(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109147793 B
(45)授权公告日 2020.11.10

(21)申请号 201810946852.8
(22)申请日 2018.08.17
(65)同一申请的已公布的文献号
申请公布号 CN 109147793 A
(43)申请公布日 2019.01.04
(73)专利权人 南京星邺汇捷网络科技有限公司
地址 210000 江苏省南京市建邺区奥体大街69号新城科技大厦01栋15层
(72)发明人 黄哲 沈鹏程 刘树权 张祖齐
(74)专利代理机构 北京超凡志成知识产权代理事务所(普通合伙) 11371
代理人 郭新娟
(51)Int.Cl.
*G10L 15/26*(2006.01)
*G10L 15/18*(2013.01)
*G10L 15/30*(2013.01)

(56)对比文件
CN 103198155 A,2013.07.10
CN 104573028 A,2015.04.29
CN 107305567 A,2017.10.31
CN 104317785 A,2015.01.28
CN 106558309 A,2017.04.05
CN 103116649 A,2013.05.22
JP 特开2007-286901 A,2007.11.01
EP 1052576 A2,2000.11.15

审查员 黄金霞

权利要求书2页 说明书12页 附图5页

(54)发明名称

# 语音数据的处理方法、装置及系统

(57)摘要

本发明提供了一种语音数据的处理方法、装置及系统，该方法首先获取当前语音数据，并将其转换为当前文本数据；对当前文本数据进行自然语言分词，得到分词数据；判断分词数据是否有特殊语义，如果有，对分词数据的语义进行去噪处理；判断去噪后的分词数据或者当前文本数据的语义与前一个文本数据的语义是否有联系；如果有联系，根据前一个文本数据的语义对应的信息，获取当前文本数据的语义对应的信息；如果没有联系，将分词数据的词组与预设的标准数据库中的词组进行比对，根据比对结果确定分词数据的词组对应的信息；最后将得到的信息返回至客户端。本发明在利用对语音数据的处理，提高了查找效率低，而且可对数据进行智能分析。


CN 109147793 B

1.一种语音数据的处理方法,其特征在于,所述方法应用于服务器,所述服务器与客户端通信连接;所述方法包括:

获取当前的语音数据;

将所述当前的语音数据转换为当前文本数据;

对所述当前文本数据进行自然语言分词,得到分词数据;

判断所述分词数据是否有特殊语义,如果有特殊语义,对所述分词数据的语义进行去噪处理;

判断去噪后的所述分词数据的语义或者所述当前文本数据的语义与前一个文本数据的语义是否有联系;

如果有联系,根据所述前一个文本数据的语义对应的信息,获取所述当前文本数据的语义对应的信息;

如果没有联系,将所述分词数据的词组与预设的标准数据库中的词组进行比对,根据比对结果确定所述分词数据的词组对应的信息;所述标准数据库中保存有预先设定的词组;

将所述当前文本数据的语义对应的信息或者所述分词数据的词组对应的信息返回至客户端;

所述方法还包括:通过ALS算法分析用户信息查询的喜好,以使不同的用户在登录客户端后,获得不同的推荐查询的信息;

根据所述前一个文本数据的语义对应的信息,获取所述当前文本数据的语义对应的信息的步骤,包括:

采用遗传算法对所述当前文本数据的语义和所述前一个文本数据的语义进行解析;

获取所述当前文本数据的语义和所述前一个文本数据的语义的维度;

如果所述当前文本数据的语义的维度与所述前一个文本数据的语义的维度相同,根据所述前一个文本数据的语义对应的信息的查找范围,查找所述当前文本数据的语义对应的信息;

如果所述当前文本数据的语义的维度少于所述前一个文本数据的语义的维度,生成提示信息,以提示输入缺少的维度;当接收到缺少的维度后,根据所述前一个文本数据的语义对应的信息的查找范围,查找当前文本数据的语义对应的信息。

2.根据权利要求1所述的方法,其特征在于,将所述当前的语音数据转换为当前文本数据的步骤,包括:通过调用Deep Speech的API接口,将所述当前的语音数据转换为当前文本数据。

3.根据权利要求1所述的方法,其特征在于,对所述当前文本数据进行自然语言分词,得到分词数据的步骤,包括:

采用jieba分词技术,以及预设的标准数据库中词组出现的权重,对当前文本数据进行分词,得到所述当前文本数据的分词数据;所述词组出现的权重根据当前行业领域,通过Trie树结构训练得到。

4.根据权利要求1所述的方法,其特征在于,将所述分词数据的词组与预设的标准数据库中的词组进行比对,根据比对结果确定所述分词数据的词组对应的信息的步骤,包括:

将所述分词数据的词组与标准数据库中的词组进行比对,得到词组的比对相识度;

如果所述比对相识度大于75％,根据所述当前文本数据的词组的维度获得所述分词数据的词组对应的信息;

如果所述比对相识度大于45％且小于75％,将所述比对相识度最高的所述当前文本数据的词组对应的信息,作为所述分词数据的词组对应的信息;

如果所述比对相识度小于45％,从日常用语资料库中查询得到所述当前文本数据的词组对应的信息。

5.一种语音数据的处理装置,其特征在于,所述装置设置于服务器,所述服务器与客户端通信连接;所述装置包括:

数据获取模块,用于获取当前的语音数据;

语音转换模块,用于将所述当前的语音数据转换为当前文本数据;

分词模块,用于对所述当前文本数据进行自然语言分词,得到分词数据;

特殊语义判断模块,用于判断所述分词数据是否有特殊语义,如果有特殊语义,对所述分词数据的语义进行去噪处理;

去噪模块,用于判断去噪后的所述分词数据的语义或者所述当前文本数据的语义与前一个文本数据的语义是否有联系;

信息获取模块,用于如果有联系,根据所述前一个文本数据的语义对应的信息,获取所述当前文本数据的语义对应的信息;

信息获取模块,还用于如果没有联系,将所述分词数据的词组与预设的标准数据库中的词组进行比对,根据比对结果确定所述分词数据的词组对应的信息;所述标准数据库中保存有预先设定的词组;

信息返回模块,用于将所述当前文本数据的语义对应的信息或者所述分词数据的词组对应的信息返回至客户端;

所述装置还包括:信息推荐模块,用于通过ALS算法分析用户信息查询的喜好,以使不同的用户在登录客户端后,获得不同的推荐查询的信息;

所述信息获取模块,还用于:如果有联系,采用遗传算法对所述当前文本数据的语义和所述前一个文本数据的语义进行解析;获取所述当前文本数据的语义和所述前一个文本数据的语义的维度;如果所述当前文本数据的语义的维度与所述前一个文本数据的语义的维度相同,根据所述前一个文本数据的语义对应的信息的查找范围,查找所述当前文本数据的语义对应的信息;如果所述当前文本数据的语义的维度少于所述前一个文本数据的语义的维度,生成提示信息,以提示输入缺少的维度;当接收到缺少的维度后,根据所述前一个文本数据的语义对应的信息的查找范围,查找当前文本数据的语义对应的信息。

6.根据权利要求5所述的装置,其特征在于,语音转换模块,还用于通过调用Deep Speech的API接口,将所述当前的语音数据转换为当前文本数据。

7.一种语音数据的处理系统,其特征在于,所述系统包括存储器以及处理器,所述存储器用于存储支持处理器执行权利要求1至4任一项所述方法的程序,所述处理器被配置为用于执行所述存储器中存储的程序。

# 语音数据的处理方法、装置及系统

## 技术领域

[0001] 本发明涉及语音数据处理技术领域，尤其是涉及语音数据的处理方法、装置及系统。

## 背景技术

[0002] 现有技术中通过浏览器或手机App登录到指标报表查询页面，找到要查询的报表或指标的菜单目录，输入要查询的地区、时间等查询条件进行查询，然而对于电信行业数千张报表和KPI(Key Performance Indicator，关键绩效指标)来说，菜单目录多，查找比较困难；而且随报表或指标的增加，需要频繁增加报表查询菜单，操作步骤复杂，菜单维护开发工作量大；同时，传统的数据结果的智能化程度较低，难以实现数据的多样化分析。

## 发明内容

[0003] 有鉴于此，本发明的目的在于提供一种语音数据的处理方法、装置及系统，以提高数据处理的效率和数据智能分析能力。

[0004] 第一方面，本发明实施例提供了一种语音数据的处理方法，该方法应用于服务器，该服务器与客户端通信连接；该方法包括：获取当前的语音数据；将当前的语音数据转换为当前文本数据；对当前文本数据进行自然语言分词，得到分词数据；判断分词数据是否有特殊语义，如果有特殊语义，对分词数据的语义进行去噪处理；判断去噪后的分词数据的语义或者当前文本数据的语义与前一个文本数据的语义是否有联系；如果有联系，根据前一个文本数据的语义对应的信息，获取当前文本数据的语义对应的信息；如果没有联系，将分词数据的词组与预设的标准数据库中的词组进行比对，根据比对结果确定分词数据的词组对应的信息；标准数据库中保存有预先设定的词组；将当前文本数据的语义对应的信息或者分词数据的词组对应的信息返回至客户端。

[0005] 进一步，上述方法还包括：通过ALS算法分析用户信息查询的喜好，以使不同的用户在登录客户端后，获得不同的推荐查询的信息。

[0006] 进一步，将当前的语音数据转换为当前文本数据的步骤，包括：通过调用Deep Speech的API接口，将当前的语音数据转换为当前文本数据。

[0007] 进一步，对当前文本数据进行自然语言分词，得到分词数据的步骤，包括：采用jieba分词技术，以及预设的标准数据库中词组出现的权重，对当前文本数据进行分词，得到当前文本数据的分词数据；词组出现的权重根据当前行业领域，通过Trie树结构训练得到。

[0008] 进一步，根据所述前一个文本数据的语义对应的信息，获取所述当前文本数据的语义对应的信息的步骤，包括：采用遗传算法对当前文本数据的语义和前一个文本数据的语义进行解析；获取当前文本数据的语义和前一个文本数据的语义的维度；如果当前文本数据的语义的维度与前一个文本数据的语义的维度相同，根据前一个文本数据的语义对应的信息的查找范围，查找当前文本数据的语义对应的信息；如果当前文本数据的语义的维

度少于前一个文本数据的语义的维度，生成提示信息，以提示输入缺少的维度；当接收到缺少的维度后，根据前一个文本数据的语义对应的信息的查找范围，查找当前文本数据的语义对应的信息。

[0009] 进一步，将分词数据的词组与预设的标准数据库中的词组进行比对，根据比对结果确定分词数据的词组对应的信息的步骤，包括：将分词数据的词组与标准数据库中的词组进行比对，得到词组的比对相识度；如果比对相识度大于75％，根据当前文本数据的词组的维度获得分词数据的词组对应的信息；如果比对相识度大于45％且小于75％，将比对相识度最高的当前文本数据的词组对应的信息，作为分词数据的词组对应的信息；如果比对相识度小于45％，从日常用语资料库中查询得到当前文本数据的词组对应的信息。

[0010] 第二方面，本发明实施例还提供一种语音数据的处理装置，该装置设置于服务器，该服务器与客户端通信连接；该装置包括：数据获取模块，用于获取当前的语音数据；语音转换模块，用于将当前的语音数据转换为当前文本数据；分词模块，用于对当前文本数据进行自然语言分词，得到分词数据；特殊语义判断模块，用于判断分词数据是否有特殊语义，如果有特殊语义，对分词数据的语义进行去噪处理；去噪模块，用于判断去噪后的分词数据的语义或者当前文本数据的语义与前一个文本数据的语义是否有联系；信息获取模块，用于如果有联系，根据前一个文本数据的语义对应的信息，获取当前文本数据的语义对应的信息；信息获取模块，还用于如果没有联系，将分词数据的词组与预设的标准数据库中的词组进行比对，根据比对结果确定分词数据的词组对应的信息；标准数据库中保存有预先设定的词组；信息返回模块，用于将当前文本数据的语义对应的信息或者分词数据的词组对应的信息返回至客户端。

[0011] 进一步，上述装置还包括：信息推荐模块，用于通过ALS算法分析用户信息查询的喜好，以使不同的用户在登录客户端后，获得不同的推荐查询的信息。

[0012] 进一步，语音转换模块，还用于通过调用Deep Speech的API接口，将当前的语音数据转换为当前文本数据。

[0013] 第三方面，本发明实施例还提供一种语音数据的处理系统，该系统包括存储器以及处理器，所述存储器用于存储支持处理器执行第一方面的方法的程序，处理器被配置为用于执行所述存储器中存储的程序。

[0014] 本发明实施例带来了以下有益效果：

[0015] 本发明提供了一种语音数据的处理方法、装置及系统，该方法首先获取当前语音数据，并将其转换为当前文本数据；对当前文本数据进行自然语言分词，得到分词数据；判断分词数据是否有特殊语义，如果有，对分词数据的语义进行去噪处理；判断去噪后的分词数据或者当前文本数据的语义与前一个文本数据的语义是否有联系；如果有联系，根据前一个文本数据的语义对应的信息，获取当前文本数据的语义对应的信息；如果没有联系，将分词数据的词组与预设的标准数据库中的词组进行比对，根据比对结果确定分词数据的词组对应的信息；最后将得到的信息返回至客户端。本发明在语音数据查找工作量大的情况下，提高了查找效率低，而且可对语音数据进行智能分析。

[0016] 本发明的其他特征和优点将在随后的说明书中阐述，或者，部分特征和优点可以从说明书推知或毫无疑义地确定，或者通过实施本发明的上述技术即可得知。

[0017] 为使本发明的上述目的、特征和优点能更明显易懂，下文特举较佳实施方式，并配

合所附附图,作详细说明如下。

## 附图说明

[0018] 为了更清楚地说明本发明具体实施方式或现有技术中的技术方案,下面将对具体实施方式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,下面描述中的附图是本发明的一些实施方式,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。

[0019] 图1为本发明实施例提供的一种语音数据的处理方法的流程图;

[0020] 图2为本发明实施例提供的另一种语音数据的处理方法中,获取当前文本数据的语义对应的信息的流程图;

[0021] 图3为本发明实施例提供的另一种语音数据的处理方法中,确定分词数据的词组对应的信息的流程图;

[0022] 图4为本发明实施例提供的另一种语音数据的处理方法的流程图;

[0023] 图5为本发明实施例提供的一种语音数据的处理系统的结构示意图。

## 具体实施方式

[0024] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合附图对本发明的技术方案进行清楚、完整地描述,显然,所描述的实施例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,都属于本发明保护的范围。

[0025] 现有的通过浏览器或者手机APP登录到指标或者报表查询页面,对所需信息进行查找的方式,由于指标或者报表的增加,使得查找工作量大,效率低,而且难以对语音数据进行智能分析,基于此,本发明实施例提供的一种语音数据的处理方法、装置及系统,该技术可以应用于电信或者其他行业数据查询的场景中。

[0026] 为便于对本实施例进行理解,首先对本发明实施例所公开的一种语音数据的处理方法进行详细介绍。

[0027] 参见图1所示的一种语音数据的处理方法的流程图,该方法应用于服务器,该服务器与客户端通信连接;该方法的具体步骤,包括:

[0028] 步骤S102,获取当前的语音数据;

[0029] 用户输入语音数据,语音数据中包含需要查询的信息的关键词,通过对该语音数据的解析,可以查询到需要的信息数据。

[0030] 步骤S104,将上述当前的语音数据转换为当前文本数据;

[0031] 将当前输入的语音数据,通过语音识别技术将其转化为文本数据,以实现对语音的识别。

[0032] 步骤S106,对当前文本数据进行自然语言分词,得到分词数据;

[0033] 转换后的文本数据通过自然语言理解(Natural Language Understanding,NLU)进行分词;该自然语言理解技术,涵盖领域非常广泛,包括句子检测、分词、词性标注、句法分析、文本分类或聚类、文字角度、信息抽取或自动摘要、机器翻译、自动问答、文本生成等多个领域。

[0034] 步骤S108,判断上述分词数据是否有特殊语义,如果有特殊语义,对该分词数据的语义进行去噪处理;

[0035] 通过语音数据转换成文本数据后,由于普通话或者方言的差异,在语音识别中识别出的文本数据并非是完全准确的,同时也存在着许多白话的差异,这些差异导致分词数据可能出现特殊语义。

[0036] 针对上述特殊语义的分词数据需要对分词后的词组去除没有用的数据,这个去除没有用的数据的处理过程为去噪处理的工程,以电信行业为例,输入语音“帮我查下,南京5月份的天翼发展量是多少”,在这个文本数据中通过分词后产生的结果中“帮”,“我”,“查询”,“多少”,“的”,这些都属于白话文,对于信息数据的查询无实际意义,因此需要对这些分词后的数据进行过滤,过滤后的主干内容为“下”,“南京”,“5月”,“天翼发展”。

[0037] 步骤S110,判断去噪后的分词数据的语义或者当前文本数据的语义与前一个文本数据的语义是否有联系;

[0038] 根据前一个文本数据对应的信息,也就是查询的信息,可以根据当前文本数据与前一个文本数据的语义的联系来得到当前需要查询的信息数据,可以简化处理过程。

[0039] 步骤S112,如果有联系,根据所述前一个文本数据的语义对应的信息,获取所述当前文本数据的语义对应的信息;

[0040] 上述当文本数据的语义与前一个文本数据的语义有联系,若当前文本数据的语义和前一个文本数据的语义的维度相同,可根据前一个文本数据对应的信息的查找范围,来查找当前文本数据对应的信息;如果当前文本数据的语义的维度比前一个文本数据的语义的维度少,则会出现语音提示或者声称文字,以提醒用户输入缺少的维度的语音数据,接收到缺少的维度后,同样可根据前一个文本数据对应的信息的查找范围,来查找当前文本数据的语义对应的信息。

[0041] 步骤S114,如果没有联系,将分词数据的词组与预设的标准数据库中的词组进行比对,根据比对结果确定分词数据的词组对应的信息;上述标准数据库中保存有预先设定的词组;

[0042] 上述标准数据库是根据所要查找的当前行业领域的专业术语,通过不断地训练得到的词组,例如电信行业的词组有:4G(the 4th Generation mobile communication technolog,第四代移动通信技术)、5G、天翼、流量、话费等等。

[0043] 将分词数据的词组与预设的标准数据库中的词组进行比对,得到比对相识度,根据比对相识度的大小,来查找可查询信息数据的词组对应的信息。

[0044] 步骤S116,将当前文本数据的语义对应的信息或者分词数据的词组对应的信息返回至客户端。

[0045] 将上述词组得到的信息返回到客户端,以便用户从客户端对该信息进行获取。

[0046] 本实施例提供了一种语音数据的处理方法,该方法首先获取当前语音数据,并将其转换为当前文本数据;对当前文本数据进行自然语言分词,得到分词数据;判断分词数据是否有特殊语义,如果有,对分词数据的语义进行去噪处理;判断去噪后的分词数据或者当前文本数据的语义与前一个文本数据的语义是否有联系;如果有联系,根据前一个文本数据的语义对应的信息,获取当前文本数据的语义对应的信息;如果没有联系,将分词数据的词组与预设的标准数据库中的词组进行比对,根据比对结果确定分词数据的词组对应的信

息;最后将得到的信息返回至客户端。本方法利用对语音数据的处理,提高了查找效率低,而且可对数据进行智能分析。

[0047] 本发明实施例还提供了另一种语音数据的处理方法,该方法在图1中所示方法基础上实现;该方法还包括:通过ALS(Alternating Least Square,交替最小二乘法)算法分析用户信息查询的喜好,以使不同的用户在登录客户端后,获得不同的推荐查询的信息。

[0048] 上述ALS算法是基于矩阵分解的CF(Collaborative Filtering,协同滤波)算法中的一种;该算法通常应用于基于矩阵分解的推荐系统中,例如:将用户(user)对商品(item)的评分矩阵分解为两个矩阵:一个是用户对商品隐含特征的偏好矩阵,另一个是商品所包含的隐含特征的矩阵,在这个矩阵分解的过程中,评分缺失项得到了填充,也就是说我们可以基于这个填充的评分来给用户推荐喜爱的商品。

[0049] 该方法是通过基于用户CF的基本思想进行信息(例如,电信行业的指标、报表)的推荐,通过分析用户的访问喜好,访问频度,进行智能的指标、报表的推荐。

[0050] 上述将当前的语音数据转换为当前文本数据的步骤,包括:通过调用Deep Speech的API(Application Programming Interface,应用程序编程接口)接口,将当前的语音数据转换为当前文本数据。

[0051] 为了对输入的语音数据转换成文本数据,需要有支撑语音解析的功能,那么上述Deep Speech是百度研发的语音识别系统,可通过调用百度发布的API接口,将语音数据内容翻译成文本数据并返回给调用方,该语音识别的准确率比较高。

[0052] 上述当前文本数据进行自然语言分词,得到分词数据的步骤,包括:采用jieba(结巴)分词技术,以及预设的标准数据库中词组出现的权重,对当前文本数据进行分词,得到所述当前文本数据的分词数据;所述词组出现的权重根据当前行业领域,通过Trie树(字典树)结构训练得到。

[0053] 上述jieba分词技术的基本原理为:

[0054] 1、基于Trie树结构实现高效的词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图(Directed Acyclic Graph,DAG);

[0055] 根据dict.txt生成trie树;字典在生成trie树的同时,也把每个词的出现次数转换成了频率。

[0056] 对待分词句子,根据dict.txt生成的trie树,生成DAG,实际上通常是指对待分词句子,根据给定的词典进行查词典操作,生成几种可能的句子切分;在DAG中记录的是句子中某个词的开始位置,从0到n-1(n为句子的长度),每个开始位置作为字典的键,value是个list,其中保存了可能的词语的结束位置(通过查字典得到词,开始位置+词语的长度得到结束位置);通常情况下,jieba支持全模式分词,能把句子中所有的可以成词的词语都扫描出来。例如:{0:[12,3]}这样一个简单的DAG,就是表示0位置开始,在1,2,3位置都是词,也就是说0～1,0～2,0～3这三个起始位置之间的字符,在dict.txt中是词语。

[0057] 例如,在基于语义解析的应用上,首先采用jieba自带的中文分词库,可以正确分析出来常用词频,但对于电信专业术语无法分词,例如“五月天翼发展量”,分词后会得到“五月天”,“易发展”,“量”,因为五月天在分词词库中出现的词频最高。因此需要针对不同的行业领域,训练不同的标准数据库,例如,在构造电信行业的术语时,将重新训练分词库中的词组出现权重(相当于上述词频),得到结果为“五月”,“天翼发展量”的分词形式。

[0058] 2、采用了动态规划查找最大概率路径，找出基于词频的最大切分组合；

[0059] 查找待分词句子中已经切分好的词语，对该词语查找该词语出现的频率(次数除以总数)，如果没有该词语，就把词典中出现频率最小的那个词语的频率作为该词的频率，也就是说P(某词语)＝FREQ.get(“某词语”,min_freq)。

[0060] 根据动态规划查找最大概率路径的方法，对句子从右往左反向计算最大概率如P(NodeN)＝1.0,P(NodeN-1)＝P(NodeN)*Max(P(倒数第一个词))…依次类推，最后得到最大概率路径，得到最大概率的切分组合。

[0061] 在语义解析上采用词频的最大切分，例如“帮我查询下南京五月份天翼发展量”，切分后的词频为“帮/我/查询/下/南京/五月份/天翼/发展/量”,“帮我/查询下/南京/五月份/天翼发展/量”,“帮我/查询/下/南京/五/月份/天翼发展/量”等。

[0062] 3、对于未登录词，采用了基于汉字成词能力的HMM(Hidden Markov Model，隐马尔科夫)模型，使用了Viterbi算法(维特比算法)。

[0063] 上述未登录词通常是指词典dict.txt中没有记录的词(也就是就算把dict.txt中所有的词汇全部删掉，jieba依然能够分词，不过分出来的词大部分的长度为2，这个就是基于HMM来预测分词的过程)。

[0064] 参见图2所示的另一种语音数据的处理方法中获取当前文本数据的语义对应的信息的流程图；根据前一个文本数据的语义对应的信息，获取当前文本数据的语义对应的信息的具体步骤，包括：

[0065] 步骤S202，采用遗传算法对当前文本数据的语义和前一个文本数据的语义进行解析；

[0066] 利用该遗传算法，对上下文(相当于当前文本数据和前一个文本数据)的语义进行交叉变异，寻找上下文之间的联系，并针对有联系的语义进行解析处理。

[0067] 上述遗传算法(Genetic Algorithm,GA)，通常是以编码空间代替问题参数空间，从代表问题可能有潜在解集的一个种群出发，按照生物进化过程中适者生存、优胜劣汰的原理，以适应度作为评价个体优劣的依据，重复使用选择、交叉、变异算子作用于群体，使之不断进化，逐渐接近最优解。

[0068] 步骤S204，获取上述当前文本数据的语义和前一个文本数据的语义的维度；

[0069] 上述维度通常是指文本数据中关键词的个数，例如，在电信行业中，对指标数据进行查询，通常需要三个维度的信息：时间、地点和指标名称，其中组成时间、地点和指标名称的词组，即为关键词。

[0070] 步骤S206，如果当前文本数据的语义的维度与前一个文本数据的语义的维度相同，根据前一个文本数据的语义对应的信息的查找范围，查找当前文本数据的语义对应的信息；

[0071] 以电信行业为例，在上一个指标查询后，如果需要重新查询当前指标的不同维度的数据量，可以直接说地区、时间，在此过程中采用遗传算法记录先前查询的记录，当新的对话中是查询维度的信息时，则不进行指标切换，而是产生遗传信息变异，产生新的查询信息。例如场景为“帮我查询下，南京5月份的天翼发展量是多少”，查询返回结果后，继续问“6月份”，则会显示6月份的数据量，继续问“南京，无锡，7月份的”，则会显示南京，无锡地区7月份的数据量；如果继续问“新装宽带发展量”由于该指标是个全新的指标不会遗传之前的

指标，系统会返回“新装宽带发展量”。同时，也可以通过语音分析出同比、环比、增量，例如“查询下3月份4月份的同比”或“3月份4月份用户增量”。

[0072] 步骤S208，如果当前文本数据的语义的维度少于前一个文本数据的语义的维度，生成提示信息，以提示输入缺少的维度；当接收到缺少的维度后，根据前一个文本数据的语义对应的信息的查找范围，查找当前文本数据的语义对应的信息。

[0073] 同样，以电信行业为例，若需要重新查询当前指标的不同维度的数据量，只说了地区，而没有说时间；由于缺少时间维度信息，则会出现语音或本文提示用户输入需要查询的时间，输入缺少的维度信息后，再次按照步骤S206的方式得到当前文本的语义对应的信息(也就是上述的指标数据)。

[0074] 参见图3所示的另一种语音数据的处理方法中确定分词数据的词组对应的信息的流程图；将分词数据的词组与预设的标准数据库中的词组进行比对，根据比对结果确定该分词数据的词组对应的信息的步骤，包括：

[0075] 步骤S302，将分词数据的词组与标准数据库中的词组进行比对，得到词组的比对相识度；

[0076] 将分词数据的词组保存到数据中，对每个数据的内容循环采用最短距离(Levenshtein)算法与标准数据库进行比对；其中，最短距离相识度算法通常用于计算两个字符串之间的Levenshtein距离，而Levenshtein距离又称为编辑距离，是指两个字符串之间，由一个转换成另一个所需的最少编辑操作次数。

[0077] 步骤S304，如果比对相识度大于75％，根据当前文本数据的词组的维度获得分词数据的词组对应的信息；

[0078] 当比对相识度大于75％时，通过该当前文本数据的词组得到对应的信息，在电信行业该信息可以是指标数据或者报表数据。

[0079] 步骤S306，如果比对相识度大于45％且小于75％，将比对相识度最高的当前文本数据的词组对应的信息，作为分词数据的词组对应的信息；

[0080] 当比对相识度大于45％且小于75％时，返回比对相识度较高的几个(例如3个)词组，作为信息数据查询的关键词，通过这几个关键词来查询其对应的信息。

[0081] 步骤S308，如果比对相识度小于45％，从日常用语资料库中查询得到当前文本数据的词组对应的信息。

[0082] 当比对相识度小于45％时，需要从日常用语言资料库中查询词组对应的关键词，并得到该关键词对应的信息。

[0083] 本实施例提供的另一种语音数据的处理方法，首先通过语音识别技术可将输入的语音数据转为文本数据，实现对语音的识别；然后通过分词技术对文本进行分词，分词后通过去噪，相识度对比等处理，并结合数据源提供方提供的API接口，将文本信息解析成API接口的输入参数，通过调用接口返回数据结果；并通过GA遗传算法，将情景对话过程中的语意进行解析，更准确的理解用户的意图，例如环比、同比、累积等处理；最后还可通过ALS协同过滤算法，按用户喜好推荐不同的报表和指标，从而提高了信息数据的处理效率，增强了数据的智能化处理。

[0084] 为了更好地理解上述语音数据的处理方法，本实施例中描述一种具体的应用场景；该实施例中，以电信行业查找指标或者报表数据为例进行说明，该语音数据的处理方法

的流程图如图4所示。

[0085] 首先输入语音数据，在调用Deep Speech的API接口通过语音转义模块将语音数据转换为文本数据，然后利用jieba分词技术对文本数据进行分词，得到分词数据的词组。

[0086] 判断上述分词数据的词组是否有特殊语义，如果有特殊语义，需要对词组进行去噪处理，去掉与查找指标没有用的词组，留下关键性的词组；对关键性的词组与没有特殊语义的词组，判断是否与前一个文本数据的词组存在联系。

[0087] 如果存在联系，用遗传算法调出上一个文本数据的查询指标记录(相当于GA遗传算法遗传记忆的内容)，再利用自然语言处理指标维度语义的转换，判断该指标维度是否全部满足该指标数据查询的维度(时间、地点、指标名称)，若该指标数据的查询维度满足要求，则解析指标的编码(相当于分词词组的编码)；若缺少指标数据查询的维度，则校验缺少哪一个指标维度，返回缺少指标维度的问题，也就是用语音或者文本的方式输出指标维度的问题，以提醒用户输入指标维度；将指标维度满足要求和维度缺失信息补充后的词组进行组装会话；判断组装后输出的会话类型是否完整，如果不完整，直接返回客户端，返回数据，输出上述词组对应的信息；如果完整，则调用电信指标API服务接口，并判断是否需要输入同比、环比、累计、方差等数据，如果需要输入则计算指标的同比、环比、累计、方差等数据，并根据用户需求判断数据展示的类型：折线、饼图或者柱形等，并返回客户端，如果不需要输入，则将需要输出的指标数据以用户所需的类型进行展示，并返回客户端。

[0088] 如果存在联系，则利用分词词组最短距离算法和信用度算法计算分词词组与标准数据库的比对相识度，其中，比对相识度大于75％时，直接获取该指标对应的分词的编码，并判断该指标维度是否满足要求，后续流程与上述判断过程一致；比对相识度大于45％小于75％时，获取指标对应的分词的编码和指标名称的列表，并组装会话，返回客户端；比对相识度小于45％时，切换到日常用语训练库中，查找对应的词组，并组装会话，返回客户端。

[0089] 对分词去噪后的词频数组，按照预先的定义的维度解析出该指标的关键字段，其中预先设定的维度都有各自的编码，例如，输入语义“帮我查询下，南京5月份的天翼发展量是多少”，解析地区“南京:025”，月份“5月:201805”，指标名称“天翼发展量:A10023”，封装后的JSON((JavaScript Object Notation,JS对象简谱))如下所示：

```
"re_data": {
    "input_value": "itv 用户数",--语音入参
    "response": {
        "indicatorsCode": "M1IU3418U",  --指标编码
        "condition": {
```

[0090]

```
                "show_type": [],  --显示类型 默认柱形图，支持柱形、饼图、
        折线图
                "city": [       --地区
                    "801"
                ],
               "account_time": [   --账期
                    "201807"
                ]
            },
[0091]      "finish": "success",
            "confidence": 0.8,  ---相识度
            "indicatorsName": "ITV 用户数(月)",
            "indicatorsType": "ZB",  --类型
            "describe": "ITV（期末到达）",--指标描述
            "occurrence": 136,--信用度
            "compute_type": "1",--计算类型 累积，同步，环比，方差等
            "format_input_value": "itv 用户数",
        }
```

[0092] 在电信行业的智能数据应用中，采用聊天对话问答的方式进行交互。场景分两类，一类是电信行业专业术语对话，一类是日常用语对话。专业术语的语料库与日常用语的语料库存储在MongoDB数据库中。在聊天对话时，当输入的语音文本相识度大于75％时去专业术语语料库(相当于上述标准数据库)中查询，当相识度低于45％时去日常用语语料库(日常用语资料库)查询。

[0093] 对话方式如下：

[0094] A:“宽带新增发展量”

[0095] S:请问要查询哪个地区？

[0096] A:南京

[0097] S:请问要查询哪个时间？

[0098] A:八月份

[0099] S:显示8月份南京的宽带发展数据

[0100] 在问答过程中如果对问题回答偏移比较大，连问3次后，不在遗传上个问题，转到日常百科进行处理，处理场景如下：

[0101] A:“宽带新增发展量”

[0102] S:请问要查询哪个地区?

[0103] A:南京

[0104] S:请问要查询哪个时间?

[0105] A:南京

[0106] S:请问要查询哪个时间?

[0107] A:南京

[0108] S:请问要查询哪个时间?

[0109] A:南京

[0110] S:南京,简称“宁”,古称金陵、建康,是江苏省会、副省级市、南京都市圈核心城市,国务院批复确定的中国东部地区重要的中心城市、全国重要的科研教育基地和综合交通枢纽。

[0111] 上述MongoDB数据库,通常为一个基于分布式文件存储的数据库。由C++语言编写。旨在为WEB(World Wide Web,全球广域网)应用提供可扩展的高性能数据存储解决方案,在智能数据应用中采用此数据库存储语料信息。

[0112] 该方法可根据ALS算法根据用户的喜好推荐指标,也就是可根据对话产生的日志作为训练数据集,生成用户访问次数的数据结构,用户访问次数的多少决定数据是否是热数据。使用UserFC进行推荐,输入:文件名,用户ID,邻居数量。通过数据的处理得到用户喜好的报表,并在登录手机App后显示推荐的指标和报表信息。

[0113] 本实施例提供的另一种语音数据的处理方法可通过语音识别技术解决复杂的查询报表操作,可以在App上按语音进行直接智能语音查询,不需要复杂的查找与输入操作;还可通过最短距离相识算法和查询信用度计算相识度,比如“帮我查询下昨天的南京、无锡、4G天翼发展量”与指标名“4G天翼发展量”、“3G天翼发展量”、“4G天翼用户数”等指标名进行相识度匹配,同时匹配该指标出现的occurrence信用度大小,相识度越大并且信用度越高说明越贴近指标;并通过分词技术对文本进行语法分析提取关键信息,如“帮我查询下昨天的南京、无锡、4G天翼发展量”进行分词,对分词后的词组进行地区和时间维度的判断,识别出地区和时间后转为系统可以调用的编码,如“南京:025”,“无锡:0510”,“昨天:20180805”。

[0114] 同时,通过GA算法解决查询的语义遗传和变异的问题,当查询一个指标后例如说:“4月份南京天翼用户发展量”,返回结果后,直接说“帮我查询下3月份”,通过GA算法可以正确识别出3月份南京天翼用户发展量数据,可以通过语音分析出同比、环比、增量,例如“查询下3月份4月份的同比”或“3月份4月份用户增量”,最后通过ALS算法分析用户的查询报表或指标的喜好,不同的用户登录到系统后查询的报表或指标,会按ITEM推荐个用户相同喜好的指标或报表。

[0115] 本发明实施例通过对语音解析后的文本检测,对识别结果进行分词,句法分析、文本分类、结合电信行业的专业术语如4G、天翼、流量、话费、等术语进行语义解析,组成查询指标或KPI的API调用接口信息,通过调用接口返回的结果动态显示文本、图表、并支持语音播报,将语音分析与智能数据相结合,方便的数据的查询和展示。

[0116] 对应于上述方法实施例,参见图5所示的一种语音数据的处理装置的结构示意图,该装置设置于服务器,所述服务器与客户端通信连接,该装置包括:

[0117] 数据获取模块50,用于获取当前的语音数据;

[0118] 语音转换模块51,用于将当前的语音数据转换为当前文本数据;

[0119] 分词模块52,用于对上述当前文本数据进行自然语言分词,得到分词数据;

[0120] 特殊语义判断模块53,用于判断分词数据是否有特殊语义,如果有特殊语义,对该分词数据的语义进行去噪处理;

[0121] 去噪模块54,用于判断去噪后的分词数据的语义或者当前文本数据的语义与前一个文本数据的语义是否有联系;

[0122] 信息获取模块55,用于如果有联系,根据前一个文本数据的语义对应的信息,获取当前文本数据的语义对应的信息;

[0123] 信息获取模块55,还用于如果没有联系,将分词数据的词组与预设的标准数据库中的词组进行比对,根据比对结果确定分词数据的词组对应的信息;该标准数据库中保存有预先设定的词组;

[0124] 信息返回模块56,用于将当前文本数据的语义对应的信息或者分词数据的词组对应的信息返回至客户端。

[0125] 进一步地,上述装置还包括:信息推荐模块,用于通过ALS算法分析用户信息查询的喜好,以使不同的用户在登录客户端后,获得不同的推荐查询的信息。

[0126] 进一步地,语音转换模块51,还用于通过调用Deep Speech的API接口,将当前的语音数据转换为当前文本数据。

[0127] 本发明实施例提供的语音数据的处理装置,与上述实施例提供的语音数据的处理方法具有相同的技术特征,所以也能解决相同的技术问题,达到相同的技术效果。

[0128] 本实施例还提供了一种与上述方法实施例相对应的一种语音数据的处理系统,该系统包括存储器以及处理器,该存储器用于存储支持处理器执行语音数据的处理方法的程序,该处理器被配置为用于执行存储器中存储的程序。

[0129] 本发明实施例所提供的语音数据的处理方法、装置和系统,该技术通过分词技术解析语义中的关键维度:包括,地区、时间、指标/报表编码、同比、环比、累积等关键字;通过GA算法,对上下文的语意进行交叉变异,包括对上个问句的指标名、地区、时间等关键字段;通过ALS算法和余弦相识度算法,对电信行业的业务指标或报表进行推荐;并基于最短路径算法,对短语音文本字段进行匹配,采用信用度提高相识度;采用MongoDB分布式数据库存储电信特有的语料信息和日常训练语料信息;数据查询方式,采用聊天对话方式,展现结果包括,数字、图表、表格等信息;而且可以通过数值和图表展示在智能终端上显示数据,并可以语音播报,从而方便的用户的使用,也提高了语音数据的提取效率。

[0130] 本发明实施例所提供的语音数据的处理方法、装置和系统的计算机程序产品,包括存储了程序代码的计算机可读存储介质,所述程序代码包括的指令可用于执行前面方法实施例中所述的方法,具体实现可参见方法实施例,在此不再赘述。

[0131] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统和/或装置的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0132] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时,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,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案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)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、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。

[0133] 最后应说明的是:以上所述实施例,仅为本发明的具体实施方式,用以说明本发明的技术方案,而非对其限制,本发明的保护范围并不局限于此,尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,其依然可以对前述实施例所记载的技术方案进行修改或可轻易想到变化,或者对其中部分技术特征进行等同替换;而这些修改、变化或者替换,并不使相应技术方案的本质脱离本发明实施例技术方案的精神和范围,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应所述以权利要求的保护范围为准。

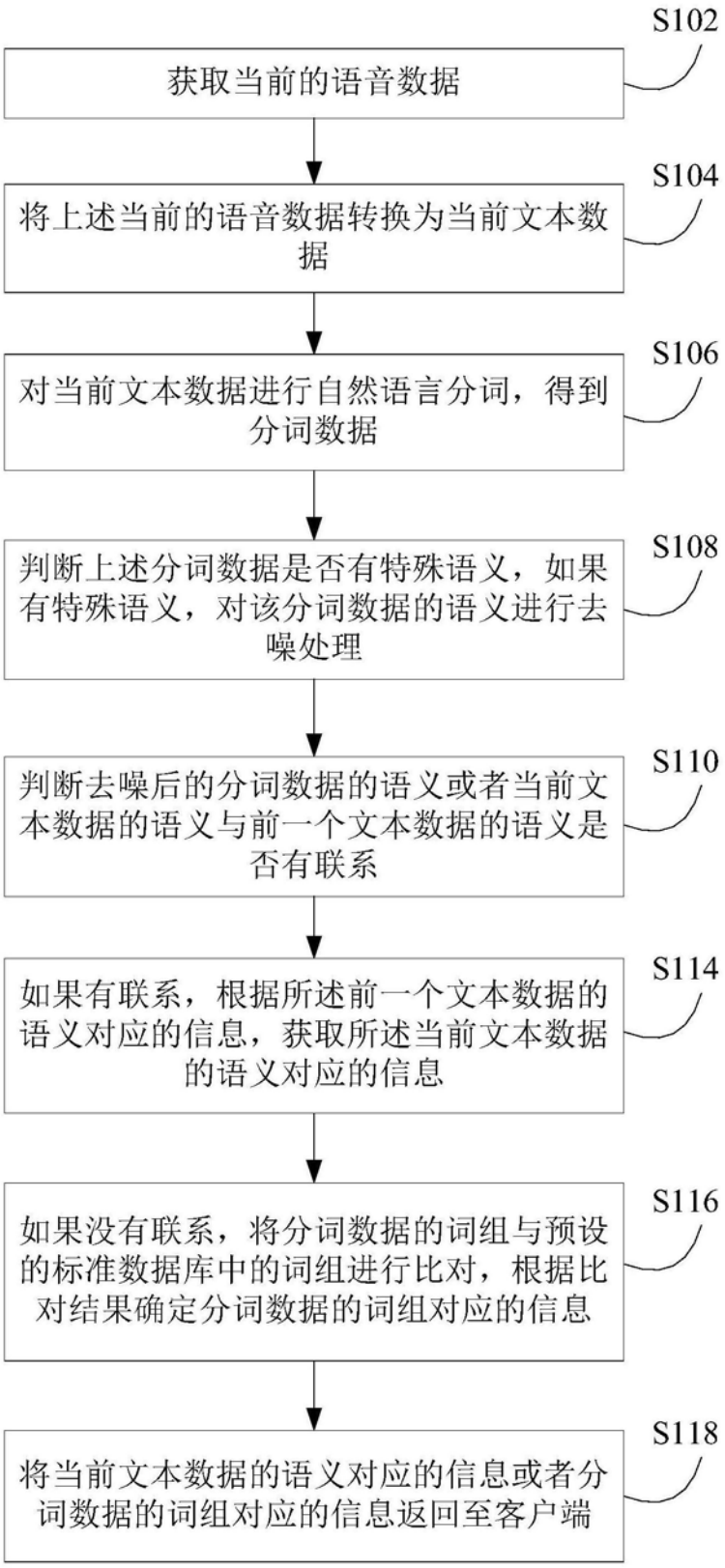
S102
获取当前的语音数据
S104
将上述当前的语音数据转换为当前文本数据
S106
对当前文本数据进行自然语言分词，得到分词数据
S108
判断上述分词数据是否有特殊语义，如果有特殊语义，对该分词数据的语义进行去噪处理
S110
判断去噪后的分词数据的语义或者当前文本数据的语义与前一个文本数据的语义是否有联系
S114
如果有联系，根据所述前一个文本数据的语义对应的信息，获取所述当前文本数据的语义对应的信息
S116
如果没有联系，将分词数据的词组与预设的标准数据库中的词组进行比对，根据比对结果确定分词数据的词组对应的信息
S118
将当前文本数据的语义对应的信息或者分词数据的词组对应的信息返回至客户端

图1

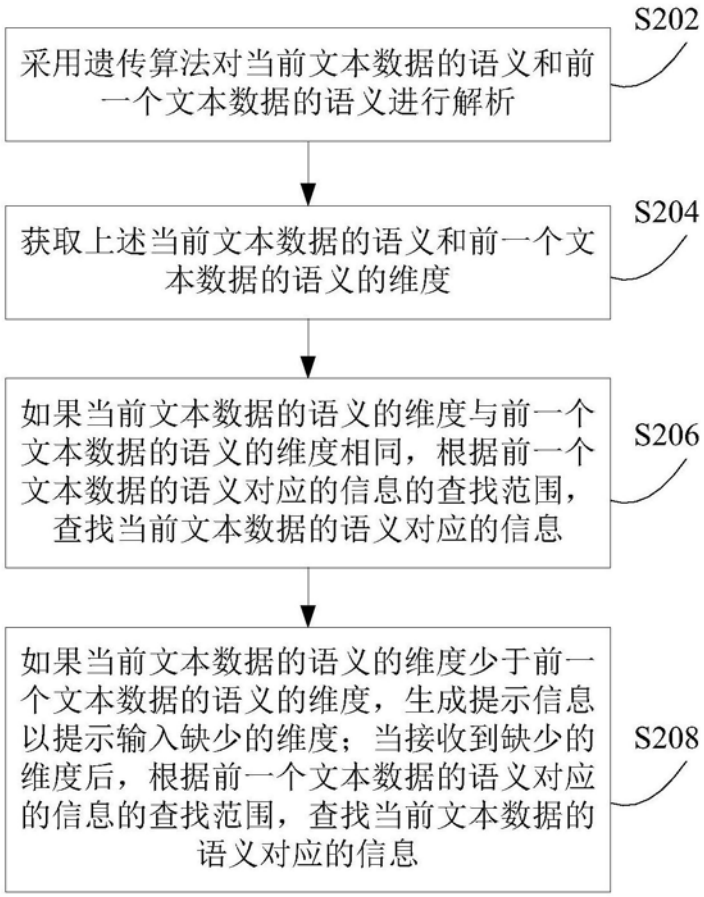
S202
采用遗传算法对当前文本数据的语义和前一个文本数据的语义进行解析
S204
获取上述当前文本数据的语义和前一个文本数据的语义的维度
S206
如果当前文本数据的语义的维度与前一个文本数据的语义的维度相同，根据前一个文本数据的语义对应的信息的查找范围，查找当前文本数据的语义对应的信息
S208
如果当前文本数据的语义的维度少于前一个文本数据的语义的维度，生成提示信息以提示输入缺少的维度；当接收到缺少的维度后，根据前一个文本数据的语义对应的信息的查找范围，查找当前文本数据的语义对应的信息

图2

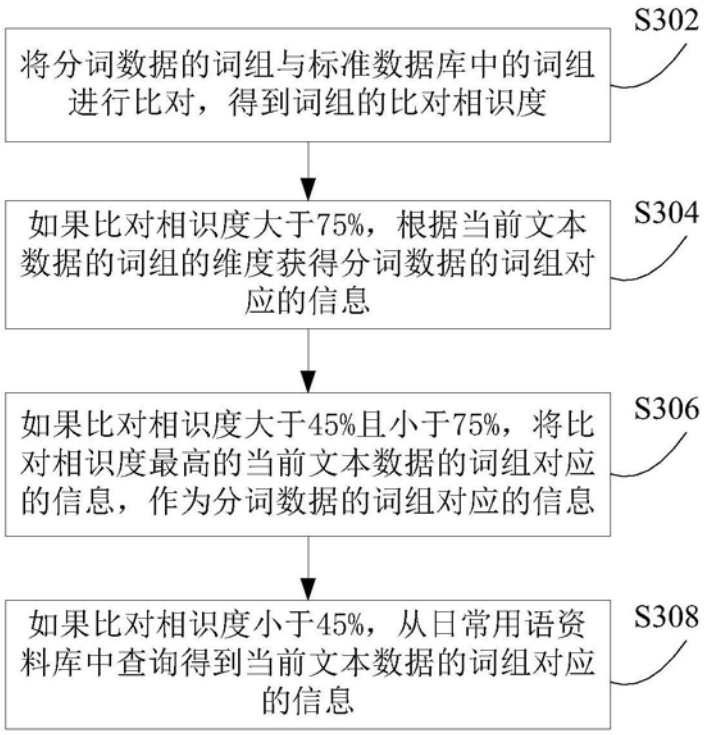
S302
将分词数据的词组与标准数据库中的词组进行比对，得到词组的比对相识度
S304
如果比对相识度大于75%，根据当前文本数据的词组的维度获得分词数据的词组对应的信息
S306
如果比对相识度大于45%且小于75%，将比对相识度最高的当前文本数据的词组对应的信息，作为分词数据的词组对应的信息
S308
如果比对相识度小于45%，从日常用语资料库中查询得到当前文本数据的词组对应的信息

图3

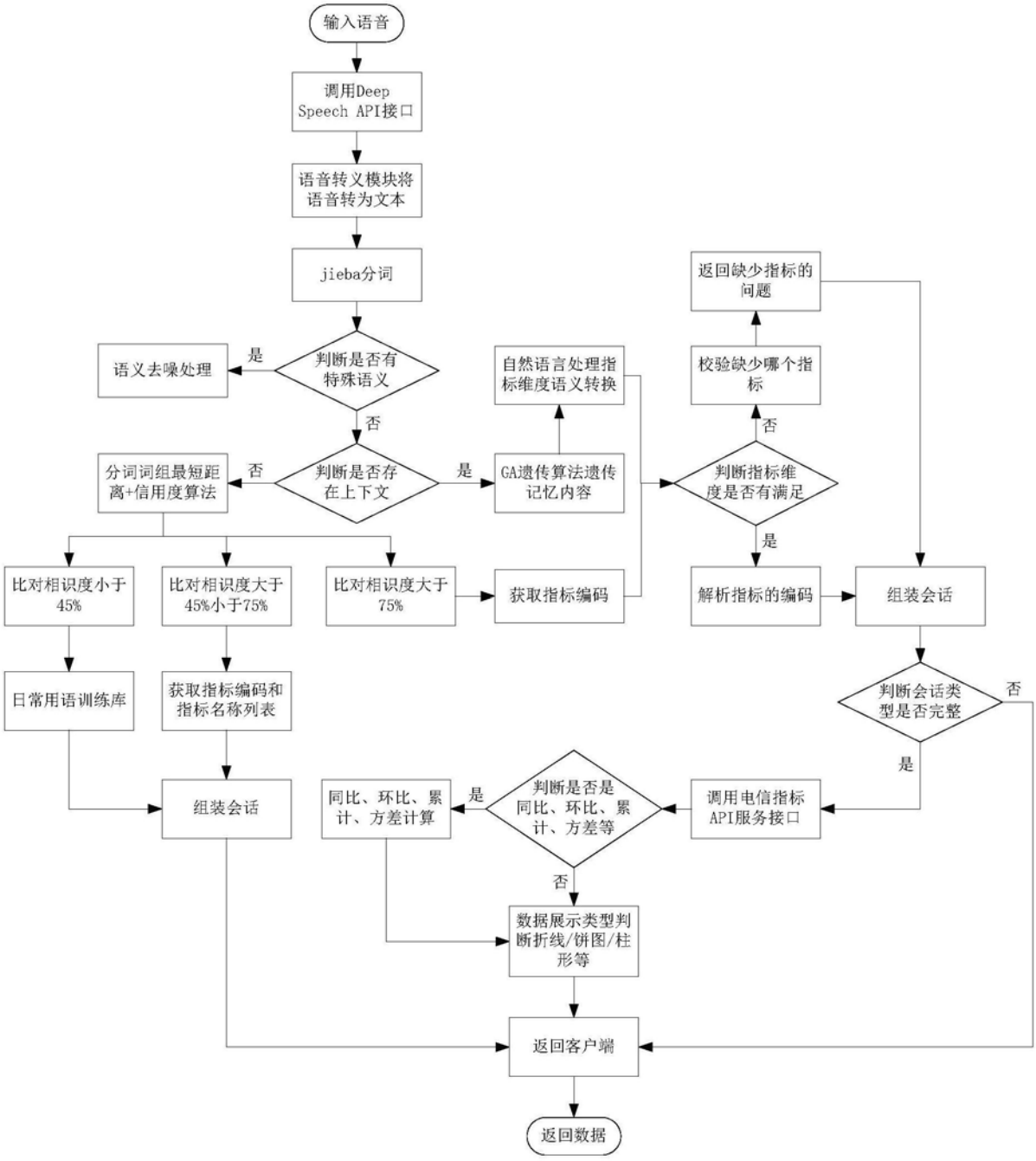
输入语音
调用Deep Speech API接口
语音转义模块将语音转为文本
jieba分词
判断是否有特殊语义
是
语义去噪处理
否
判断是否存在上下文
是
GA遗传算法遗传记忆内容
否
分词词组最短距离+信用度算法
比对相识度小于45%
比对相识度大于45%小于75%
比对相识度大于75%
获取指标编码
自然语言处理指标维度语义转换
判断指标维度是否有满足
否
校验缺少哪个指标
返回缺少指标的问题
是
解析指标的编码
组装会话
判断会话类型是否完整
否
是
日常用语训练库
获取指标编码和指标名称列表
组装会话
同比、环比、累计、方差计算
是
判断是否是同比、环比、累计、方差等
调用电信指标API服务接口
否
数据展示类型判断折线/饼图/柱形等
返回客户端
返回数据

图4

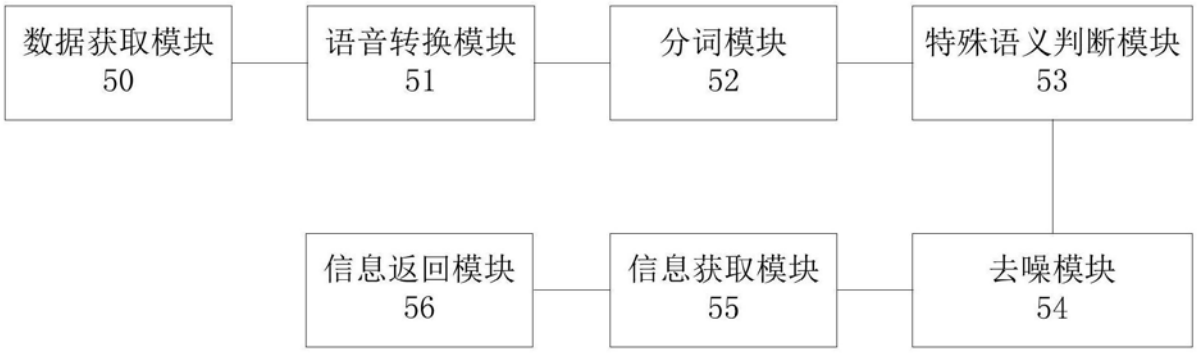
数据获取模块
50
语音转换模块
51
分词模块
52
特殊语义判断模块
53
去噪模块
54
信息获取模块
55
信息返回模块
56

图5