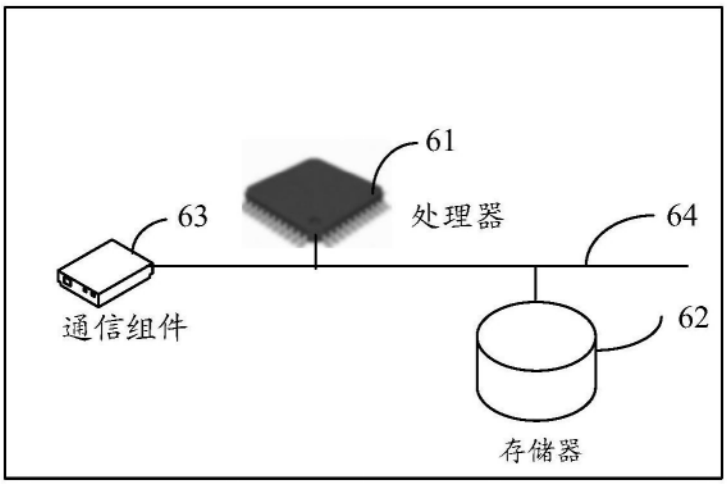


图8