(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110069631 A
(43)申请公布日 2019.07.30

(21)申请号 201910277438.7

(22)申请日 2019.04.08

(71)申请人 腾讯科技(深圳)有限公司
地址 518057 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 陈曦 赖盛章 曹行 张淳 乔倩倩

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

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/36(2019.01)
G06F 16/903(2019.01)
G06F 17/27(2006.01)

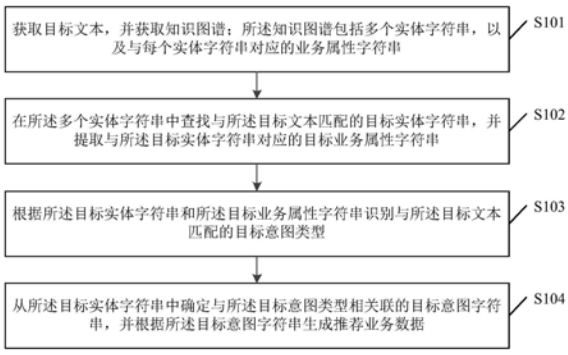
权利要求书4页 说明书25页 附图14页

(54)发明名称

一种文本处理方法、装置以及相关设备

(57)摘要

本发明实施例公开了一种文本处理方法、装置以及相关设备，方法包括：获取目标文本，并获取知识图谱；知识图谱包括多个实体字符串，以及与每个实体字符串对应的业务属性字符串；在多个实体字符串中查找与目标文本匹配的目标实体字符串，并提取与目标实体字符串对应的目标业务属性字符串；根据目标实体字符串和目标业务属性字符串识别与目标文本匹配的目标意图类型；从目标实体字符串中确定与目标意图类型相关联的目标意图字符串，并根据目标意图字符串生成推荐业务数据。采用本发明，可以提高获取业务数据的效率。



CN 110069631 A

1.一种文本处理方法,其特征在于,包括:

获取目标文本,并获取知识图谱;所述知识图谱包括多个实体字符串,以及与每个实体字符串对应的业务属性字符串;

在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串,并提取与所述目标实体字符串对应的目标业务属性字符串;

根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型;

从所述目标实体字符串中确定与所述目标意图类型相关联的目标意图字符串,并根据所述目标意图字符串生成推荐业务数据。

2.根据权利要求1所述的方法,其特征在于,所述在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串,包括:

从所述目标文本中提取目标关键词;

将所述目标关键词映射为图谱标注实体字符串;

从所述多个实体字符串中查找与所述图谱标注实体字符串相同的实体字符串,作为与所述目标文本匹配的目标实体字符串。

3.根据权利要求2所述的方法,其特征在于,所述从所述目标文本中提取目标关键词,包括:

将所述目标文本划分为多个目标单位字符,并将每个目标单位字符转换为目标单位字符向量;

基于第一循环神经网络模型中的编码层,将多个目标单位字符向量进行双向循环编码,得到正向编码矩阵和反向编码矩阵;

将所述正向编码矩阵和所述反向编码矩阵拼接为隐藏状态矩阵;

基于所述第一循环神经网络中的条件随机场,对所述隐藏状态矩阵进行序列标注,确定所述每个目标单位字符分别对应的词性标签;

根据所述每个目标单位字符分别对应的词性标签,确定所述目标关键词。

4.根据权利要求2所述的方法,其特征在于,所述将所述目标关键词映射为图谱标注实体字符串,包括:

将所述目标关键词划分为多个关键单位字符,并将每个关键单位字符转换为关键单位字符向量;

基于第二循环神经网络模型中的编码层,将多个关键单位字符向量进行编码,得到所述目标关键词的上下文向量;

基于所述第二循环神经网络模型中的解码层,将所述上下文向量进行解码,得到所述上下文向量的隐藏状态向量;

识别所述隐藏状态向量,得到与所述隐藏状态向量对应的字符序列,并将所述字符序列确定为所述图谱标注实体字符串。

5.根据权利要求1所述的方法,其特征在于,所述根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型,包括:

将所述目标实体字符串转换为目标实体词向量,并将所述目标业务属性字符串转换为目标业务属性词向量;

将所述目标实体词向量和所述目标业务属性词向量组合为输入向量，基于意图识别模型中的卷积层和池化层，对所述输入向量进行卷积和池化，得到所述目标文本的意图特征向量；

基于所述意图识别模型中的分类器，识别所述意图特征向量与所述分类器中多种意图类型之间的匹配概率，从多个匹配概率中将最大匹配概率对应的意图类型确定为所述目标意图类型。

6.根据权利要求5所述的方法，其特征在于，所述将所述目标实体字符串转换为目标实体词向量，包括：

从实体词袋中查找与所述目标实体字符串对应的独热码，作为第一向量；所述实体词袋包括所述知识图谱中的所述多个实体字符串，以及与所述每个实体字符串分别对应的独热码；

基于实体词向量转换模型中的隐藏层，将所述第一向量进行降维，得到所述目标实体词向量。

7.根据权利要求1所述的方法，其特征在于，所述从所述目标实体字符串中确定与所述目标意图类型相关联的目标意图字符串，并根据所述目标意图字符串生成推荐业务数据，包括：

获取所述目标意图类型所表征的业务属性，作为意图业务属性；

将具有所述意图业务属性的目标实体字符串作为所述目标意图字符串；

查找与所述目标意图字符串相关联的业务数据，并将与所述目标意图字符串相关联的业务数据作为所述推荐业务数据。

8.根据权利要求7所述的方法，其特征在于，所述查找与所述目标意图字符串相关联的业务数据，并将与所述目标意图字符串相关联的业务数据作为所述推荐业务数据，包括：

当所述目标意图类型属于语义推理类型时，从所述知识图谱的多个实体字符串中提取多个第一实体字符串；所述第一实体字符串是在所述知识图谱的多个实体字符串中除具有所述意图业务属性的实体字符串以外的实体字符串；

在所述知识图谱中，确定所述目标意图字符串与每个第一实体字符串之间的目标相似系数，根据多个目标相似系数，从所述多个第一实体字符串中选择与所述目标意图类型匹配的目标第一实体字符串；

查找与所述目标第一实体字符串相关联的业务数据，并将与所述目标第一实体字符串相关联的业务数据作为所述推荐业务数据。

9.根据权利要求8所述的方法，其特征在于，所述知识图谱还包括多个实体字符串之间的关联匹配系数；所述关联匹配系数是通过与所述知识图谱关联的数据集合统计得到；

所述在所述知识图谱中，确定所述目标意图字符串与每个第一实体字符串之间的目标相似系数，包括：

从所述多个第一实体字符串中提取用于轮询的测试第一实体字符串；

将所述目标意图字符串与所述测试第一实体字符串之间的关联匹配系数，作为所述目标意图字符串与所述测试第一实体字符串之间的目标相似系数；

当每个第一实体字符串均被确定为测试第一实体字符串时，停止轮询。

10.根据权利要求8所述的方法，其特征在于，所述在所述知识图谱中，确定所述目标意

图字符串与每个第一实体字符串之间的目标相似系数，包括：

将所述知识图谱中具有所述意图业务属性的多个实体字符串，均作为第二实体字符串；

获取每个第一实体字符串的第一图向量，获取每个第二实体字符串的第二词向量和第二图向量；

从多个第二词向量中提取所述目标意图字符串对应的第二词向量，作为目标词向量，并从多个第二图向量中提取所述目标意图字符串对应的第二图向量，作为目标图向量；

根据所述目标实体词向量、所述目标图向量、多个第一图向量和所述多个第二词向量，确定所述目标意图字符串和每个第一实体字符串之间的目标相似系数。

11.根据权利要求10所述的方法，其特征在于，所述知识图谱还包括所述多个实体字符串之间的关联匹配系数；

所述获取每个第一实体字符串的第一图向量，包括：

初始化知识图谱中每个实体字符串分别对应的原始图向量；

从所述知识图谱中采样多个样本实体字符串，并获取每个样本实体字符串分别对应的原始图向量，作为样本图向量，根据所述多个样本实体字符串之间的关联匹配系数，采用梯度下降规则分别更新每个样本图向量，得到调整图向量，并将所述调整图向量确定为所述原始图向量；

当采样次数达到次数阈值时，将所述每个第一实体字符串分别对应的更新后的原始图向量，作为所述第一图向量。

12.根据权利要求11所述的方法，其特征在于，所述根据所述目标实体词向量、所述目标图向量、多个第一图向量和所述多个第二词向量，确定所述目标意图字符串和每个第一实体字符串之间的目标相似系数，包括：

从所述多个第一实体字符串中提取用于轮询的测试第一实体字符串，将所述测试第一实体字符串对应第一图向量作为测试图向量；

根据所述关联匹配系数，确定在所述知识图谱中与所述测试第一实体字符串相邻的第二实体字符串的第二词向量与所述目标词向量之间的第一相似系数；

根据所述关联匹配系数，确定所述测试图向量与所述目标图向量之间的第二相似系数；

根据所述第一相似系数与所述第二相似系数，生成所述目标意图字符串与所述测试第一实体字符串之间的目标相似系数；

当每个第一实体字符串均被确定为测试第一实体字符串时，停止轮询。

13.一种文本处理装置，其特征在于，包括：

获取模块，用于获取目标文本，并获取知识图谱；所述知识图谱包括多个实体字符串，以及与每个实体字符串对应的业务属性字符串；

查找模块，用于在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串；

所述获取模块，还用于提取与所述目标实体字符串对应的目标业务属性字符串；

识别模块，用于根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型；

生成模块，用于从所述目标实体字符串中确定与所述目标意图类型相关联的目标意图字符串，并根据所述目标意图字符串生成推荐业务数据。

14.一种电子设备，其特征在于，包括：处理器和存储器；

所述处理器和存储器相连，其中，所述存储器用于存储程序代码，所述处理器用于调用所述程序代码，以执行如权利要求1-12任一项所述的方法。

15.一种计算机存储介质，其特征在于，所述计算机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，执行如权利要求1-12任一项所述的方法。

一种文本处理方法、装置以及相关设备

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种文本处理方法、装置以及相关设备。

背景技术

[0002] 随着信息技术的飞速发展，用户足不出户在互联网上就可以实现远程挂号。然而对于大多数用户来说，自己应该挂哪一个科室，做什么检查等，往往是通过自己以往的经验，或者通过自己的症状查询相关资料来确定的。用户自我诊断后，在互联网上找到对应的医院以及科室挂号预约。

[0003] 然而由于用户自我诊断时的非专业性，造成挂号科室与用户的症状并不匹配，导致用户需要进行多次挂号操作、以及多次看诊后才可以找到与症状最匹配的科室，进而降低用户找到正确科室的效率。

发明内容

[0004] 本发明实施例提供一种文本处理方法、装置以及相关设备，可以提高获取业务数据的效率。

[0005] 本发明实施例一方面提供了一种文本处理方法，包括：

[0006] 获取目标文本，并获取知识图谱；所述知识图谱包括多个实体字符串，以及与每个实体字符串对应的业务属性字符串；

[0007] 在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串，并提取与所述目标实体字符串对应的目标业务属性字符串；

[0008] 根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型；

[0009] 从所述目标实体字符串中确定与所述目标意图类型相关联的目标意图字符串，并根据所述目标意图字符串生成推荐业务数据。

[0010] 其中，所述在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串，包括：

[0011] 从所述目标文本中提取目标关键词；

[0012] 将所述目标关键词映射为图谱标注实体字符串；

[0013] 从所述多个实体字符串中查找与所述图谱标注实体字符串相同的实体字符串，作为与所述目标文本匹配的目标实体字符串。

[0014] 其中，所述从所述目标文本中提取目标关键词，包括：

[0015] 将所述目标文本划分为多个目标单位字符，并将每个目标单位字符转换为目标单位字符向量；

[0016] 基于第一循环神经网络模型中的编码层，将多个目标单位字符向量进行双向循环编码，得到正向编码矩阵和反向编码矩阵；

[0017] 将所述正向编码矩阵和所述反向编码矩阵拼接为隐藏状态矩阵；

[0018] 基于所述第一循环神经网络中的条件随机场，对所述隐藏状态矩阵进行序列标注，确定所述每个目标单位字符分别对应的词性标签；

[0019] 根据所述每个目标单位字符分别对应的词性标签，确定所述目标关键词。

[0020] 其中，所述将所述目标关键词映射为图谱标注实体字符串，包括：

[0021] 将所述目标关键词划分为多个关键单位字符，并将每个关键单位字符转换为关键单位字符向量；

[0022] 基于第二循环神经网络模型中的编码层，将多个关键单位字符向量进行编码，得到所述目标关键词的上下文向量；

[0023] 基于所述第二循环神经网络模型中的解码层，将所述上下文向量进行解码，得到所述上下文向量的隐藏状态向量；

[0024] 识别所述隐藏状态向量，得到与所述隐藏状态向量对应的字符序列，并将所述字符序列确定为所述图谱标注实体字符串。

[0025] 其中，所述根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型，包括：

[0026] 将所述目标实体字符串转换为目标实体词向量，并将所述目标业务属性字符串转换为目标业务属性词向量；

[0027] 将所述目标实体词向量和所述目标业务属性词向量组合为输入向量，基于意图识别模型中的卷积层和池化层，对所述输入向量进行卷积和池化，得到所述目标文本的意图特征向量；

[0028] 基于所述意图识别模型中的分类器，识别所述意图特征向量与所述分类器中多种意图类型之间的匹配概率，从多个匹配概率中将最大匹配概率对应的意图类型确定为所述目标意图类型。

[0029] 其中，所述将所述目标实体字符串转换为目标实体词向量，包括：

[0030] 从实体词袋中查找与所述目标实体字符串对应的独热码，作为第一向量；所述实体词袋包括所述知识图谱中的所述多个实体字符串，以及与所述每个实体字符串分别对应的独热码；

[0031] 基于实体词向量转换模型中的隐藏层，将所述第一向量进行降维，得到所述目标实体词向量。

[0032] 其中，所述从所述目标实体字符串中确定与所述目标意图类型相关联的目标意图字符串，并根据所述目标意图字符串生成推荐业务数据，包括：

[0033] 获取所述目标意图类型所表征的业务属性，作为意图业务属性；

[0034] 将具有所述意图业务属性的目标实体字符串作为所述目标意图字符串；

[0035] 查找与所述目标意图字符串相关联的业务数据，并将与所述目标意图字符串相关联的业务数据作为所述推荐业务数据。

[0036] 其中，所述查找与所述目标意图字符串相关联的业务数据，并将与所述目标意图字符串相关联的业务数据作为所述推荐业务数据，包括：

[0037] 当所述目标意图类型属于语义推理类型时，从所述知识图谱的多个实体字符串中提取多个第一实体字符串；所述第一实体字符串是在所述知识图谱的多个实体字符串中除具有所述意图业务属性的实体字符串以外的实体字符串；

[0038] 在所述知识图谱中，确定所述目标意图字符串与每个第一实体字符串之间的目标相似系数，根据多个目标相似系数，从所述多个第一实体字符串中选择与所述目标意图类型匹配的目标第一实体字符串；

[0039] 查找与所述目标第一实体字符串相关联的业务数据，并将与所述目标第一实体字符串相关联的业务数据作为所述推荐业务数据。

[0040] 其中，所述知识图谱还包括多个实体字符串之间的关联匹配系数；所述关联匹配系数是通过与所述知识图谱关联的数据集合统计得到；

[0041] 所述在所述知识图谱中，确定所述目标意图字符串与每个第一实体字符串之间的目标相似系数，包括：

[0042] 从所述多个第一实体字符串中提取用于轮询的测试第一实体字符串；

[0043] 将所述目标意图字符串与所述测试第一实体字符串之间的关联匹配系数，作为所述目标意图字符串与所述测试第一实体字符串之间的目标相似系数；

[0044] 当每个第一实体字符串均被确定为测试第一实体字符串时，停止轮询。

[0045] 其中，所述在所述知识图谱中，确定所述目标意图字符串与每个第一实体字符串之间的目标相似系数，包括：

[0046] 将所述知识图谱中具有所述意图业务属性的多个实体字符串，均作为第二实体字符串；

[0047] 获取每个第一实体字符串的第一图向量，获取每个第二实体字符串的第二词向量和第二图向量；

[0048] 从多个第二词向量中提取所述目标意图字符串对应的第二词向量，作为目标词向量，并从多个第二图向量中提取所述目标意图字符串对应的第二图向量，作为目标图向量；

[0049] 根据所述目标实体词向量、所述目标图向量、多个第一图向量和所述多个第二词向量，确定所述目标意图字符串和每个第一实体字符串之间的目标相似系数。

[0050] 其中，所述知识图谱还包括所述多个实体字符串之间的关联匹配系数；

[0051] 所述获取每个第一实体字符串的第一图向量，包括：

[0052] 初始化知识图谱中每个实体字符串分别对应的原始图向量；

[0053] 从所述知识图谱中采样多个样本实体字符串，并获取每个样本实体字符串分别对应的原始图向量，作为样本图向量，根据所述多个样本实体字符串之间的关联匹配系数，采用梯度下降规则分别更新每个样本图向量，得到调整图向量，并将所述调整图向量确定为所述原始图向量；

[0054] 当采样次数达到次数阈值时，将所述每个第一实体字符串分别对应的更新后的原始图向量，作为所述第一图向量。

[0055] 其中，所述根据所述目标实体词向量、所述目标图向量、多个第一图向量和所述多个第二词向量，确定所述目标意图字符串和每个第一实体字符串之间的目标相似系数，包括：

[0056] 从所述多个第一实体字符串中提取用于轮询的测试第一实体字符串，将所述测试第一实体字符串对应第一图向量作为测试图向量；

[0057] 根据所述关联匹配系数，确定在所述知识图谱中与所述测试第一实体字符串相邻的第二实体字符串的第二词向量与所述目标词向量之间的第一相似系数；

[0058] 根据所述关联匹配系数，确定所述测试图向量与所述目标图向量之间的第二相似系数；

[0059] 根据所述第一相似系数与所述第二相似系数，生成所述目标意图字符串与所述测试第一实体字符串之间的目标相似系数；

[0060] 当每个第一实体字符串均被确定为测试第一实体字符串时，停止轮询。

[0061] 本发明实施例另一方面提供了一种文本处理装置，包括：

[0062] 获取模块，用于获取目标文本，并获取知识图谱；所述知识图谱包括多个实体字符串，以及与每个实体字符串对应的业务属性字符串；

[0063] 查找模块，用于在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串；

[0064] 所述获取模块，还用于提取与所述目标实体字符串对应的目标业务属性字符串；

[0065] 识别模块，用于根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型；

[0066] 生成模块，用于从所述目标实体字符串中确定与所述目标意图类型相关联的目标意图字符串，并根据所述目标意图字符串生成推荐业务数据。

[0067] 其中，所述查找模块，包括：

[0068] 提取单元，用于从所述目标文本中提取目标关键词；

[0069] 映射单元，用于将所述目标关键词映射为图谱标注实体字符串；

[0070] 查找单元，用于从所述多个实体字符串中查找与所述图谱标注实体字符串相同的实体字符串，作为与所述目标文本匹配的目标实体字符串。

[0071] 其中，所述提取单元，包括：

[0072] 划分子单元，用于将所述目标文本划分为多个目标单位字符，并将每个目标单位字符转换为目标单位字符向量；

[0073] 编码子单元，用于基于第一循环神经网络模型中的编码层，将多个目标单位字符向量进行双向循环编码，得到正向编码矩阵和反向编码矩阵；

[0074] 拼接子单元，用于将所述正向编码矩阵和所述反向编码矩阵拼接为隐藏状态矩阵；

[0075] 所述拼接子单元，还用于基于所述第一循环神经网络中的条件随机场，对所述隐藏状态矩阵进行序列标注，确定所述每个目标单位字符分别对应的词性标签；

[0076] 所述拼接子单元，还用于根据所述每个目标单位字符分别对应的词性标签，确定所述目标关键词。

[0077] 其中，所述映射单元，包括：

[0078] 转换子单元，用于将所述目标关键词划分为多个关键单位字符，并将每个关键单位字符转换为关键单位字符向量；

[0079] 所述转换子单元，还用于基于第二循环神经网络模型中的编码层，将多个关键单位字符向量进行编码，得到所述目标关键词的上下文向量；

[0080] 解码子单元，用于基于所述第二循环神经网络模型中的解码层，将所述上下文向量进行解码，得到所述上下文向量的隐藏状态向量；

[0081] 所述解码子单元，还用于识别所述隐藏状态向量，得到与所述隐藏状态向量对应

的字符序列，并将所述字符序列确定为所述图谱标注实体字符串。

[0082] 其中，所述识别模块，包括：

[0083] 转换单元，用于将所述目标实体字符串转换为目标实体词向量；

[0084] 组合单元，用于将所述目标业务属性字符串转换为目标业务属性词向量；

[0085] 所述组合单元，还用于将所述目标实体词向量和所述目标业务属性词向量组合为输入向量，基于意图识别模型中的卷积层和池化层，对所述输入向量进行卷积和池化，得到所述目标文本的意图特征向量；

[0086] 识别单元，用于基于所述意图识别模型中的分类器，识别所述意图特征向量与所述分类器中多种意图类型之间的匹配概率，从多个匹配概率中将最大匹配概率对应的意图类型确定为所述目标意图类型。

[0087] 其中，所述转换单元，包括：

[0088] 查找子单元，用于从实体词袋中查找与所述目标实体字符串对应的独热码，作为第一向量；所述实体词袋包括所述知识图谱中的所述多个实体字符串，以及与所述每个实体字符串分别对应的独热码；

[0089] 降维子单元，用于基于实体词向量转换模型中的隐藏层，将所述第一向量进行降维，得到所述目标实体词向量。

[0090] 其中，所述生成模块，包括：

[0091] 获取单元，用于获取所述目标意图类型所表征的业务属性，作为意图业务属性；

[0092] 确定单元，用于将具有所述意图业务属性的目标实体字符串作为所述目标意图字符串；

[0093] 生成单元，用于查找与所述目标意图字符串相关联的业务数据，并将与所述目标意图字符串相关联的业务数据作为所述推荐业务数据。

[0094] 其中，所述生成单元，包括：

[0095] 第一提取子单元，用于当所述目标意图类型属于语义推理类型时，从所述知识图谱的多个实体字符串中提取多个第一实体字符串；所述第一实体字符串是在所述知识图谱的多个实体字符串中除具有所述意图业务属性的实体字符串以外的实体字符串；

[0096] 选择子单元，用于在所述知识图谱中，确定所述目标意图字符串与每个第一实体字符串之间的目标相似系数；

[0097] 所述第一提取子单元，还用于根据多个目标相似系数，从所述多个第一实体字符串中选择与所述目标意图类型匹配的目标第一实体字符串；

[0098] 确定子单元，用于查找与所述目标第一实体字符串相关联的业务数据，并将与所述目标第一实体字符串相关联的业务数据作为所述推荐业务数据。

[0099] 其中，所述知识图谱还包括多个实体字符串之间的关联匹配系数；所述关联匹配系数是通过与所述知识图谱关联的数据集合统计得到；

[0100] 所述选择子单元，包括：

[0101] 开始子单元，用于从所述多个第一实体字符串中提取用于轮询的测试第一实体字符串；

[0102] 所述开始子单元，还用于将所述目标意图字符串与所述测试第一实体字符串之间的关联匹配系数，作为所述目标意图字符串与所述测试第一实体字符串之间的目标相似系

数；

[0103] 停止子单元，用于当每个第一实体字符串均被确定为测试第一实体字符串时，停止轮询。

[0104] 其中，所述选择子单元，包括：

[0105] 第一获取子单元，用于将所述知识图谱中具有所述意图业务属性的多个实体字符串，均作为第二实体字符串；

[0106] 第二获取子单元，用于获取每个第一实体字符串的第一图向量；

[0107] 所述第一获取子单元，还用于获取每个第二实体字符串的第二词向量和第二图向量；

[0108] 第二提取子单元，用于从多个第二词向量中提取所述目标意图字符串对应的第二词向量，作为目标词向量，并从多个第二图向量中提取所述目标意图字符串对应的第二图向量，作为目标图向量；

[0109] 第三提取子单元，用于根据所述目标实体词向量、所述目标图向量、多个第一图向量和所述多个第二词向量，确定所述目标意图字符串和每个第一实体字符串之间的目标相似系数。

[0110] 其中，所述知识图谱还包括所述多个实体字符串之间的关联匹配系数；

[0111] 所述第二获取子单元，包括：

[0112] 初始化子单元，用于初始化知识图谱中每个实体字符串分别对应的原始图向量；

[0113] 采样子单元，用于从所述知识图谱中采样多个样本实体字符串，并获取每个样本实体字符串分别对应的原始图向量，作为样本图向量，根据所述多个样本实体字符串之间的关联匹配系数，采用梯度下降规则分别更新每个样本图向量，得到调整图向量，并将所述调整图向量确定为所述原始图向量；

[0114] 所述采样子单元，还用于当采样次数达到次数阈值时，将所述每个第一实体字符串分别对应的更新后的原始图向量，作为所述第一图向量。

[0115] 其中，所述第三提取子单元，包括：

[0116] 第一生成子单元，用于从所述多个第一实体字符串中提取用于轮询的测试第一实体字符串，将所述测试第一实体字符串对应第一图向量作为测试图向量；

[0117] 所述第一生成子单元，还用于根据所述关联匹配系数，确定在所述知识图谱中与所述测试第一实体字符串相邻的第二实体字符串的第二词向量与所述目标词向量之间的第一相似系数；

[0118] 所述第一生成子单元，还用于根据所述关联匹配系数，确定所述测试图向量与所述目标图向量之间的第二相似系数；

[0119] 第二生成子单元，用于根据所述第一相似系数与所述第二相似系数，生成所述目标意图字符串与所述测试第一实体字符串之间的目标相似系数；

[0120] 所述第二生成子单元，还用于当每个第一实体字符串均被确定为测试第一实体字符串时，停止轮询。

[0121] 本发明实施例另一方面提供了一种电子设备，包括：处理器和存储器；

[0122] 所述处理器和存储器相连，其中，所述存储器用于存储程序代码，所述处理器用于调用所述程序代码，以执行如本发明实施例中一方面中的方法。

[0123] 本发明实施例另一方面提供了一种计算机存储介质，所述计算机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，执行如本发明实施例中一方面中的方法。

[0124] 本发明实施例通过获取目标文本并获取包含多个实体字符串以及业务属性字符串的知识图谱，在知识图谱中确定目标实体字符串以及目标业务属性字符串，后续根据目标实体字符串以及目标业务属性字符串确定目标意图类型，并根据目标意图类型确定目标意图字符串，进而确定推荐业务数据。上述可知，用户只需要输入文本，终端就可以识别出用户的意图，并生成满足用户意图的业务数据，相比通过手动查找资料并人工执行多次业务行为以确定业务数据，本发明所采用的自动化的方式生成满足用户心理预期的业务数据，可以节约用户执行多次业务行为所耗费的时间，进而可以提高获取业务数据的效率。

附图说明

[0125] 为了更清楚地说明本发明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。

[0126] 图1是本发明实施例提供的一种文本处理的系统架构图；

[0127] 图2a-图2b是本发明实施例提供的一种文本处理的场景示意图；

[0128] 图3是本发明实施例提供的一种文本处理方法的流程示意图；

[0129] 图4是本发明实施例提供的一种知识图谱的示意图；

[0130] 图5是本发明实施例提供的一种文本处理的界面示意图；

[0131] 图6本发明实施例提供的一种确定推荐业务数据的场景示意图；

[0132] 图7本发明实施例提供的另一种确定推荐业务数据的场景示意图；

[0133] 图8是本发明实施例提供的另一种文本处理方法的流程示意图；

[0134] 图9是本发明实施例提供的另一个文本识别方法的流程示意图；

[0135] 图10是本发明实施例提供的另一种知识图谱的示意图；

[0136] 图11是本发明实施例提供的一种挂号转化率对比图；

[0137] 图12是本发明实施例提供的一种文本处理的系统结构图；

[0138] 图13是本发明实施例提供的一种文本处理装置的结构示意图；

[0139] 图14是是本发明实施例提供的一种电子设备的结构示意图。

具体实施方式

[0140] 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作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0141] 请参见图1，是本发明实施例提供的一种文本处理的系统架构图。服务器10f通过交换机10e和通信总线10d与用户终端集群建立连接，用户终端集群可以包括：用户终端10a、用户终端10b、...、用户终端10c。

[0142] 以用户终端10a为例，当用户终端10a获取到用户输入的文本时，用户终端10a将上述文本通过交换机10e和通信总线10d发送至服务器10f。服务器10f对应的数据库10g中存储了知识图谱，该知识图谱包括多个实体字符串以及每个实体字符串对应的业务属性字符串。服务器10f根据目标文本以及知识图谱确定用户的意图类型，进而根据意图类型确定对应的推荐业务数据，例如，推荐业务数据可以为医生排班表信息、科室挂号链接等。服务器10f可以将确定的推荐业务数据发送至用户终端10a，后续用户终端10a可以在屏幕中显示该推荐业务数据。

[0143] 当然，若用户终端10a本地存储了知识图谱，也可以直接由用户终端10a根据该知识图谱以及目标文本确定用户的意图类型，同样地根据意图类型确定推荐业务数据。下述以用户终端10a如何确定意图类型，以及如何确定推荐业务数据为例进行具体的说明。其中，图1所示的用户终端10a、用户终端10b、用户终端10c等可以包括手机、平板电脑、笔记本电脑、掌上电脑、移动互联网设备(MID,mobile internet device)、可穿戴设备(例如智能手表、智能手环等)等。

[0144] 请参见图2a-图2b，是本发明实施例提供的一种文本处理的场景示意图。下述以医疗导诊系统为例，如图2a中的智能导诊应答页20a所示，用户在智能导诊应答页20a中输入“肚子鼓鼓的”、“恶心”，用户终端10a将“肚子鼓鼓的，恶心”作为目标文本20b。

[0145] 用户终端10a可以基于实体识别模型，提取目标文本20b中的关键词集合20c：“肚子鼓鼓”、“恶心”。其中，实体识别模型可以是根据LSTM(Long Short-Term Memory，长短期记忆网络)算法+CRF(Conditional Random Field，条件随机场)算法训练得到的。

[0146] 提取关键词集合20c的具体过程为：用户终端10a首先将目标文本20b中的每个字符进行划分，采用one-hot编码方式将每个字符编码为字符向量，用户终端10a将多个字符向量按照前后顺序依次输入实体识别模型中的编码层，即是采用LSTM分别对每个字符向量进行前向编码和反向编码，得到前向编码矩阵和反向编码矩阵，其中该矩阵的行数对应字符数量，矩阵的列数对应每个字符基于LSTM编码后的向量。将前向编码矩阵和反向编码矩阵拼接为隐藏状态矩阵，分别对隐藏状态矩阵中的每一行进行识别，确定对应的词性标签，即是确定目标文本20b中每个字符的词性标签，其中词性标签可以包括：B-Person(人名的开始部分)、I-Person(人名的中间部分)、B-Organization(组织机构的开始部分)、I-Organization(组织机构的中间部分)、O(非实体信息)。

[0147] 用户终端10a确定了目标文本20b中的每个字符的词性标签后，可以将词性标签为“B-Person”到“I-Person”之间的字符序列作为关键词，将词性标签为“B-Organization”到“I-Organization”之间的字符序列作为关键词，词性标签为“O”的字符是非实体字符，可以直接忽略。

[0148] 由于关键词集合20c是采用自然语言描述的，因此还需要用户终端10a基于转换模型将该关键词集合20c中的关键词转换为书面化的专业词汇，即将关键词集合20c中的“肚子鼓鼓”转换为“腹胀”，由于“恶心”已经是书面化的专业词汇，因此可以不作转换。即，转换后的“腹胀”、“恶心”可以组合为标准实体字符串集合20d。

[0149] 上述转换过程类似于机器语言翻译过程，例如，将中文翻译为同语义的英文。上述转换模型可以是根据LSTM以及大量的语料训练得到的，为了提高后续与知识图谱的匹配准确率，构建知识图谱所使用的语料与转换模型所使用的语料可以相同。

[0150] 当然也可以构建多个X→Y映射关系，其中X是口语化的关键词，Y是与该关键词同义的书面化的专业术语，通过查找对应的映射关系，也可以将关键词集合20c转换为标准实体字符串集合20d。

[0151] 用户终端10a获取知识图谱20e，知识图谱20e包括多个实体字符串，以及每个实体字符串对应的业务属性字符串。由于当前场景是医疗场景，因此知识图谱20e是关于医疗的知识图谱，即知识图谱中的实体字符串是关于医生名称、症状名称、疾病名称、药物名称的字符串。业务属性字符串是用于标识对应实体字符串的实体所在的业务领域，例如知识图谱20e中，实体字符串“腹泻”的业务属性字符串为：症状；实体字符串“消化科”的业务属性字符串为：科室；实体字符串“张三”的业务属性字符串为：医生。

[0152] 用户终端10a在知识图谱20e所包含的多个实体字符串中，查找与标准实体字符串集合20d相同的实体字符串，作为目标实体字符串，并在知识图谱20e中查找与目标实体字符串对应的业务属性字符串，即在知识图谱20e中查找到目标实体字符串“腹胀”对应的业务属性字符串为：症状；目标实体字符串“恶心”对应的业务属性字符串为：症状。可以将上述目标实体字符串以及对应的业务属性字符串组合为文本20f。

[0153] 用户终端10a首先将文本20f中的字符串“腹胀”转换为实体词向量20g，将文本20f中的字符串“症状”转换为属性词向量20h，将文本20f中的字符串“恶心”转换为实体词向量20m，将实体词向量20g以及属性词向量20h拼接为第一输入向量，可以知道上述第一输入向量是关于文本20f中“腹胀、症状”的词向量；将实体词向量20m以及属性词向量20h拼接为第二输入向量，上述第二输入向量是关于文本20f中“恶心、症状”的词向量。

[0154] 用户终端10a将上述第一输入向量和第二输入向量组合为输入矩阵，并获取意图识别模型20k，意图识别模型可以识别输入矩阵的意图类型，意图类型可以包括：找医生、找科室、挂号科室推荐、检验检查、流程相关、其它。意图识别模型20k包括卷积层、池化层和分类器，其中卷积层和池化层是用于对输入矩阵进行卷积运算和池化运算，经过卷积运算和池化运算后，可以提取出目标文本的意图特征向量。

[0155] 意图识别模型20k中的分类器可以识别提取出来的意图特征向量与多个意图类型之间的相关性，即是识别出目标文本与多种意图类型之间的匹配概率，如图2a所示，目标文本20b与意图类型“找医生”的匹配概率为：0.05；目标文本20b与意图类型“找科室”的匹配概率为：0.03；目标文本20b与意图类型“挂号科室推荐”的匹配概率为：0.75；目标文本20b与意图类型“检验检查”的匹配概率为：0.07；目标文本20b与意图类型“其他”的匹配概率为：0.07。

[0156] 根据上述多个匹配概率，用户终端10a可以确定与目标文本20b匹配的意图类型是：挂号科室推荐。

[0157] 用户终端10a确定了意图类型后，根据意图类型“挂号科室推荐”确定推荐业务数据。

[0158] 确定推荐业务数据的具体过程为：用户终端10a获取知识图谱20e中各实体字符串之间的匹配系数，该匹配系数是用于表示两个实体字符串的共现概率，共现概率越高，说明两个实体字符串对应的实体越匹配。用户终端10a在知识图谱20e中，计算各实体字符串分别与目标实体字符串“腹胀”以及与目标实体字符串“恶心”之间的匹配系数，将业务属性字符串为：“科室”，且与目标实体字符串“腹胀”以及与目标实体字符串“恶心”之间的匹配系

数最高的实体字符串，作为推荐实体字符串，即“消化科”为推荐实体字符串。在数据库中查找推荐实体字符串“消化科”的挂号链接，将上述挂号链接作为推荐业务数据。

[0159] 如图2b中的界面20x所示，当用户终端10a在确定意图类型以及查找推荐业务数据的过程中可以在屏幕上播放预设的动画。如图2b中的界面20y所示，当用户终端10a确定了推荐业务数据后，可以在屏幕上显示推荐实体字符串“消化科”，并显示推荐业务数据，即显示“消化科”的挂号链接。

[0160] 请参见图3，是本发明实施例提供的一种文本处理方法的流程示意图，如图3所示，该方法可以包括：

[0161] 步骤S101，获取目标文本，并获取知识图谱；所述知识图谱包括多个实体字符串，以及与每个实体字符串对应的业务属性字符串。

[0162] 具体的，终端设备(如上述图2a对应实施例中的用户终端10a)获取目标文本(如上述图2a对应实施例中的目标文本20b)，目标文本的获取方式可以是由用户向终端设备手动输入文本，终端设备将用户输入的文本作为目标文本；也可以是终端设备采集用户的语音，将上述语音转换为文本，并将转换后的文本称为目标文本。

[0163] 终端设备获取知识图谱(如上述图2a对应实施例中的知识图谱20e)，其中知识图谱是用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。知识图谱包括多个实体对应的实体字符串，以及每个实体字符串对应的业务属性的业务属性字符串；其中业务属性表示对应的实体所在的业务领域。

[0164] 例如，知识图谱中的实体字符串“乏力”对应的业务属性字符串为：症状；知识图谱中的实体字符串“眼科”对应的业务属性字符串为：科室；知识图谱中的实体字符串“皮疹”对应的业务属性字符串为：疾病；知识图谱中的实体字符串“头孢克洛芬”对应的业务属性字符串为：药物；知识图谱中的实体字符串“磁共振成像(Magnetic Resonance Imaging，MRI)”对应的业务属性字符串为：检查；知识图谱中的实体字符串“张三”对应的业务属性字符串为：医生；知识图谱中的实体字符串“阑尾切除术”对应的业务属性字符串为：手术；知识图谱中的实体字符串“复诊”对应的业务属性字符串为：流程。

[0165] 请参见图4，是本发明实施例提供的一种知识图谱的示意图。以构建疾病“皮疹”的知识图谱为例，与疾病“皮疹”所关联的实体字符串可以来自于电子病历(EHR，Electronic Health Record)中所涉及的实体，例如，科室名称、年龄等，由于构建“皮疹”涉及大量的专业知识，使用EHR相比从互联网上爬取更加可靠。每两个实体字符串之间的连线表示对应的逻辑关系，也就是标识两个实体字符串各自的业务属性，图4对应的疾病“皮疹”的知识图谱中，包含了1种疾病：皮疹；4个科室：儿内科、皮肤科、儿童急诊和隔离门；2种性别：男、女；7种年龄：0岁、1岁、2岁、3岁、4岁、5岁、9岁。因此，疾病-科室逻辑关系的数量为4个，疾病-性别逻辑关系的数量为2个，疾病-年龄逻辑关系的数量为7个。

[0166] 可选的，获取输入目标文本的用户(称为登录用户)的用户信息(称为登录用户信息)，登录用户信息可以包括年龄信息和性别信息。当终端设备获取到目标文本，并检测到目标文本中也包括年龄信息或性别信息时，从目标文本中提取出该年龄信息或者性别信息，若提取出来的年龄信息或者性别信息，与登录用户信息中的年龄信息或性别信息不符，终端设备生成提示消息，该提示消息用于提示登录用户调整上述登录用户信息，或者调整目标文本中的年龄信息或性别信息。

[0167] 这是因为在医疗场景下，部分疾病的诊断在很大程度上依赖于年龄和性别，因此提前确定患者的年龄和性别可以提高后续疾病诊断的准确率，进而提高向用户推荐业务数据的准确率。

[0168] 请参见图5，是本发明实施例提供的一种文本处理的界面示意图，如界面30a所示，登录用户在智能导诊登录页设置的登录用户信息为：年龄：29岁；性别：男性。登录用户在智能导诊应答页中输入的目标文本为：“小孩子肚子不舒服”，终端设备识别到目标文本中“小孩子”的年龄信息为：0-10岁，因此目标文本中的年龄信息与登录用户信息中的年龄信息不符，进而终端设备可以生成提示消息：“当前咨询人的个人信息与输入的个人信息不符，请重新选择咨询人，或者重新输入”。登录用户重新设置咨询人的登录用户信息为：1岁、男性，后续终端设备可以在新的登录用户信息的基础上，接收新的目标文本并确定目标意图类型。

[0169] 步骤S102，在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串，并提取与所述目标实体字符串对应的目标业务属性字符串。

[0170] 具体的，终端设备从目标文本中提取出关键词(称为目标关键词，如前述图2a对应实施例中的关键词集合20c中的关键词)，提取目标关键词是可以摒弃目标文本中与用户意图无关的字符，提高后续识别用户意图的准确性。例如，目标文本为：“昨天晚上半夜开始拉肚子”，那么目标关键词就是：“拉肚子”。

[0171] 由于目标关键词是用户输入的，可能存在用词不规范或者是不专业的情况，如前述举例中的“拉肚子”，因此为了提高后续识别意图的准确性，终端设备还需要将目标关键词映射为专业的、规范的同义词，可以将映射得到的同义词称为图谱标识实体字符串(如前述图2a对应实施例中的标准实体字符串集合20d中的标准实体字符串)。终端设备在知识图谱的多个实体字符串中查找与图谱标识实体字符串相同的实体字符串，作为目标实体字符串。

[0172] 例如，知识图谱包括的实体字符串为：“腹泻”、“早饱”、“消化科”、“血常规”，图谱标识实体字符串为：“腹泻”、“张三”，那么在上述4个实体字符串中查找与图谱标识实体字符串“腹泻”、“张三”相同的实体字符串为为：“腹泻”，即是目标实体字符串就是：“腹泻”。

[0173] 下面对如何从目标文本中提取目标关键词的过程进行详细的描述：终端设备将目标文本划分为多个单位字符，均称为目标单位字符。终端设备可以通过one-hot编码方式将每个目标单位字符转换为一维的向量，称为目标单位字符向量，并获取第一循环神经网络模型，该第一循环神经网络模型可以标注出每个字符的词性标签，词性标签可以包括：B-Person(人名的开始部分)、I-Person(人名的中间部分)、B-Organization(组织机构的开始部分)、I-Organization(组织机构的中间部分)、O(非实体信息)。其中，第一循环神经网络模型中的编码层可以是基于RNN(Recurrent Neural Network，循环神经网络)，或者是LSTM训练得到的，第一循环神经网络模型中的条件随机场是由CRF训练得到的。

[0174] 首先对所有的目标单位字符向量进行正向编码，终端设备初始化隐藏状态向量h10，在t11时刻，将在目标文本中位于首位的目标单位字符向量x1、隐藏状态向量h10输入第一循环神经网络模型中的编码层，根据公式(1)计算t11时刻的隐藏状态向量h11，

[0175]
$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
h_t &= o_t \tanh(c_t)
\end{aligned}
\qquad (1)
$$

[0176] 其中，σ(·)是σ函数，i,f,o分别表示输入门、遗忘门、输出门。所有的W表示两门之间的权重矩阵。在编码过程中，整个第一循环神经网络模型中的参数是共享的，即是在计算每一个时刻的隐藏状态向量时，上述参数都不变。

[0177] 在t12时刻，将目标文本中位于第二位的目标单位字符向量x2、t11时刻的隐藏状态向量h11输入编码层，根据公式(1)再计算t12时刻的隐藏状态向量h12；在t13时刻，同样根据公式(1)计算t13时刻的隐藏状态向量h13。换句话说，t时刻的隐藏状态向量ht是由t-1时刻的隐藏状态向量h(t-1)和t时刻的目标单位字符向量xt决定的，不断地迭代，直至最后一次迭代得到隐藏状态向量h1n，将隐藏状态向量h11、h12、...、h1n组合为正向编码矩阵，正向编码矩阵的尺寸可以表示为：n×m，其中n表示目标单位字符的数量，m表示每个目标单位字符正向编码后的隐藏状态向量的维数。

[0178] 然后对所有的目标单位字符向量进行反向编码，终端设备初始化隐藏状态向量h20，在t21时刻，将目标文本中位于末位的目标单位字符串向量xn、隐藏状态向量h20输入第一循环神经网络模型中的编码层，根据公式(1)计算t21时刻的隐藏状态向量h21；在t22时刻，将目标文本中位于倒数第二位的目标单位字符串向量x(n-1)、t21时刻的隐藏状态向量h21输入编码层，根据公式(1)计算t22时刻的隐藏状态向量h22。和正向循码相同，t时刻的隐藏状态向量ht是由t-1时刻的隐藏状态向量h(t-1)和t时刻的目标单位字符向量x(n+1-t)决定的，不断地迭代，直至最后一次迭代得到隐藏状态向量h2n，将隐藏状态向量h21、h22、...、h2n组合为反向编码矩阵。反向编码矩阵的尺寸可以表示为：n×m，其中n表示目标单位字符的数量，m表示每个目标单位字符反向编码后的隐藏状态向量的维数。

[0179] 正向编码和反向循环编码的过程除了输入目标单位字符向量的顺序不同(正向编码是从前往后输入，反向编码是从后往前输入)，其余的过程都相同，这是为了保证即使目标文本中包含的目标单位字符非常多，目标单位字符的语义信息和时序信息也不会随着序列而消减。

[0180] 终端设备将正向编码矩阵与反向编码矩阵拼接为隐藏状态矩阵，可以知道隐藏状态矩阵的尺寸为n×2m。基于第一循环神经网络中的条件随机场，对隐藏状态矩阵中的每个行分别进行序列标注，即是识别每一行与多种词性标签之间的相似度，将相似度最大的词性标签作为对应目标单位字符的词性标签。

[0181] 终端设备可以将词性标签为“B-Person”到“I-Person”之间的字符序列作为关键词，将词性标签为“B-Organization”到“I-Organization”之间的字符序列作为关键词，词性标签为“O”的字符是非实体字符，可以直接忽略。

[0182] 下面对终端设备如何将一个目标关键词映射为一个图谱标注实体字符串进行具体的说明，若存在多个目标关键词可以基于相同的处理过程确定与之对应的图谱标注实体字符串。

[0183] 终端设备将目标关键词划分为多个单位字符，称为关键单位字符，终端设备可以

通过one-hot编码方式将每个关键单位字符转换为一维的向量，称为关键单位字符向量，并获取第二循环神经网络模型，该第二循环神经网络模型可以将目标关键词翻译为规范化的图谱标注实体字符串。其中，第二循环神经网络模型包括编码层和解码层，且编码层和解码层都是由RNN或者LSTM训练得到的。

[0184] 按照每个关键单位字符向量在目标关键词中的前后位置关系，第二循环神经网络模型中的编码层对每个关键单位字符向量依次进行编码，编码的过程可以参数上述前向编码的过程。终端设备可以将位于目标关键词末尾的关键单位字符向量确定的隐藏状态向量，作为该目标关键词的上下文向量，即是将编码层最后一次运算得到的隐藏状态向量作为上下文向量。基于第二循环神经网络模型的解码层，对上述上下文向量进行解码，解码过程中，t时刻的输出y_t是由t-1时刻的隐藏状态向量h(t-1)以及t-1时刻的输出y_{t-1}决定的，终端设备将解码后的字符序列作为图谱标注实体字符串。

[0185] 将目标关键词转换为图谱标注实体字符串的过程就类似于翻译模型将中文字符序列翻译为同义的英文字符序列。

[0186] 步骤S103，根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型。

[0187] 具体的，终端设备从实体词袋中查找与目标实体字符串对应的独热码(one-hot code)，称为第一向量。其中，实体词袋中包括知识图谱中的多个实体字符串，以及每个实体字符串对应到的独热码，独热码是向量中只包含一个1，其余均为0的向量。

[0188] 举例来说，知识图谱中包括4个实体字符串，分别为：消化科，腹胀，早饱以及张三；那么实体词袋中，实体字符串“消化科”的独热码为：[1,0,0,0]；实体字符串“腹胀”的独热码为：[0,1,0,0]；实体字符串“早饱”的独热码为：[0,0,1,0]；实体字符串“张三”的独热码为：[0,0,0,1]。可以知道，知识图谱中所包含的实体字符串的数量就等于每一个独热码的维数。

[0189] 终端设备获取实体词向量转换模型，该实体词向量转换模型可以将高维度的第一向量，降低为低维度的词向量，基于实体词向量转换模型中的隐藏层对应的隐藏矩阵，将第一向量与该隐藏矩阵进行矩阵的乘法运算，乘法运算后得到向量称为目标实体词向量(如图2a对应实施例中的实体词向量20g)，其中，隐藏矩阵的行数等于第一向量的维数，隐藏矩阵的列数等于目标实体词向量的维数。例如，第一向量的尺寸为：1×1000，隐藏矩阵的尺寸为：1000×100，那么目标实体词向量的尺寸为：1×100。

[0190] 实体词向量转换模型可以是根据fasttext(快速文本分类器)、word2vec(词向量转换模型)的训练得到的，训练过程中还需要在隐藏层后加一个分类器，用于计算分类误差，并将该分类误差通过前向传播传递至隐藏层，用于更新隐藏层中隐藏矩阵的参数值。

[0191] 若存在多个目标实体字符串，可以基于实体词袋以及实体词向量转换模型将每个目标实体字符串分别转换为与之对应的目标实体词向量，t个目标实体词向量可以表示为：{x-1,x-2...x-t}，其中x-n表示第n个尺寸为1×p的目标实体词向量，1≤n≤t，其中p可以等于256。

[0192] 终端设备从业务属性词袋中查找与目标业务属性字符串对应的独热码(one-hot code)，称为第二向量。其中，业务属性词袋中包括知识图谱对应的多个业务属性字符串，以及每个业务属性字符串对应到的独热码。

[0193] 终端设备获取业务属性词向量转换模型，该业务属性词向量转换模型可以将高维度的第二向量，降低为低维度的词向量，基于业务属性词向量转换模型中的隐藏层对应的隐藏矩阵，将第二向量与该隐藏矩阵进行矩阵的乘法运算，乘法运算后得到向量称为目标业务属性词向量(如图2a对应实施例中的属性词向量20h)。

[0194] 业务属性词向量转换模型同样可以是根据fasttext、word2vec的训练得到的，训练过程中还需要在隐藏层后加一个分类器，用于计算分类误差，并将该分类误差通过前向传播传递至隐藏层，用于更新隐藏层中隐藏矩阵的参数值。

[0195] 若存在多个目标业务属性字符串，可以基于业务属性词袋以及业务属性词向量转换模型将每个目标业务属性字符串分别转换为与之对应的目标业务属性词向量，t个目标业务属性词向量可以表示为：{g-1,g-2...g-t}，其中g-n表示第n个尺寸为1×p的目标业务属性词向量，1≤n≤t，其中p可以等于256。

[0196] 终端设备将目标实体词向量与目标业务属性词向量组合为输入向量，t个目标实体字符串对应的输入向量可以表示为：{d-1,d-2...d-t}，其中d-n表示第n个尺寸为1×2p的输入向量，1≤n≤t。

[0197] 终端设备获取意图识别模型(如图2a对应实施例中的意图识别模型20k)，该意图识别模型可以识别输入向量对应的意图类型，意图识别模型包括卷积层、池化层以及分类器。卷积层对应1个或者多个卷积核(kernel，也可以称为滤波器，或者称为感受野)，卷积运算是指卷积核与位于输入向量不同位置的子矩阵进行矩阵乘法运算，卷积运算后的输出矩阵的行数H_{out}和列数W_{out}是由输入向量的尺寸、卷积核的尺寸、步长(stride)以及边界填充(padding)共同决定的，即$H_{out}=(H_{in}-H_{kernel}+2*padding)/stride+1$，$W_{out}=(W_{in}-W_{kernel}+2*padding)/stride+1$。$H_{in}$，$H_{kernel}$分别表示输向量的数量和卷积核的行数；$W_{in}$，$W_{kernel}$分别表示每个输入向量的维数和卷积核的列数。

[0198] 卷积后，还需基于池化层对输出矩阵进行池化运算，池化运算是指对提取出来的输出矩阵进行聚合统计，池化运算可以包括平均池化运算和最大池化运算。平均池化运算方法是在输出矩阵的每一个行(或者是列)中计算出一个平均数值代表该行(或者列)；最大池化运算是在输出矩阵的每一个行(或者是列)中提取出最大数值代表该行(或者列)。通过卷积运算和池化运算，可以提取出目标文本最显著的意图特征，该意图特征可以称为意图特征向量。

[0199] 基于意图分类模型中的分类器识别意图特征向量与分类器中多种意图类型之间的匹配概率，从多个匹配概率中提取出最大匹配概率对应的意图类型，作为目标意图类型。

[0200] 若当前业务场景是医疗导诊场景，意图类型可以包括：找医生、找科室、疾病诊断、检验检查、流程相关以及其他。其中，找医生、找科室以及检验检查等都是用户已经知道具体的业务对象，例如，用户已经知道找“张三”医生，用户已经知道挂“消化科”的号，用户已经知道要去做“血常规”检查等；对应地，疾病诊断是需要由终端设备根据用户的意图进行推理，确定业务对象。

[0201] 步骤S104，从所述目标实体字符串中确定与所述目标意图类型相关联的目标意图字符串，并根据所述目标意图字符串生成推荐业务数据。

[0202] 具体的，从目标实体字符串中提取与目标意图类型相关联的实体字符串，作为目标意图字符串。终端设备查找与目标意图字符串对应的业务数据，作为推荐业务数据。

[0203] 通过理解用户的意图(即是目标意图类型),可以将具有不同意图的用户数据流分发至不同的业务处理模块,该业务处理模型可以根据目标意图字符串查找对应的推荐业务数据,以提高推荐业务数据的推荐正确率以及效率,同时通过识别用户的意图,将用户数据流分流,这种分流方式使得整个系统具有合理的层级架构。

[0204] 例如,目标文本为:“我肚子痛,还犯恶心,想找张三医生”,从目标文本中确定出目标实体字符串为:“腹痛”、“恶心”、“张三”,且识别出目标文本的目标意图类型为:找医生。因此,在上述3个目标实体字符串中,与目标意图类型“找医生”关联的目标意图字符串为:张三。终端设备可以在数据库中查找“张三”医生的值班表、个人简介以及挂号链接等,并将上述数据封装为推荐业务数据,显示在屏幕上。在数据库中还可以将每个医生擅长的病种作为标签存储在数据库中,当查找到对应医生时,可以将该医生擅长的病种一并封装为推荐业务数据,这样可以向用户反馈最佳的个性化回答。

[0205] 请一并参见图6,本发明实施例提供的一种确定推荐业务数据的场景示意图,如图6所示,在智能导诊应答页中,用户输入的目标文本为:“我要找张三”,终端设备获取到上述目标文本后,提取出目标实体字符串为:“张三”,且确定目标业务属性字符串为:“医生”,根据上述目标实体字符串以及目标业务属性字符串确定目标意图类型为:找医生,以及目标意图字符串为:“张三”,进而在数据库中查找“张三”医生的排班表以及“张三”医生的挂号链接,并将上述数据封装为推荐业务数据,向用户展示该推荐业务数据。

[0206] 请一并参见图7,本发明实施例提供的另一种确定推荐业务数据的场景示意图,如图7所示,在智能导诊应答页中,用户输入的目标文本为:“骨科”,终端设备获取到上述目标文本后,确定出目标实体字符串为:“创伤骨科”,且确定目标业务属性字符串为:“科室”,根据上述目标实体字符串以及目标业务属性字符串确定目标意图类型为:找科室,以及目标意图字符串:“创伤骨科”,进而在数据库中查找“创伤骨科”的科室简介以及“创伤骨科”的挂号链接,并将上述数据封装为推荐业务数据,向用户展示该推荐业务数据。后续还可以继续向用户提供推荐医生功能,当用户选择推荐此科室的医生时,可以在知识图谱中查找属于创伤骨科室的医生,并按照查找到的医生的就诊画像生成推荐医生列表,并将该列表显示在屏幕上,以供用户选择。

[0207] 请参见图8,是本发明实施例提供的另一种文本处理方法的流程示意图,用户输入的目标文本40a为:“最近肚子鼓鼓的,不舒服”,通过实体识别算法,识别目标文本40a中的目标关键词40b:“肚子鼓鼓”,提取目标关键词的具体过程可以参见上述图3对应实施例中的步骤S102。通过实体链接,将目标关键词40b映射为目标实体字符串40c:“腹胀”,在知识图谱中查找目标实体字符串40c的目标业务属性字符串:“症状”,确定目标实体字符串以及目标业务属性字符串的具体过程可以参见上述图3对应实施例中的步骤S102。通过意图识别模型,识别目标文本的目标意图类型,意图类型可以包括:找医生、找科室、疾病诊断、检验检查、流程相关以及其他。当目标意图类型是疾病诊断类型时,终端设备根据医学知识图谱、目标实体字符串40c:“腹胀”进行推理,确定与目标文本40a匹配的推荐业务数据,其中推荐业务数据可以是:疾病、科室、药物、症状等。

[0208] 上述可知,用户只需要输入文本,终端就可以识别出用户的意图,并生成满足用户意图的业务数据,相比通过手动查找资料并人工执行多次业务行为以确定业务数据,本发明所采用的自动化的方式生成满足用户心理预期的业务数据,可以节约用户执行多次业务

行为所耗费的时间，进而可以提高获取业务数据的效率。

[0209] 请参见图9，是本发明实施例提供的另一个文本识别方法的流程示意图，文本识别方法可以包括如下步骤S201-步骤S206：

[0210] 步骤S201，获取目标文本，并获取知识图谱；所述知识图谱包括多个实体字符串，以及与每个实体字符串对应的业务属性字符串。

[0211] 步骤S202，在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串，并提取与所述目标实体字符串对应的目标业务属性字符串。

[0212] 步骤S203，根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型。

[0213] 其中，步骤S201-步骤S203的具体过程可以参见上述图3对应实施例中的步骤S101-步骤S103，此处不再详述。

[0214] 步骤S204，获取所述目标意图类型所表征的业务属性，确定为意图业务属性。

[0215] 具体的，获取目标意图类型所标识的业务属性，并将获取的业务属性称为意图业务属性，例如，目标意图类型“找医生”所标识的业务属性为：医生；目标意图类型“找科室”所标识的业务属性为：科室；目标意图类型“检验检查”所标识的业务属性为：检查；目标意图类型“流程相关”所标识的业务属性为：流程；目标意图类型“疾病判断”所标识的业务属性为：症状。

[0216] 步骤S205，将具有所述意图业务属性的目标实体字符串作为所述目标意图字符串。

[0217] 终端设备将具有意图业务属性的目标实体字符串作为目标意图字符串，这样进一步筛选出与用户意图最相关的字符串，后续基于该字符串查找到的业务数据是符合用户的心里预期的业务数据，或者说是符合用户意图的业务数据。

[0218] 步骤S206，查找与所述目标意图字符串相关联的业务数据，并将与所述目标意图字符串相关联的业务数据作为所述推荐业务数据。

[0219] 具体的，当目标意图类型是语义确定类型时，在数据库中查找与目标意图字符串相关联的业务数据，将查找的业务数据作为推荐业务数据。属于语义确定类型的目标意图类型是指用户已经知道具体的业务对象，例如，找医生、找科室、检验检查、流程相关等均属于语义确定类型。

[0220] 例如，当目标意图字符串为“消化科”时，可以在数据库中查找消化科挂号链接，以及属于消化科医生的简介，并将上述数据封装为推荐业务数据；当目标意图字符串为“血常规”是，可以在数据库中查找“血常规”检查的注意事项，以及“血常规”检查的预约链接等，并将上述数据封装为推荐业务数据；当目标意图字符串为“复诊”时，可以在数据库中查找复诊的流程简介，以及复诊的注意事项，并将上述数据封装为推荐业务数据。

[0221] 当目标意图类型是语义推理类型时，终端设备可以将知识图谱中的所有实体字符串划分为多个第一实体字符串，以及多个第二实体字符串，其中不具有意图业务属性的实体字符串称为第一实体字符串，具有意图业务属性的实体字符串称为第二实体字符串。属于语义推理类型的目标意图类型是指需要由终端设备根据当前的业务场景以及目标文本进一步推理出业务对象。

[0222] 例如，目标意图类型“疾病判断”属于语义推理类型，且上述目标意图类型所标识

的业务属性为:症状,那么知识图谱对应的多个实体字符串,可以将不具有“症状”业务属性的实体字符串作为第一实体字符串,对应地将具有“症状”业务属性的实体字符串作为第二实体字符串。

[0223] 在知识图谱中,确定目标意图字符串与每个第一实体字符串在知识图谱中的相似系数(称为目标相似系数),由于目标意图字符串也是属于知识图谱的,也即是是确定目标意图字符串在知识图谱中的节点,与第一实体字符串在知识图谱中的节点之间的相关性。

[0224] 从多个目标相似系数中,终端设备可以将目标相似系数大于系数阈值的第一实体字符串,作为与目标意图类型匹配的目标第一实体字符串;或者是从多个目标相似系数中,将具有最大目标相似系数的第一实体字符作为目标第一实体字符串;或者是根据当前不同的业务场景以及多个目标相似系数,从多个第一实体字符串中选择与目标意图类型匹配的目标第一实体字符串;同时还可以将登录用户的登录用户信息作为辅助条件,按照上述方式一并确定目标第一实体字符串;例如,若第一实体字符串包括:“成人湿疹”和“婴儿湿疹”,根据登录用户信息中的年龄信息:“1岁”,从上述第一实体字符串中选择与登录用户信息相符合的目标实体字符串:“婴儿湿疹”。

[0225] 例如,若当前的业务场景是疾病问询系统,且意图业务属性为:症状,那么可以在业务属性为:“疾病”对应的第一实体字符串的目标相似系数中,选择具有最高目标相似系数的第一实体字符串作为目标第一实体字符串;若当前的业务场景是药店药物推荐系统,那么可以在业务属性为:“药物”对应的第一实体字符串的目标相似系数中,选择具有最高目标相似系数的第一实体字符串作为目标第一实体字符串。

[0226] 请参见图10,是本发明实施例提供的另一种知识图谱的示意图,图10是以症状“面部皮疹”、科室“皮肤科”、科室“儿科”、疾病“婴儿湿疹”、疾病“婴儿痤疮”、疾病“湿疹”以及疾病“皮疹”所生成的知识图谱,且该知识图谱中还包括症状“面部皮疹”与2个科室之间的关联匹配系数以及,症状“面部皮疹”与4个疾病之间的关联匹配系数。

[0227] 若目标意图类型为:疾病诊断,该目标语义类型属于语义推理类型,且目标意图字符串为“面部皮疹”,那么图10的知识图谱中的“皮肤科”、“儿科”、“婴儿湿疹”、“婴儿痤疮”、“湿疹”以及“皮疹”均属于第一实体字符串,可以将关联匹配系数作为目标意图字符串“面部皮疹”与对应的第一实体字符串之间的目标相似系数。

[0228] 可以从上述6个目标相似系数中,将目标相似系数大于系数阈值0.4的第一实体字符串“皮肤科”和第一实体字符串“湿疹”作为目标第一实体字符串;也可以从上述6个目标相似系数中,选择具有最大目标相似系数的第一实体字符串“皮肤科”作为目标第一实体字符串;若当前场景是疾病问询系统,还可以在业务属性为:“疾病”对应的第一实体字符串的目标相似系数中,选择具有最高目标相似系数的第一实体字符串“湿疹”作为目标第一实体字符串;若当前场景是科室推荐系统,还可以在业务属性为:“科室”对应的第一实体字符串的目标相似系数中,选择具有最高目标相似系数的第一实体字符串“皮肤科”作为目标第一实体字符串。

[0229] 确定知识图谱中目标意图字符串与第一实体字符串之间的目标相似系数有两种方式,下面其中一种方式进行详细的说明:

[0230] 知识图谱还包括多个实体字符串之间的关联匹配系数,其中,关联匹配系数是通过与知识图谱关联的数据集合统计得到。从多个第一实体字符串中任意提取一个实体字符

串作为测试第一实体字符串，将目标意图字符串与测试第一实体字符串之间的关联匹配系数，作为目标意图字符串与测试第一实体字符串之间的目标相似系数，若目标意图字符串的数量不止一个，那么就将每个目标意图字符串与测试第一实体字符串之前的关联匹配系数之和，作为目标意图字符串与测试第一实体字符串之间的目标相似系数。再从剩余的第一实体字符串中，再提取测试第一实体字符串，再确定目标意图字符串与新的测试第一实体字符串之间的目标相似系数，不断循环，当每个第一实体字符串均被确定为测试第一实体字符串时，停止轮询，即确定了目标意图字符串与每个第一实体字符串之间的目标相似系数。

[0231] 下述以医疗导诊为例，如何确定知识图谱中多个实体字符串之间的关联匹配系数进行说明，知识图谱可以是基于大量的电子病历(EHR，Electronic Health Record)构建的，首先从大量的RHR中筛选出满足数据要求的条目，例如，每个数据条目必须包含患者性别、年龄、治疗科室、症状、用药、医生、疾病等；然后对不同药物名称对应同一种药物的条目进行合并，合并为统一的药物名称；对不同科室名称对应同一个科室的条目进行合并，合并为统一的科室名称。以症状、药物、医生、疾病为中心，计算中心与其余的列之间的共现概率。

[0232] 以疾病-性别为例，设疾病的集合为{d_1,d_2...d_n}，性别的集合为{s1,s2}，以T为阈值，根据公式(2)计算疾病与性别之间的共现概率：

[0233]
$$r_{pop}(d_i,s_j)=\begin{cases}count(d_i\cap s_j),count(d_i\cap s_j)>T\\0,else\end{cases}$$
$$r_{rat}(d_i,s_j)=\frac{r_{pop}(d_i,s_j)}{count(d_i)},i=1,2...,n;j=1,2 \qquad (2)$$

[0234] 也就是说，对于任意一个疾病和性别，计算它们同时出现的次数，若大于一定阈值T，则保留共现次数并记为r_{pop}。用共现次数除以中心向量的出现次数，即为共现概率r_{rat}。

[0235] 同理，可以计算疾病-性别、疾病-症状、症状-科室、医生-疾病、疾病-科室等共现对的共现系数和共现概率。

[0236] 终端设备可以将EHR中的实体对应的字符串作为知识图谱中的实体字符串，并将任意两个实体之间的共现概率，作为对应实体字符串之间的关联匹配系数；通过统计学的方式，确定知识图谱中任意两个节点(实体字符串)之间的相关性，用于后续在知识图谱上的推理提供具有科学性的数据支持。

[0237] 下面对如何确定知识图谱中目标意图字符串与第一实体字符串之间的目标相似系数的另一种方式进行详细的说明：该方式是通过随机游走+skipgram(语言模型)生成每个实体字符串的图向量。知识图谱中不具有意图业务属性的实体字符串是第一实体字符串，具有意图业务属性的实体字符串是第二实体字符串。获取知识图谱中每个实体字符串的图向量，可以将第一实体字符串的图向量称为第一图向量，将第二实体字符串的图向量称为第二图像了，并获取第二实体字符串的词向量，其中获取词向量的具体过程可以参见上述图3对应实施例中的步骤S103中对获取目标实体词向量、以及目业务属性词向量的描述。

[0238] 从多个第二词向量中提取目标意图字符串的第二词向量，作为目标词向量，可以知道，目标意图字符串是属于目标实体字符串的，对应的目标词向量也是属于前述中目标

实体词向量的。从多个第二图向量中提取目标意图字符串的第二图向量，作为目标图向量。从多个第一实体字符串中提取一个用于当前轮询第一实体字符串，作为测试第一实体字符串，并将测试第一实体字符串对应的第一图向量，作为测试图向量。

[0239] 根据公式(3)计算目标意图字符串与测试第一实体字符串之间的第一相似系数score1(d_t)：

[0240] $$score1(d_t) = w_{ti} * \sum_{i=1}^{n} \max(sim(T_{tj}, T_i)) \quad (3)$$

[0241] 其中，n表示目标意图字符串的数量；d_t表示测试第一实体字符串；T_i表示目标词向量；T_{tj}表示在知识图谱中与测试第一实体字符串相邻的第二实体字符串的第二词向量，0≤j≤m，m表示与测试第一实体字符串相邻的第二实体字符串的数量；w_{ti}表示目标意图字符串与测试第一实体字符串之间的关联匹配系数；sim()表示余弦相似度。

[0242] 根据公式(4)计算目标意图字符串与测试第一实体字符串之间的第二相似系数score2(d_t)：

[0243] $$score2(d_t) = w_{ti} * \sum_{i=1}^{n} sim(G_i, G(d_t)) \quad (4)$$

[0244] 其中，n表示目标意图字符串的数量；G_i表示目标图向量；G(d_t)表示测试第一实体字符串的图向量(即是测试图向量)；w_{ti}表示目标意图字符串与测试第一实体字符串之间的关联匹配系数；sim()表示余弦相似度。

[0245] 将第一相似系数score1(d_t)和第二相似系数score2(d_t)之和，作为目标意图字符串与测试第一实体字符串之间的目标相似系数：

[0246] score(d_t)＝score1(d_t)+score2(d_t) (5)

[0247] 再从剩余的第一实体字符串中，提取测试第一实体字符串，再根据公式(3)、公式(4)、公式(5)确定目标意图字符串与新的测试第一实体字符串之间的目标相似系数，不断循环，当每个第一实体字符串均被确定为测试第一实体字符串时，停止轮询，即确定了目标意图字符串与每个第一实体字符串之间的目标相似系数。

[0248] 公式(3)可以看作是计算所有用户输入的症状与所有疾病对应的症状、科室对应的症状、药物对应的症状、手术对应的症状的相似性。公式(4)可以看作是通过知识图谱得到的，所有输入症状与某一疾病、科室、药物、手术等的相似性。前一部分通过输入的症状与疾病对应的症状、科室对应的症状、药物对应的症状、手术对应的症状进行比对，来考察症状和疾病、疾病、科室、药物、手术等的相关性；后一部分利用知识图谱，直接计算症状和疾病、疾病、科室、药物、手术等的相似性。两种方法相辅相成，既充分利用了知识图谱的数据，又增强了系统的灵活性。

[0249] 下面对如何确定知识图谱中每个实体字符串的图向量进行详细说明：知识图谱可以看作是一种拓扑图(graph)，知识图谱中的每个实体字符串即是拓扑图的顶点；若两个实体字符串之间的关联匹配系数大于0，那么上述两个实体字符串之间存在连接边，若两个实体字符串之间的关联匹配系数等于0，那么上述两个实体字符串之间不存在连接边。初始化知识图谱中每个实体字符串的图向量，作为原始图向量，采用随机游走的方式，从知识图谱所包含的所有实体字符串中，采样多个实体字符串，作为样本实体字符串，并获取每个样本实体字符串的原始图向量，分别作为样本图向量。

[0250] 根据公式(6)确定损失值：

[0251] $$P(\{v_{i-w},\ldots,v_{i+w}\}\mid v_i)=\prod_{j=i-w}^{i+w}P(\phi(v_j)\mid v_i) \qquad (6)$$

[0252] 其中，φ(v_i)表示顶点i的图向量表示函数，P(v_j|v_i)表示第j个实体字符串的与第i个实体字符串在训练语料中同时出现的概率。通过梯度下降规则调整图向量表示函数φ，最终根据新的图向量表示函数可得到每个样本实体字符串新的图向量，称为调整图向量，并将样本实体字符串的原始图向量替换为调整图向量。再从所有的实体字符串中进行采样，再根据公式(6)以及梯度下降规则确定新的调整图向量，在将新的调整图向量作为原始图向量，不断地更新每个实体字符串对应的原始图向量，当采样次数达到次数阈值时，将最后一次确定的所有实体字符串的原始图向量，作为每个实体字符串的图向量。

[0253] 请参见图11，是本发明实施例提供的一种挂号转化率对比图，从图11可以看出，在12月1号-12月9号期间，识别目标意图类型后再确定业务推荐数据的挂号转化率处于4.10%-6.30%之间，不识别目标意图类型而直接根据目标文本确定业务推荐数据的挂号转化率处于3.00%-4.10%之间；因此，识别目标意图类型后，挂号转换率提升1%以上，性能提升20%以上。

[0254] 请参见图12，是本发明实施例提供的一种文本处理的系统结构图。当文本处理的场景处于医疗场景下时，文本识别的系统分为：数据、算法支持和应用。数据主要来源于患者的口述、权威知识和EHR三部分组成。患者口述的作用是生成用于训练第一循环神经网络模型、第二循环神经网络模型以及意图识别模型的语料；EHR的作用是为自动快速构建知识图谱提供数据来源；权威疾病知识的作用是对知识图谱、系统的结果进行调整，提出指导性的意见，来自于专业人士与书籍。

[0255] 算法支持部分由关键词提取模块、知识图谱模块、意图识别模块、匹配模块组成。当系统检测到用户的输入目标文本时，首先通过关键词提取模块中的中文分词、实体识别、词性标注、实体修饰、句法解析和实体关系共同确定目标文本中的目标实体字符串；同时还可以根据目标文本中的年龄信息和性别信息可以判断输入目标文本中的登录用户是否是患者本人。

[0256] 再通过知识图谱模块提取出与目标实体字符串对应的目标业务属性字符串；每个目标实体字符串都是存在与之对应的业务属性的，例如，业务属性字符串为：“疾病”、“症状”、“检查”、“治疗”等；最后将目标实体字符串与目标业务属性字符串输入意图识别模块进行意图类型分类，确定目标意图类型。得到目标意图类型后，再利用目标意图类型以及目标实体字符串确定最佳的推荐业务数据，并输出推荐业务数据。例如，可以根据目标意图类型以及目标文本进行疾病的诊断，诊断出来的疾病就可以是推荐业务数据，同时根据诊断到的疾病与目标实体字符串之间的目标相似系数，确定此次疾病诊断的分类是粗糙诊断、中等程度诊断或者是精细诊断。

[0257] 在应用层面，不仅考虑到对用户的意图理解和对疾病分类体系的借鉴，还对医生的就诊情况进行了分析，得到每个医生的就诊画像，后续当需要向用户推荐医生(即生成推荐业务数据)时，可以根据用户的意图、疾病分类体系以及医生的就诊画像生成推荐医生列表，以向用户推荐合适的医生，同时还可以将每个医生擅长病种作为标签嵌入数据库中，在确定推荐业务数据时，为用户的问题返回最佳的个性化回答。

[0258] 上述可知，用户只需要输入文本，终端就可以识别出用户的意图，并生成满足用户意图的业务数据，相比通过手动查找资料并人工执行多次业务行为以确定业务数据，本发明所采用的自动化的方式生成满足用户心理预期的业务数据，可以节约用户执行多次业务行为所耗费的时间，进而可以提高获取业务数据的效率。

[0259] 进一步的，请参见图13，是本发明实施例提供的一种文本处理装置的结构示意图。如图13所示，文本处理装置1可以应用于上述图3-图12对应实施例中的终端设备，文本处理装置1可以包括：获取模块11、查找模块12、识别模块13、生成模块14。

[0260] 获取模块11，用于获取目标文本，并获取知识图谱；所述知识图谱包括多个实体字符串，以及与每个实体字符串对应的业务属性字符串；

[0261] 查找模块12，用于在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串；

[0262] 所述获取模块11，还用于提取与所述目标实体字符串对应的目标业务属性字符串；

[0263] 识别模块13，用于根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型；

[0264] 生成模块14，用于从所述目标实体字符串中确定与所述目标意图类型相关联的目标意图字符串，并根据所述目标意图字符串生成推荐业务数据。

[0265] 其中，获取模块11、查找模块12、识别模块13、生成模块14的具体功能实现方式可以参见上述图3对应实施例中的步骤S101-步骤S104，这里不再进行赘述。

[0266] 请参见图13，查找模块12可以包括：提取单元121、映射单元122、查找单元123。

[0267] 提取单元121，用于从所述目标文本中提取目标关键词；

[0268] 映射单元122，用于将所述目标关键词映射为图谱标注实体字符串；

[0269] 查找单元123，用于从所述多个实体字符串中查找与所述图谱标注实体字符串相同的实体字符串，作为与所述目标文本匹配的目标实体字符串。

[0270] 其中，提取单元121、映射单元122、查找单元123的具体功能实现方式可以参见上述图3对应实施例中的步骤S102，这里不再进行赘述。

[0271] 请参见图13，提取单元121可以包括：划分子单元1211、编码子单元1212、拼接子单元1213。

[0272] 划分子单元1211，用于将所述目标文本划分为多个目标单位字符，并将每个目标单位字符转换为目标单位字符向量；

[0273] 编码子单元1212，用于基于第一循环神经网络模型中的编码层，将多个目标单位字符向量进行双向循环编码，得到正向编码矩阵和反向编码矩阵；

[0274] 拼接子单元1213，用于将所述正向编码矩阵和所述反向编码矩阵拼接为隐藏状态矩阵；

[0275] 所述拼接子单元1213，还用于基于所述第一循环神经网络中的条件随机场，对所述隐藏状态矩阵进行序列标注，确定所述每个目标单位字符分别对应的词性标签；

[0276] 所述拼接子单元1213，还用于根据所述每个目标单位字符分别对应的词性标签，确定所述目标关键词。

[0277] 其中，划分子单元1211、编码子单元1212、拼接子单元1213的具体功能实现方式可

以参见上述图3对应实施例中的步骤S102,这里不再进行赘述。

[0278] 请参见图13,映射单元122可以包括:转换子单元1221、解码子单元1222。

[0279] 转换子单元1221,用于将所述目标关键词划分为多个关键单位字符,并将每个关键单位字符转换为关键单位字符向量;

[0280] 所述转换子单元1221,还用于基于第二循环神经网络模型中的编码层,将多个关键单位字符向量进行编码,得到所述目标关键词的上下文向量;

[0281] 解码子单元1222,用于基于所述第二循环神经网络模型中的解码层,将所述上下文向量进行解码,得到所述上下文向量的隐藏状态向量;

[0282] 所述解码子单元1222,还用于识别所述隐藏状态向量,得到与所述隐藏状态向量对应的字符序列,并将所述字符序列确定为所述图谱标注实体字符串。

[0283] 其中,转换子单元1221、解码子单元1222的具体功能实现方式可以参见上述图3对应实施例中的步骤S102,这里不再进行赘述。

[0284] 请参见图13,识别模块13:转换单元131、组合单元132、识别单元133。

[0285] 转换单元131,用于将所述目标实体字符串转换为目标实体词向量;

[0286] 组合单元132,用于将所述目标业务属性字符串转换为目标业务属性词向量;

[0287] 所述组合单元132,还用于将所述目标实体词向量和所述目标业务属性词向量组合为输入向量,基于意图识别模型中的卷积层和池化层,对所述输入向量进行卷积和池化,得到所述目标文本的意图特征向量;

[0288] 识别单元133,用于基于所述意图识别模型中的分类器,识别所述意图特征向量与所述分类器中多种意图类型之间的匹配概率,从多个匹配概率中将最大匹配概率对应的意图类型确定为所述目标意图类型。

[0289] 其中,转换单元131、组合单元132、识别单元133的具体功能实现方式可以参见上述图3对应实施例中的步骤S103,这里不再进行赘述。

[0290] 请参见图13,转换单元131可以包括:查找子单元1311、降维子单元1312。

[0291] 查找子单元1311,用于从实体词袋中查找与所述目标实体字符串对应的独热码,作为第一向量;所述实体词袋包括所述知识图谱中的所述多个实体字符串,以及与所述每个实体字符串分别对应的独热码;

[0292] 降维子单元1312,用于基于实体词向量转换模型中的隐藏层,将所述第一向量进行降维,得到所述目标实体词向量。

[0293] 其中,查找子单元1311、降维子单元1312的具体功能实现方式可以参见上述图3对应实施例中的步骤S103。

[0294] 请参见图13,生成模块14可以包括:获取单元141、确定单元142、生成单元143。

[0295] 获取单元141,用于获取所述目标意图类型所表征的业务属性,作为意图业务属性;

[0296] 确定单元142,用于将具有所述意图业务属性的目标实体字符串作为所述目标意图字符串;

[0297] 生成单元143,用于查找与所述目标意图字符串相关联的业务数据,并将与所述目标意图字符串相关联的业务数据作为所述推荐业务数据。

[0298] 其中,获取单元141、确定单元142、生成单元143的具体功能实现方式可以参见上

述图9对应实施例中的步骤S204-步骤S206,这里不再进行赘述。

[0299] 请参见图13,生成单元143可以包括:第一提取子单元1431、选择子单元1432、确定子单元1433。

[0300] 第一提取子单元1431,用于当所述目标意图类型属于语义推理类型时,从所述知识图谱的多个实体字符串中提取多个第一实体字符串;所述第一实体字符串是在所述知识图谱的多个实体字符串中除具有所述意图业务属性的实体字符串以外的实体字符串;

[0301] 选择子单元1432,用于在所述知识图谱中,确定所述目标意图字符串与每个第一实体字符串之间的目标相似系数;

[0302] 所述第一提取子单元1431,还用于根据多个目标相似系数,从所述多个第一实体字符串中选择与所述目标意图类型匹配的目标第一实体字符串;

[0303] 确定子单元1433,用于查找与所述目标第一实体字符串相关联的业务数据,并将与所述目标第一实体字符串相关联的业务数据作为所述推荐业务数据。

[0304] 其中,第一提取子单元1431、选择子单元1432、确定子单元1433的具体功能实现方式可以参见上述图9对应实施例中的步骤S206,这里不再进行赘述。

[0305] 请参见图13,知识图谱还包括多个实体字符串之间的关联匹配系数;所述关联匹配系数是通过与所述知识图谱关联的数据集合统计得到;

[0306] 选择子单元1432可以包括:开始子单元14321、停止子单元14322。

[0307] 开始子单元14321,用于从所述多个第一实体字符串中提取用于轮询的测试第一实体字符串;

[0308] 所述开始子单元14321,还用于将所述目标意图字符串与所述测试第一实体字符串之间的关联匹配系数,作为所述目标意图字符串与所述测试第一实体字符串之间的目标相似系数;

[0309] 停止子单元14322,用于当每个第一实体字符串均被确定为测试第一实体字符串时,停止轮询。

[0310] 其中,开始子单元14321、停止子单元14322的具体功能实现方式可以参见上述图9对应实施例中的步骤S206,这里不再进行赘述。

[0311] 请参见图13,选择子单元1432可以包括:第一获取子单元14323、第二获取子单元14324、第二提取子单元14325、第三提取子单元14326。

[0312] 第一获取子单元14323,用于将所述知识图谱中具有所述意图业务属性的多个实体字符串,均作为第二实体字符串;

[0313] 第二获取子单元14324,用于获取每个第一实体字符串的第一图向量;

[0314] 所述第一获取子单元14323,还用于获取每个第二实体字符串的第二词向量和第二图向量;

[0315] 第二提取子单元14325,用于从多个第二词向量中提取所述目标意图字符串对应的第二词向量,作为目标词向量,并从多个第二图向量中提取所述目标意图字符串对应的第二图向量,作为目标图向量;

[0316] 第三提取子单元14326,用于根据所述目标实体词向量、所述目标图向量、多个第一图向量和所述多个第二词向量,确定所述目标意图字符串和每个第一实体字符串之间的目标相似系数。

[0317] 其中，第一获取子单元14323、第二获取子单元14324、第二提取子单元14325、第三提取子单元14326的具体功能实现方式可以参见上述图9对应实施例中的步骤S206，这里不再进行赘述。

[0318] 请参见图13，知识图谱还包括所述多个实体字符串之间的关联匹配系数；

[0319] 第二获取子单元14324可以包括：初始化子单元143241、采样子单元143242。

[0320] 初始化子单元143241，用于初始化知识图谱中每个实体字符串分别对应的原始图向量；

[0321] 采样子单元143242，用于从所述知识图谱中采样多个样本实体字符串，并获取每个样本实体字符串分别对应的原始图向量，作为样本图向量，根据所述多个样本实体字符串之间的关联匹配系数，采用梯度下降规则分别更新每个样本图向量，得到调整图向量，并将所述调整图向量确定为所述原始图向量；

[0322] 所述采样子单元143242，还用于当采样次数达到次数阈值时，将所述每个第一实体字符串分别对应的更新后的原始图向量，作为所述第一图向量。

[0323] 其中，初始化子单元143241、采样子单元143242的具体功能实现方式可以参见上述图9对应实施例中的步骤S206，这里不再进行赘述。

[0324] 请参见图13，第三提取子单元14326可以包括：第一生成子单元143261、第二生成子单元143262。

[0325] 第一生成子单元143261，用于从所述多个第一实体字符串中提取用于轮询的测试第一实体字符串，将所述测试第一实体字符串对应第一图向量作为测试图向量；

[0326] 所述第一生成子单元143261，还用于根据所述关联匹配系数，确定在所述知识图谱中与所述测试第一实体字符串相邻的第二实体字符串的第二词向量与所述目标词向量之间的第一相似系数；

[0327] 所述第一生成子单143261元，还用于根据所述关联匹配系数，确定所述测试图向量与所述目标图向量之间的第二相似系数；

[0328] 第二生成子单元143262，用于根据所述第一相似系数与所述第二相似系数，生成所述目标意图字符串与所述测试第一实体字符串之间的目标相似系数；

[0329] 所述第二生成子单元143262，还用于当每个第一实体字符串均被确定为测试第一实体字符串时，停止轮询。

[0330] 其中，第一生成子单元143261、第二生成子单元143262的具体功能实现方式可以参见上述图9对应实施例中的步骤S206，这里不再进行赘述。

[0331] 进一步地，请参见图14，是本发明实施例提供的一种电子设备的结构示意图。上述图3-图11对应实施例中的终端设备可以为电子设备1000，如图14所示，所述电子设备1000可以包括：用户接口1002、处理器1004、编码器1006以及存储器1008。信号接收器1016用于经由蜂窝接口1010、WIFI接口1012、...、或NFC接口1014接收或者发送数据。编码器1006将接收到的数据编码为计算机处理的数据格式。存储器1008中存储有计算机程序，处理器1004被设置为通过计算机程序执行上述任一项方法实施例中的步骤。存储器1008可包括易失性存储器(例如，动态随机存取存储器DRAM)，还可以包括非易失性存储器(例如，一次性可编程只读存储器OTPROM)。在一些实例中，存储器1008可进一步包括相对于处理器1004远程设置的存储器，这些远程存储器可以通过网络连接至电子设备1000。用户接口1002可以

包括:键盘1018和显示器1020。

[0332] 在图14所示的电子设备1000中,处理器1004可以用于调用存储器1008中存储计算机程序,以实现:

[0333] 获取目标文本,并获取知识图谱;所述知识图谱包括多个实体字符串,以及与每个实体字符串对应的业务属性字符串;

[0334] 在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串,并提取与所述目标实体字符串对应的目标业务属性字符串;

[0335] 根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型;

[0336] 从所述目标实体字符串中确定与所述目标意图类型相关联的目标意图字符串,并根据所述目标意图字符串生成推荐业务数据。

[0337] 应当理解,本发明实施例中所描述的电子设备1000可执行前文图3到图12所对应实施例中对所述文本处理方法的描述,也可执行前文图13所对应实施例中对所述文本处理装置1的描述,在此不再赘述。另外,对采用相同方法的有益效果描述,也不再进行赘述。

[0338] 此外,这里需要指出的是:本发明实施例还提供了一种计算机存储介质,且所述计算机存储介质中存储有前文提及的文本处理装置1所执行的计算机程序,且所述计算机程序包括程序指令,当所述处理器执行所述程序指令时,能够执行前文图3到图12所对应实施例中对所述文本处理方法的描述,因此,这里将不再进行赘述。另外,对采用相同方法的有益效果描述,也不再进行赘述。对于本发明所涉及的计算机存储介质实施例中未披露的技术细节,请参照本发明方法实施例的描述。

[0339] 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,所述的程序可存储于一计算机可读取存储介质中,该程序在执行时,可包括如上述各方法的实施例的流程。其中,所述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory,ROM)或随机存储记忆体(Random Access Memory,RAM)等。

[0340] 以上所揭露的仅为本发明较佳实施例而已,当然不能以此来限定本发明之权利范围,因此依本发明权利要求所作的等同变化,仍属本发明所涵盖的范围。

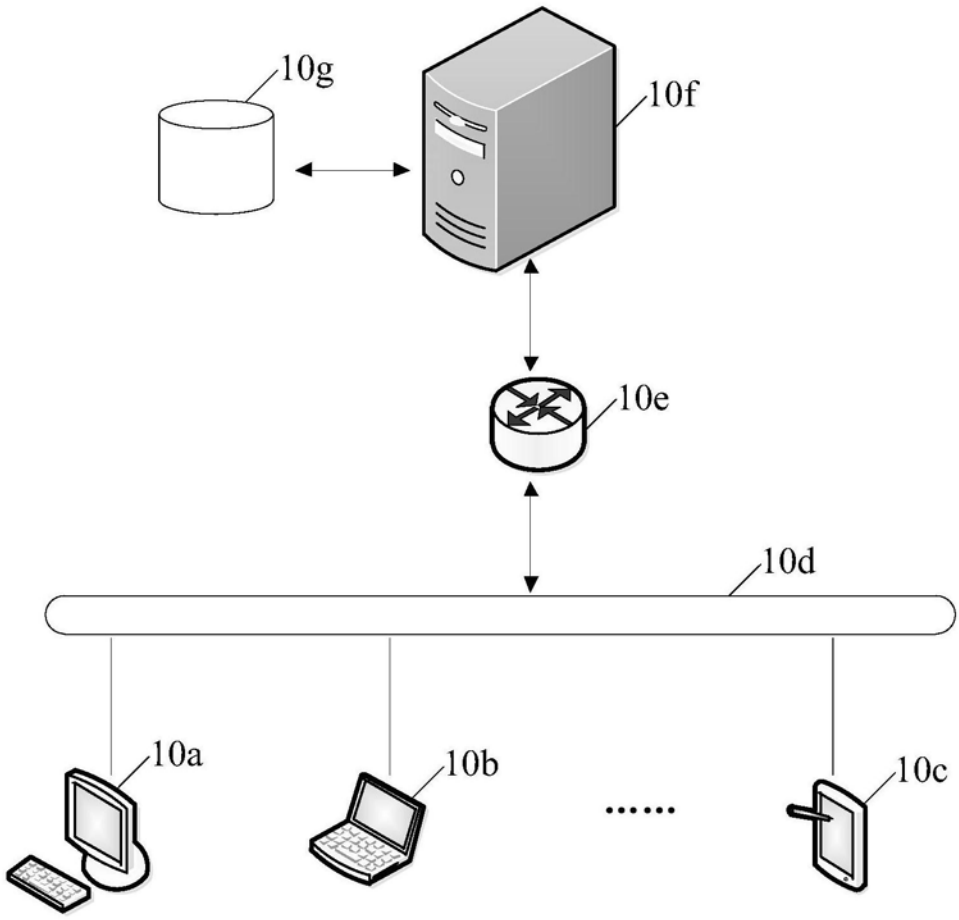


图1

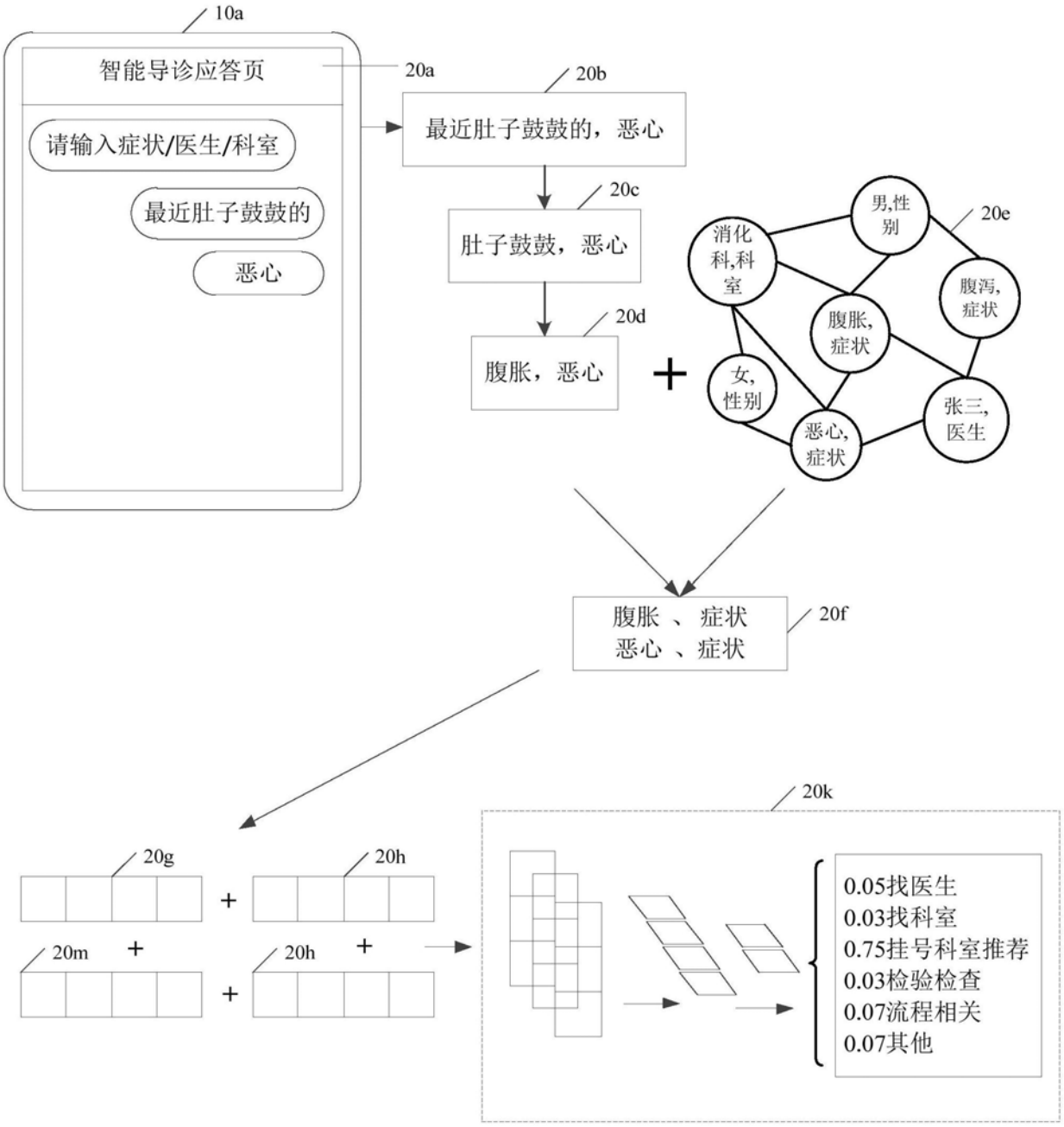

10a
智能导诊应答页
请输入症状/医生/科室
最近肚子鼓鼓的
恶心
20a
20b
最近肚子鼓鼓的，恶心
20c
肚子鼓鼓，恶心
20d
腹胀，恶心
20e
男,性别
消化科,科室
腹泻,症状
腹胀,症状
女,性别
恶心,症状
张三,医生
腹胀 、 症状
恶心 、 症状
20f
20g
20h
20m
20h
20k
0.05找医生
0.03找科室
0.75挂号科室推荐
0.03检验检查
0.07流程相关
0.07其他

图2a

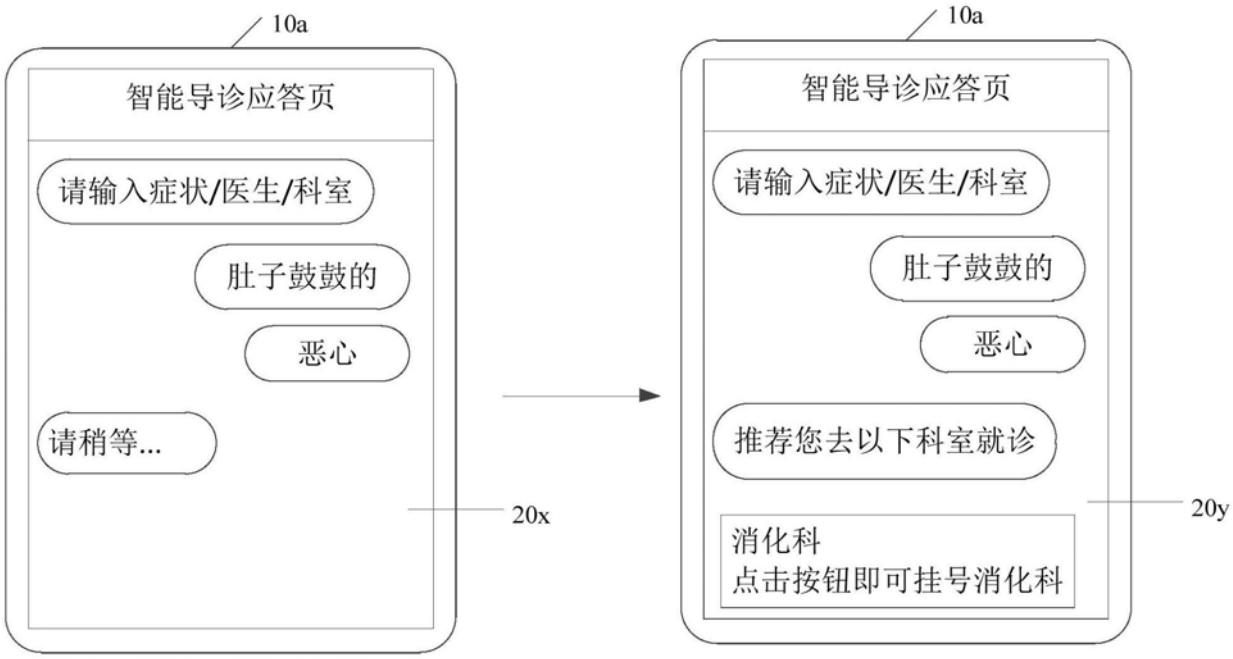

10a
智能导诊应答页
请输入症状/医生/科室
肚子鼓鼓的
恶心
请稍等...
20x
10a
智能导诊应答页
请输入症状/医生/科室
肚子鼓鼓的
恶心
推荐您去以下科室就诊
消化科
点击按钮即可挂号消化科
20y

图2b

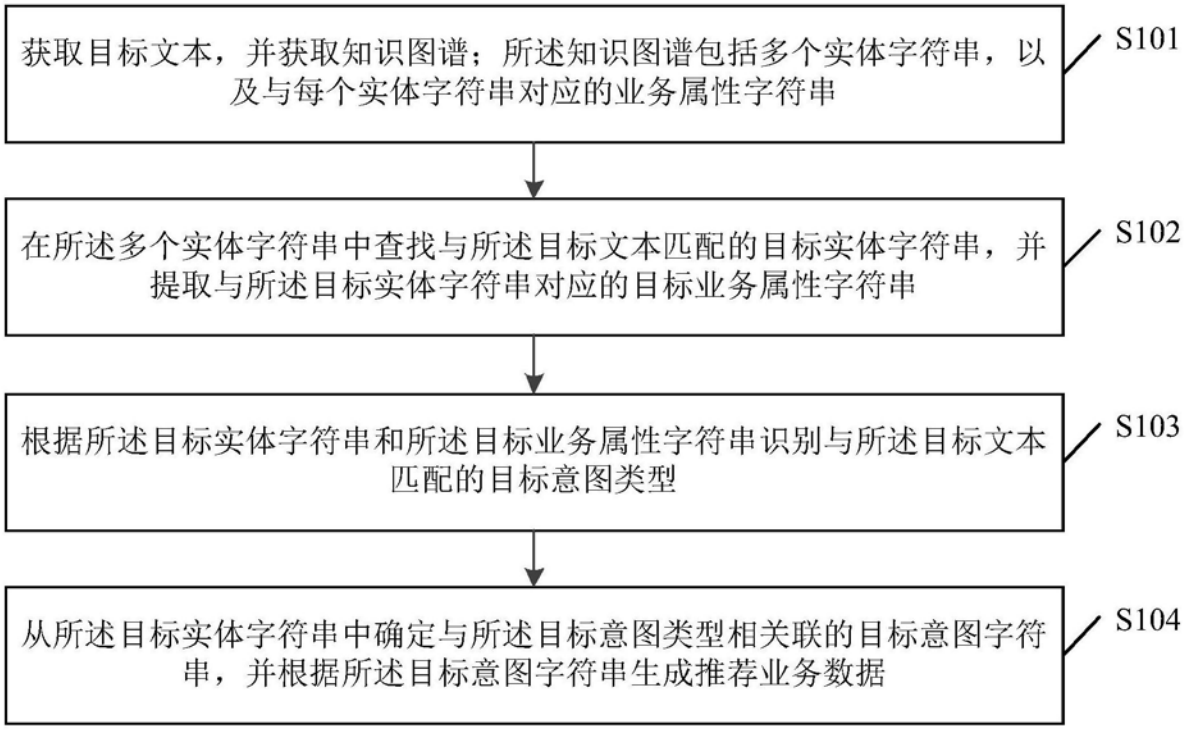

获取目标文本，并获取知识图谱；所述知识图谱包括多个实体字符串，以及与每个实体字符串对应的业务属性字符串
S101
在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串，并提取与所述目标实体字符串对应的目标业务属性字符串
S102
根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型
S103
从所述目标实体字符串中确定与所述目标意图类型相关联的目标意图字符串，并根据所述目标意图字符串生成推荐业务数据
S104

图3

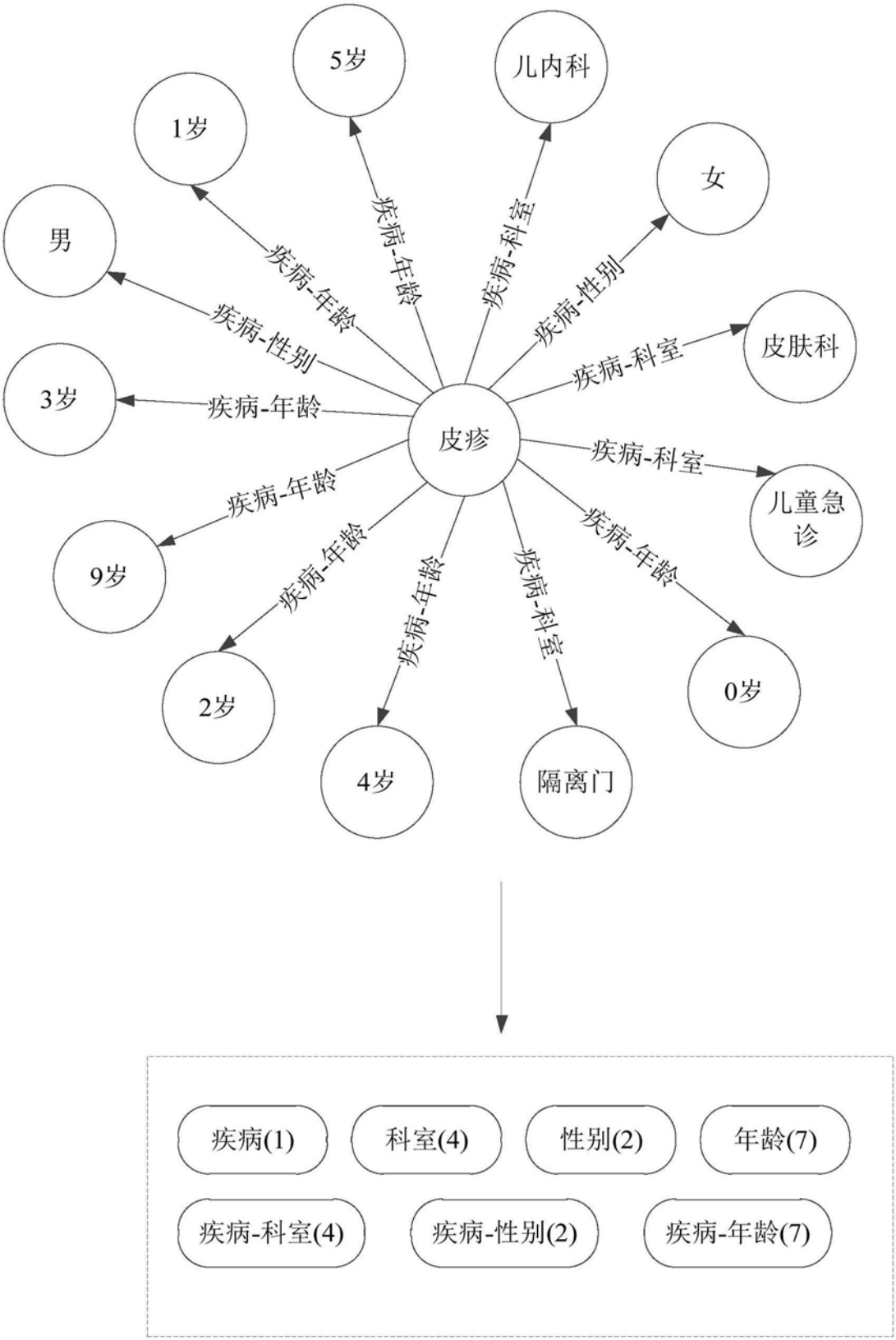

5岁
儿内科
1岁
女
男
疾病-年龄
疾病-性别
疾病-科室
疾病-性别
疾病-科室
皮肤科
3岁
疾病-年龄
皮疹
疾病-科室
儿童急诊
疾病-年龄
9岁
疾病-年龄
疾病-年龄
疾病-年龄
疾病-科室
2岁
4岁
隔离门
0岁
疾病(1)
科室(4)
性别(2)
年龄(7)
疾病-科室(4)
疾病-性别(2)
疾病-年龄(7)

图4

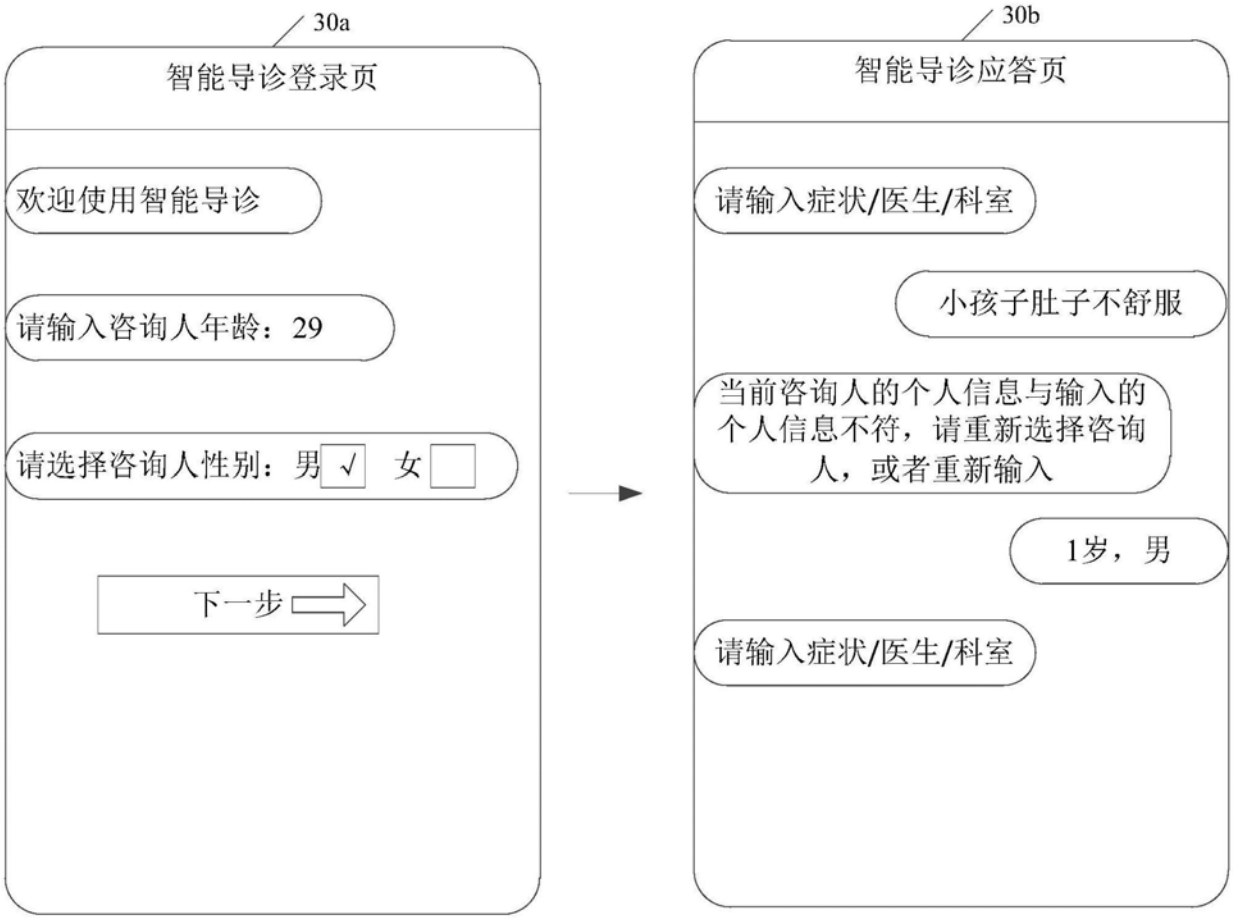

30a
智能导诊登录页
欢迎使用智能导诊
请输入咨询人年龄：29
请选择咨询人性别：男 √ 女
下一步
30b
智能导诊应答页
请输入症状/医生/科室
小孩子肚子不舒服
当前咨询人的个人信息与输入的个人信息不符，请重新选择咨询人，或者重新输入
1岁，男
请输入症状/医生/科室

图5


智能导诊应答页
请输入症状/医生/科室
我要找张三
已为您找到创伤骨科主任医师张三
创伤骨科
张三
主任医师
点击此处查看此医生排班表
点击此处挂号此医生

图6

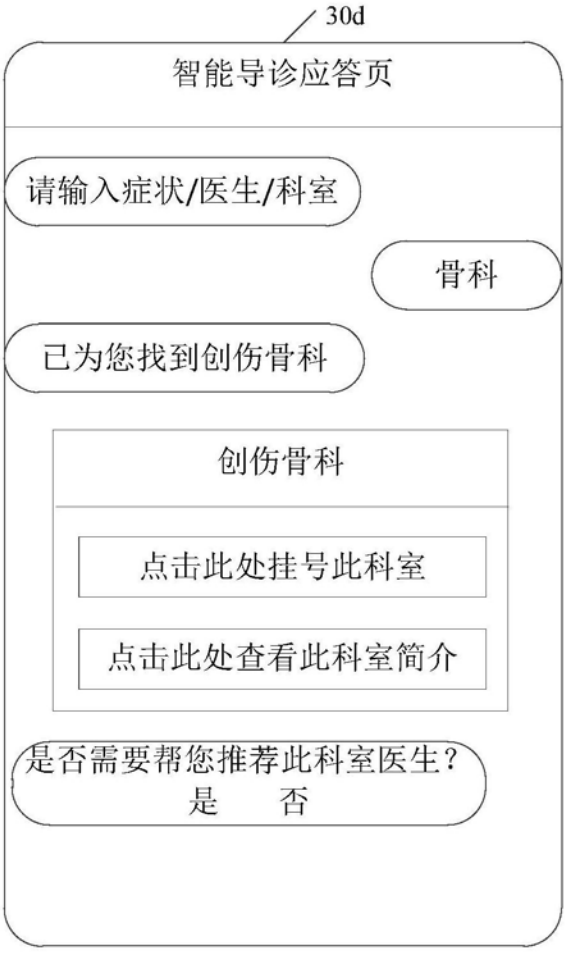

30d
智能导诊应答页
请输入症状/医生/科室
骨科
已为您找到创伤骨科
创伤骨科
点击此处挂号此科室
点击此处查看此科室简介
是否需要帮您推荐此科室医生?
是 否

图7

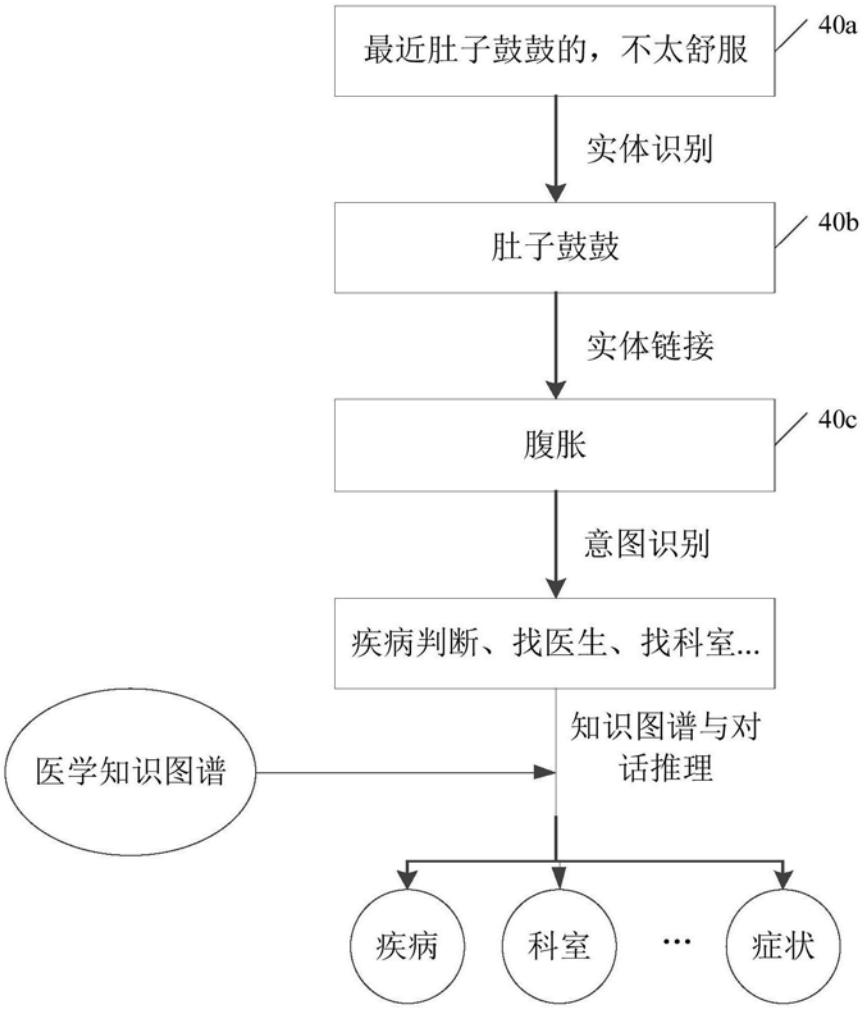

最近肚子鼓鼓的，不太舒服
40a
实体识别
肚子鼓鼓
40b
实体链接
腹胀
40c
意图识别
疾病判断、找医生、找科室...
医学知识图谱
知识图谱与对话推理
疾病
科室
…
症状

图8

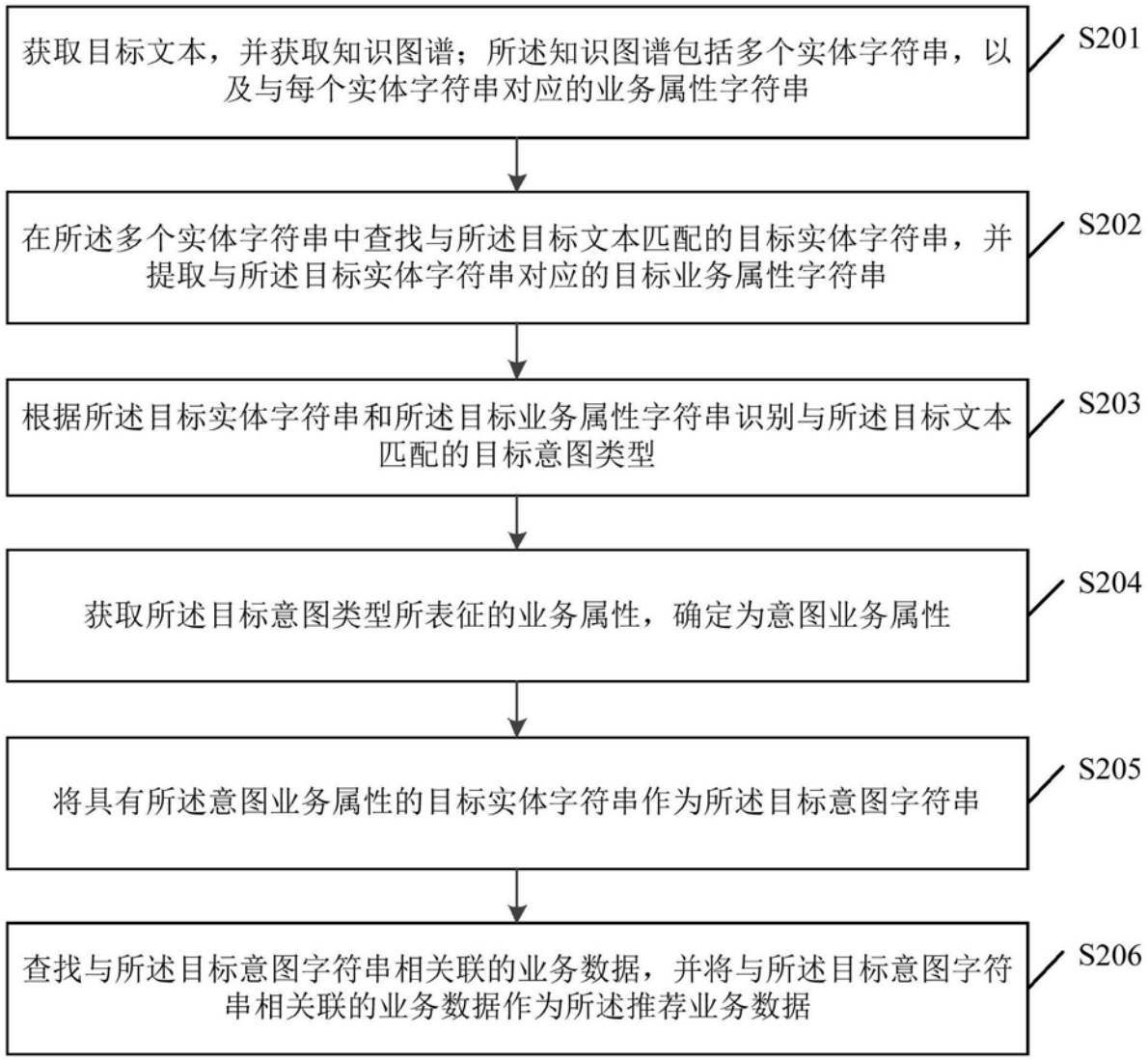

获取目标文本，并获取知识图谱；所述知识图谱包括多个实体字符串，以及与每个实体字符串对应的业务属性字符串
S201
在所述多个实体字符串中查找与所述目标文本匹配的目标实体字符串，并提取与所述目标实体字符串对应的目标业务属性字符串
S202
根据所述目标实体字符串和所述目标业务属性字符串识别与所述目标文本匹配的目标意图类型
S203
获取所述目标意图类型所表征的业务属性，确定为意图业务属性
S204
将具有所述意图业务属性的目标实体字符串作为所述目标意图字符串
S205
查找与所述目标意图字符串相关联的业务数据，并将与所述目标意图字符串相关联的业务数据作为所述推荐业务数据
S206

图9

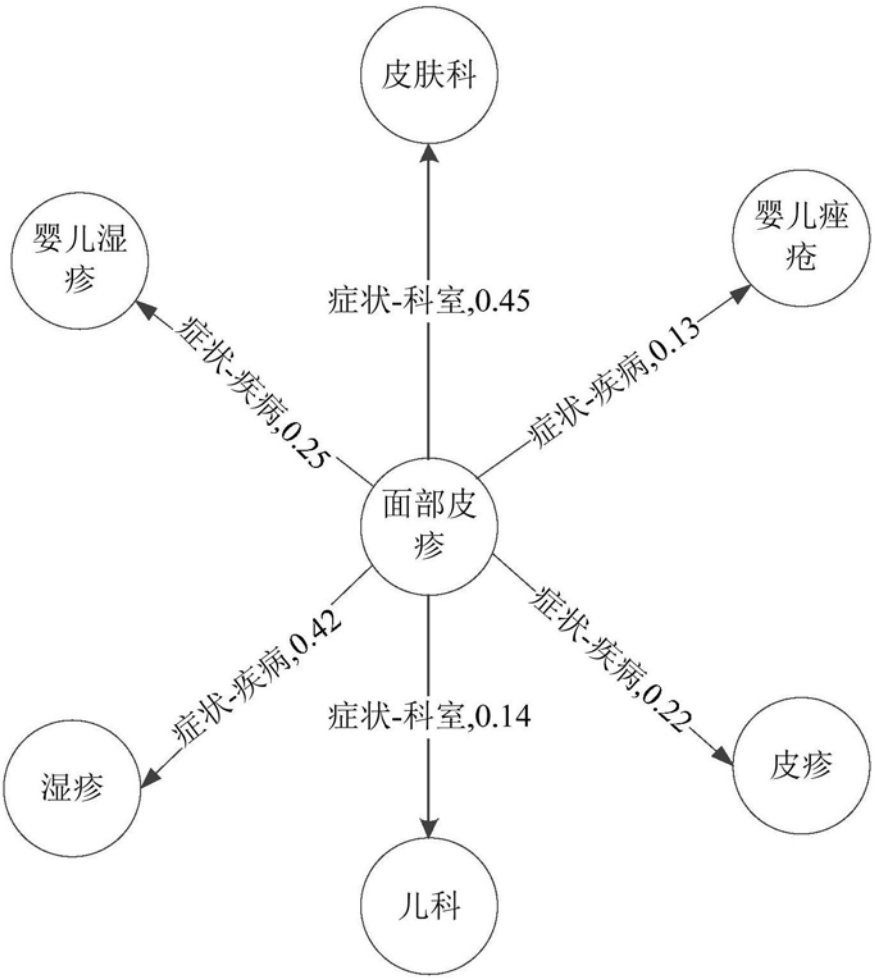

皮肤科
婴儿湿疹
婴儿痤疮
症状-科室,0.45
症状-疾病,0.25
症状-疾病,0.13
面部皮疹
症状-疾病,0.42
症状-疾病,0.22
症状-科室,0.14
湿疹
皮疹
儿科

图10

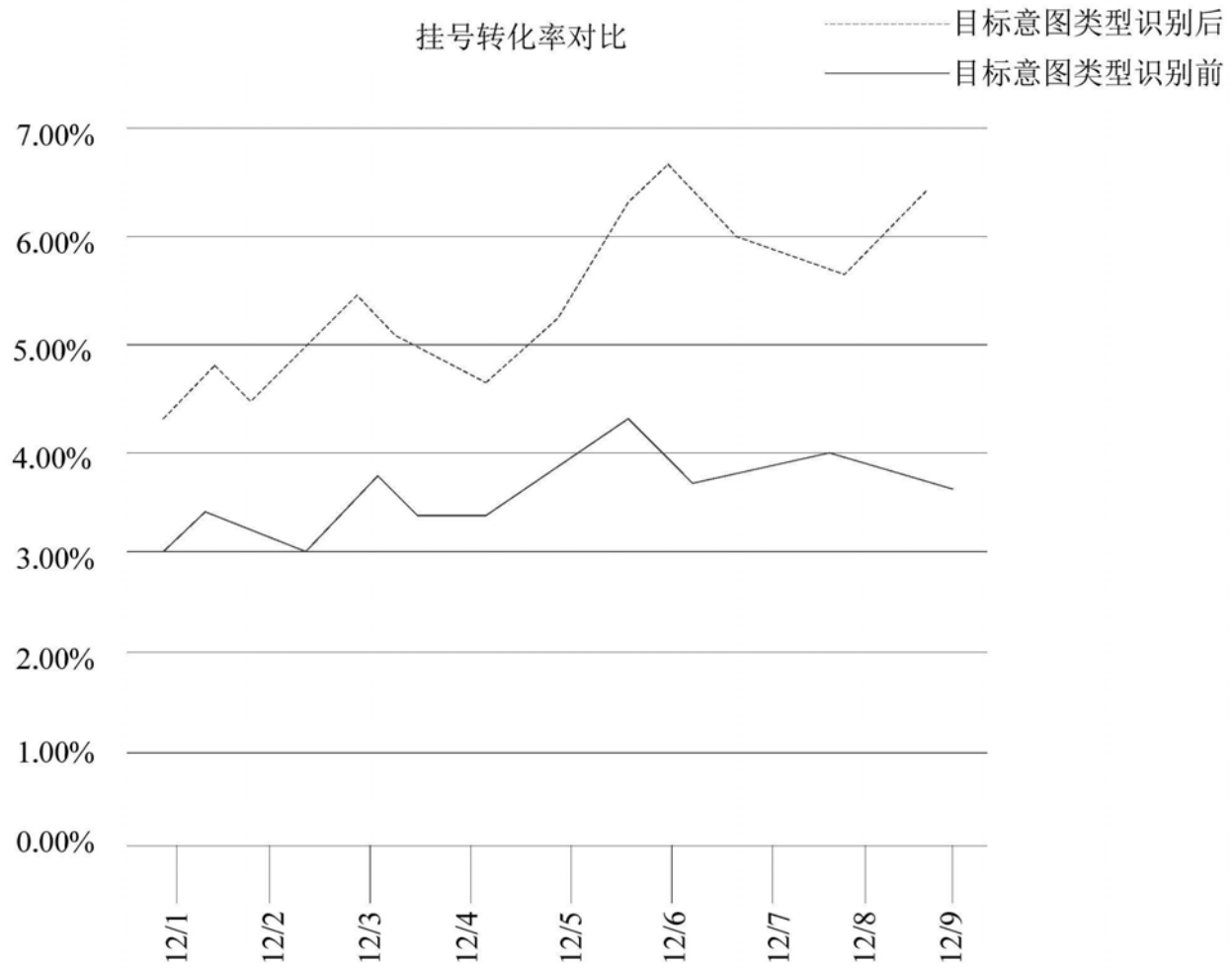

挂号转化率对比
目标意图类型识别后
目标意图类型识别前
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
12/1
12/2
12/3
12/4
12/5
12/6
12/7
12/8
12/9

图11

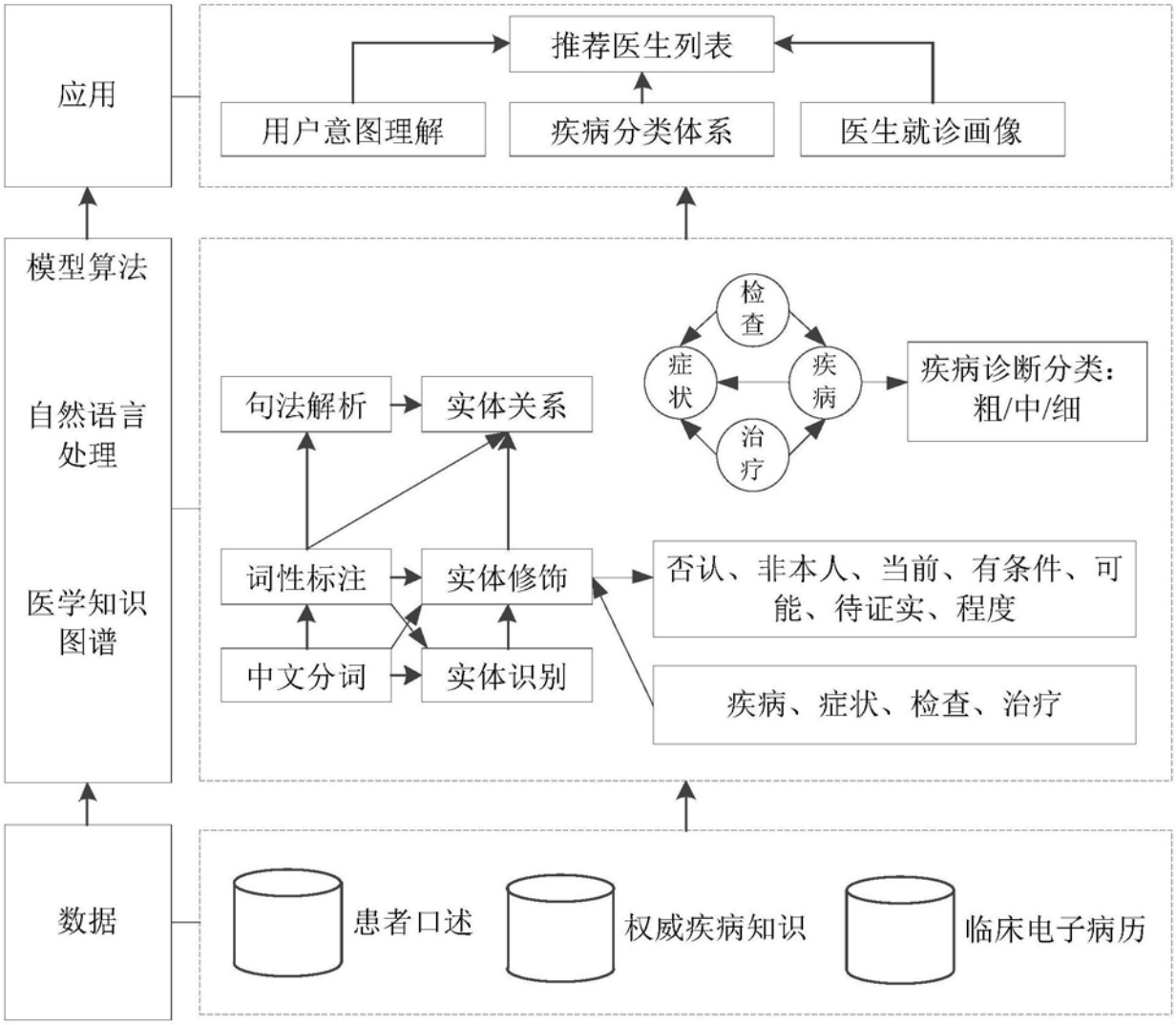

应用
推荐医生列表
用户意图理解
疾病分类体系
医生就诊画像
模型算法
自然语言处理
医学知识图谱
句法解析
实体关系
检查
症状
疾病
治疗
疾病诊断分类：粗/中/细
词性标注
实体修饰
否认、非本人、当前、有条件、可能、待证实、程度
中文分词
实体识别
疾病、症状、检查、治疗
数据
患者口述
权威疾病知识
临床电子病历

图12

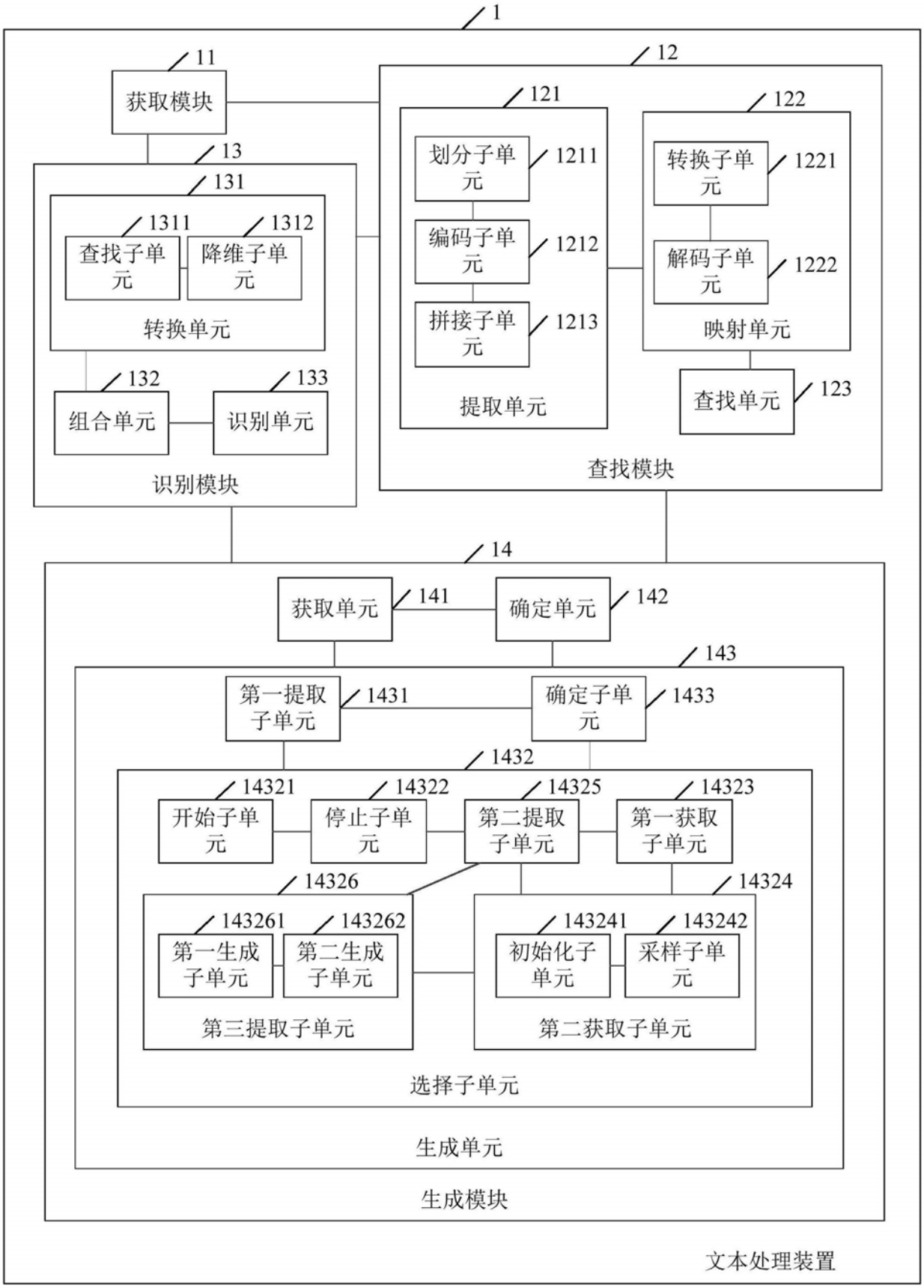

1
11
获取模块
12
121
划分子单元
1211
编码子单元
1212
拼接子单元
1213
提取单元
122
转换子单元
1221
解码子单元
1222
映射单元
123
查找单元
查找模块
13
131
1311
查找子单元
1312
降维子单元
转换单元
132
组合单元
133
识别单元
识别模块
14
获取单元
141
确定单元
142
143
第一提取子单元
1431
确定子单元
1433
1432
14321
开始子单元
14322
停止子单元
14325
第二提取子单元
14323
第一获取子单元
14326
14324
143261
第一生成子单元
143262
第二生成子单元
第三提取子单元
143241
初始化子单元
143242
采样子单元
第二获取子单元
选择子单元
生成单元
生成模块
文本处理装置

图13

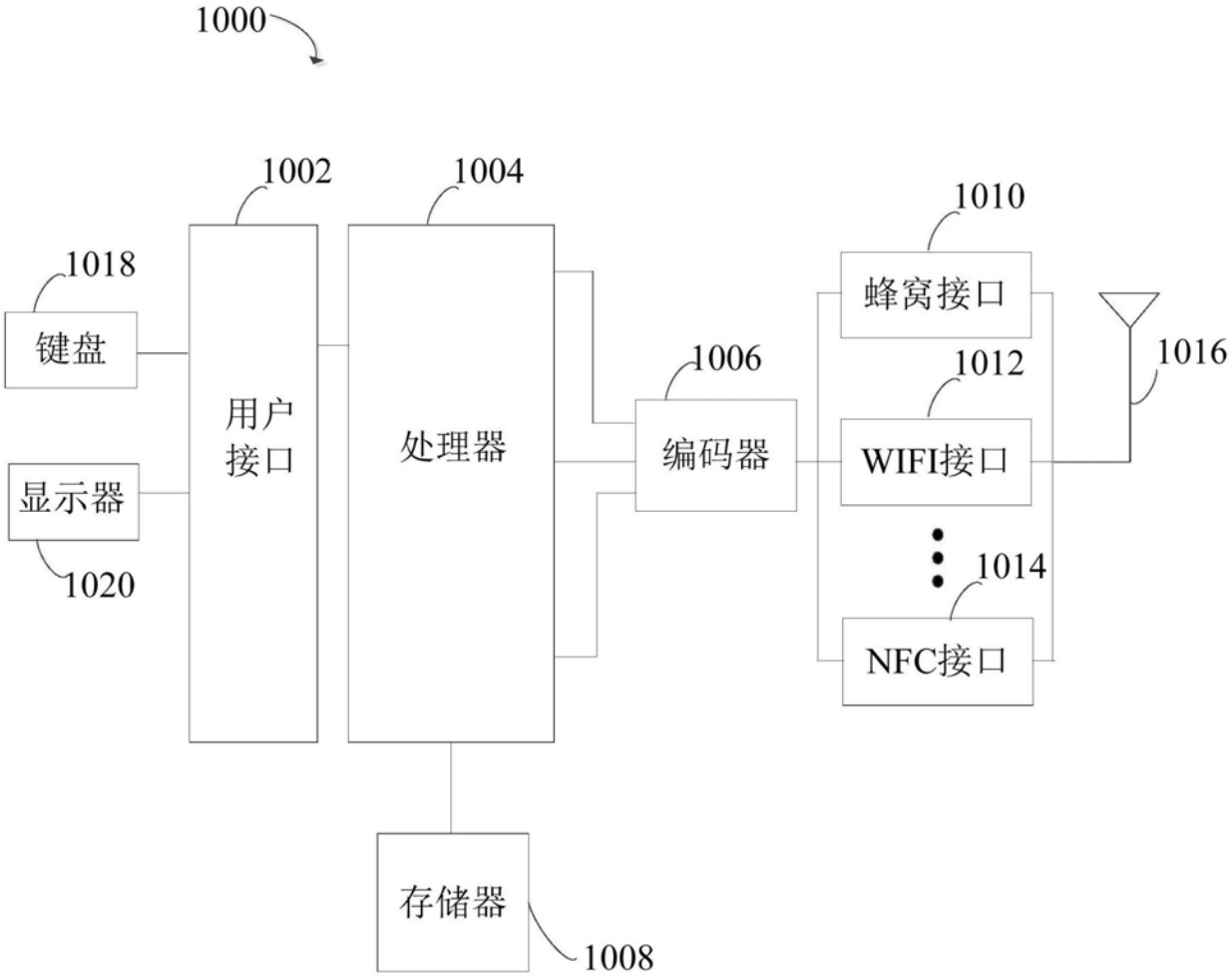

1000
1002
1004
1010
1018
蜂窝接口
1016
键盘
1006
1012
用户
接口
处理器
编码器
WIFI接口
显示器
1014
1020
NFC接口
存储器
1008

图14