(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 111737499 B
(45)授权公告日 2020.11.27

(21)申请号 202010727532.0

(22)申请日 2020.07.27

(65)同一申请的已公布的文献号
申请公布号 CN 111737499 A

(43)申请公布日 2020.10.02

(73)专利权人 平安国际智慧城市科技股份有限公司
地址 518000 广东省深圳市前海深港合作区妈湾兴海大道3048号前海自贸大厦1-34层

(72)发明人 袁小力

(74)专利代理机构 深圳市赛恩倍吉知识产权代理有限公司 44334
代理人 何春兰 迟珊珊

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 16/332(2019.01)
G06F 16/33(2019.01)

(56)对比文件
CN 102713865 A,2012.10.03
CN 110727930 A,2020.01.24
CN 110008234 A,2019.07.12
CN 111416789 A,2020.07.14
CN 107480551 A,2017.12.15

审查员 邝小青

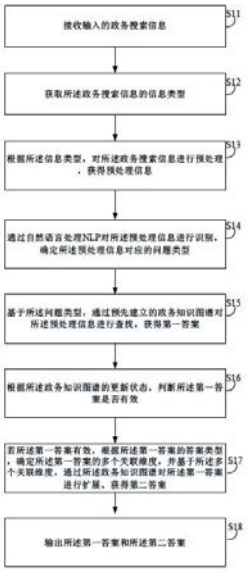
权利要求书2页 说明书12页 附图2页

(54)发明名称

基于自然语言处理的数据搜索方法及相关设备

(57)摘要

本发明涉及人工智能技术领域，提供一种基于自然语言处理的数据搜索方法，包括：接收政务搜索信息；获取政务搜索信息的信息类型；根据信息类型，对政务搜索信息进行预处理，获得预处理信息；对预处理信息进行识别，确定问题类型；基于问题类型，通过政务知识图谱对预处理信息进行查找，获得第一答案；根据政务知识图谱的更新状态，判断第一答案是否有效；若有效，根据第一答案的答案类型，确定第一答案的多个关联维度，并基于多个关联维度，通过政务知识图谱对第一答案进行扩展，获得第二答案；输出第一答案和第二答案。本发明还涉及区块链技术，可以将第一答案和第二答案上传至区块链。本发明应用于智慧政务场景，从而推动智慧城市的发展。



CN 111737499 B

1.一种基于自然语言处理的数据搜索方法,其特征在于,所述基于自然语言处理的数据搜索方法包括:

接收输入的政务搜索信息;

获取所述政务搜索信息的信息类型;

根据所述信息类型,对所述政务搜索信息进行预处理,获得预处理信息;

通过自然语言处理NLP对所述预处理信息进行识别,确定所述预处理信息对应的问题类型;

获取所述问题类型的机密级别,以及获取当前输入用户的用户级别;

判断所述用户级别与所述机密级别是否匹配;

若所述用户级别与所述机密级别匹配,基于所述问题类型,通过预先建立的政务知识图谱对所述预处理信息进行查找,获得第一答案;

根据所述政务知识图谱的更新状态,判断所述第一答案是否有效;

当所述第一答案有效时,获取所述输入用户的用户信息,并根据所述用户信息判断所述输入用户是否具备扩展答案的权限;

当所述输入用户具备扩展答案的权限时,根据所述第一答案的答案类型,确定所述第一答案的多个关联维度,并基于所述多个关联维度,通过所述政务知识图谱对所述第一答案进行扩展,获得第二答案;

输出所述第一答案和所述第二答案。

2.根据权利要求1所述的基于自然语言处理的数据搜索方法,其特征在于,所述根据所述信息类型,对所述政务搜索信息进行预处理,获得预处理信息包括:

若所述信息类型为文本类型,判断所述政务搜索信息中是否存在错误文字;

若所述政务搜索信息中存在错误文字,根据编辑距离算法从预设词库中查找与所述错误文字相似度高的第一文字;

使用所述第一文字替换所述错误文字,并将替换后的政务搜索信息确定为预处理信息。

3.根据权利要求1所述的基于自然语言处理的数据搜索方法,其特征在于,所述根据所述信息类型,对所述政务搜索信息进行预处理,获得预处理信息包括:

若所述信息类型为图片类型,采用图像去模糊算法,对所述政务搜索信息进行图片处理,获得第一图片;

若所述第一图片存在边缘无关信息,删除所述边缘无关信息,并将删除后的第一图片确定为预处理信息。

4.根据权利要求1所述的基于自然语言处理的数据搜索方法,其特征在于,所述根据所述信息类型,对所述政务搜索信息进行预处理,获得预处理信息包括:

若所述信息类型为语音类型,识别所述政务搜索信息所属的区域口音类型;

基于所述区域口音类型,对所述政务搜索信息进行语音纠正处理,并将处理后的政务搜索信息确定为预处理信息。

5.根据权利要求1所述的基于自然语言处理的数据搜索方法,其特征在于,所述基于自然语言处理的数据搜索方法还包括:

通过网络爬虫技术,从各个政务网站获取政务信息;

从所述政务信息中确定多个实体，并基于所述多个实体的标签，分析所述多个实体的关联关系；

根据所述多个实体以及所述关联关系，建立政务知识图谱。

6.根据权利要求1所述的基于自然语言处理的数据搜索方法，其特征在于，所述输出所述第一答案和所述第二答案包括：

获取与所述信息类型匹配的输出模式；

若所述输出模式为文字模式，获取所述政务搜索信息的文字属性，采用所述文字属性输出所述第一答案和所述第二答案；或

若所述输出模式为图片模式，获取与所述答案类型匹配的图形模板，采用所述图形模板对所述第一答案和所述第二答案进行可视化展示；或

若所述输出模式为语音模式，将所述第一答案和所述第二答案转换为与所述政务搜索信息的口音匹配的第二语音，并输出所述第二语音。

7.一种数据搜索装置，其特征在于，所述数据搜索装置包括：

接收模块，用于接收输入的政务搜索信息；

获取模块，用于获取所述政务搜索信息的信息类型；

处理模块，用于根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息；

确定模块，用于通过自然语言处理NLP对所述预处理信息进行识别，确定所述预处理信息对应的问题类型，获取所述问题类型的机密级别，以及获取当前输入用户的用户级别；判断所述用户级别与所述机密级别是否匹配；

查找模块，用于若所述用户级别与所述机密级别匹配，基于所述问题类型，通过预先建立的政务知识图谱对所述预处理信息进行查找，获得第一答案；

判断模块，用于根据所述政务知识图谱的更新状态，判断所述第一答案是否有效，当所述第一答案有效时，获取所述输入用户的用户信息，并根据所述用户信息判断所述输入用户是否具备扩展答案的权限；

所述确定模块，还用于当所述输入用户具备扩展答案的权限时，根据所述第一答案的答案类型，确定所述第一答案的多个关联维度，并基于所述多个关联维度，通过所述政务知识图谱对所述第一答案进行扩展，获得第二答案；输出模块，用于输出所述第一答案和所述第二答案。

8.一种电子设备，其特征在于，所述电子设备包括处理器和存储器，所述处理器用于执行存储器中存储的计算机程序以实现如权利要求1至6中任意一项所述的基于自然语言处理的数据搜索方法。

9.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有至少一个指令，所述至少一个指令被处理器执行时实现如权利要求1至6任意一项所述的基于自然语言处理的数据搜索方法。

基于自然语言处理的数据搜索方法及相关设备

技术领域

[0001] 本发明涉及人工智能技术领域，尤其涉及一种基于自然语言处理的数据搜索方法及相关设备。

背景技术

[0002] 目前，从互联网上进行信息搜索已经成为人们获取信息的一种趋势。然而，实践中发现，互联网输出的通常是一些相关的网页，由于互联网上的数据量巨大，而且数据繁杂，需要用户一个个去点击并查找相关答案，有时候由于关键词错误，还需要多次搜索，这无疑浪费大量的时间和精力，搜索效率较低，而且搜索的准确度也较低。

发明内容

[0003] 鉴于以上内容，有必要提供一种基于自然语言处理的数据搜索方法及相关设备，能够提高搜索效率以及搜索的准确度。

[0004] 本发明的第一方面提供一种基于自然语言处理的数据搜索方法，所述基于自然语言处理的数据搜索方法包括：

[0005] 接收输入的政务搜索信息；

[0006] 获取所述政务搜索信息的信息类型；

[0007] 根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息；

[0008] 通过自然语言处理NLP对所述预处理信息进行识别，确定所述预处理信息对应的问题类型；

[0009] 基于所述问题类型，通过预先建立的政务知识图谱对所述预处理信息进行查找，获得第一答案；

[0010] 根据所述政务知识图谱的更新状态，判断所述第一答案是否有效；

[0011] 若所述第一答案有效，根据所述第一答案的答案类型，确定所述第一答案的多个关联维度，并基于所述多个关联维度，通过所述政务知识图谱对所述第一答案进行扩展，获得第二答案；

[0012] 输出所述第一答案和所述第二答案。

[0013] 在一种可能的实现方式中，所述根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息包括：

[0014] 若所述信息类型为文本类型，判断所述政务搜索信息中是否存在错误文字；

[0015] 若所述政务搜索信息中存在错误文字，根据编辑距离算法从预设词库中查找与所述错误文字相似度高的第一文字；

[0016] 使用所述第一文字替换所述错误文字，并将替换后的政务搜索信息确定为预处理信息。

[0017] 在一种可能的实现方式中，所述根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息包括：

[0018] 若所述信息类型为图片类型，采用图像去模糊算法，对所述政务搜索信息进行图片处理，获得第一图片；

[0019] 若所述第一图片存在边缘无关信息，删除所述边缘无关信息，并将删除后的第一图片确定为预处理信息。

[0020] 在一种可能的实现方式中，所述根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息包括：

[0021] 若所述信息类型为语音类型，识别所述政务搜索信息所属的区域口音类型；

[0022] 基于所述区域口音类型，对所述政务搜索信息进行语音纠正处理，并将处理后的政务搜索信息确定为预处理信息。

[0023] 在一种可能的实现方式中，所述通过自然语言处理NLP对所述预处理信息进行识别，确定所述预处理信息对应的问题类型之后，所述基于自然语言处理的数据搜索方法还包括：

[0024] 获取所述问题类型的机密级别，以及获取当前输入用户的用户级别；

[0025] 判断所述用户级别与所述机密级别是否匹配；

[0026] 若所述用户级别与所述机密级别匹配，基于所述问题类型，通过预先建立的政务知识图谱对所述预处理信息进行查找，获得第一答案。

[0027] 在一种可能的实现方式中，所述基于自然语言处理的数据搜索方法还包括：

[0028] 通过网络爬虫技术，从各个政务网站获取政务信息；

[0029] 从所述政务信息中确定多个实体，并基于所述多个实体的标签，分析所述多个实体的关联关系；

[0030] 根据所述多个实体以及所述关联关系，建立政务知识图谱。

[0031] 在一种可能的实现方式中，所述输出所述第一答案和所述第二答案包括：

[0032] 获取与所述信息类型匹配的输出模式；

[0033] 若所述输出模式为文字模式，获取所述政务搜索信息的文字属性，采用所述文字属性输出所述第一答案和所述第二答案；或

[0034] 若所述输出模式为图片模式，获取与所述答案类型匹配的图形模板，采用所述图形模板对所述第一答案和所述第二答案进行可视化展示；或

[0035] 若所述输出模式为语音模式，将所述第一答案和所述第二答案转换为与所述政务搜索信息的口音匹配的第二语音，并输出所述第二语音。

[0036] 本发明的第二方面提供一种数据搜索装置，所述数据搜索装置包括：

[0037] 接收模块，用于接收输入的政务搜索信息；

[0038] 获取模块，用于获取所述政务搜索信息的信息类型；

[0039] 处理模块，用于根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息；

[0040] 确定模块，用于通过自然语言处理NLP对所述预处理信息进行识别，确定所述预处理信息对应的问题类型；

[0041] 查找模块，用于基于所述问题类型，通过预先建立的政务知识图谱对所述预处理信息进行查找，获得第一答案；

[0042] 判断模块，用于根据所述政务知识图谱的更新状态，判断所述第一答案是否有效；

[0043] 所述确定模块，还用于若所述第一答案有效，根据所述第一答案的答案类型，确定所述第一答案的多个关联维度，并基于所述多个关联维度，通过所述政务知识图谱对所述第一答案进行扩展，获得第二答案；

[0044] 输出模块，用于输出所述第一答案和所述第二答案。

[0045] 本发明的第三方面提供一种电子设备，所述电子设备包括处理器和存储器，所述处理器用于执行所述存储器中存储的计算机程序时实现所述的基于自然语言处理的数据搜索方法。

[0046] 本发明的第四方面提供一种计算机可读存储介质，所述计算机可读存储介质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所述的基于自然语言处理的数据搜索方法。

[0047] 本发明中，根据政务搜索信息的信息类型对政务搜索信息进行预处理，通过NLP进行语义识别，获得用户的真实意图，并基于政务知识图谱搜索用户想到的答案，同时，对答案进行关联搜索，可以给用户提供更多的关联信息，不仅可以提高搜索的效率和搜索的准确度，同时，还可以提高用户搜索的满意度，提高用户体验。

附图说明

[0048] 图1是本发明公开的一种基于自然语言处理的数据搜索方法的较佳实施例的流程图。

[0049] 图2是本发明公开的一种数据搜索装置的较佳实施例的功能模块图。

[0050] 图3是本发明实现基于自然语言处理的数据搜索方法的较佳实施例的电子设备的结构示意图。

具体实施方式

[0051] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。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。

[0052] 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0053] 需要说明的是，在本发明中涉及“第一”、“第二”等的描述仅用于描述目的，而不能理解为指示或暗示其相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。另外，各个实施例之间的技术方案可以相互结合，但是必须是以本领域普通技术人员能够实现为基础，当技术方案的结合出现相互矛盾或无法实现时应当认为这种技术方案的结合不存在，也不在本发明要求

的保护范围之内。

[0054] 其中，所述电子设备是一种能够按照事先设定或存储的指令，自动进行数值计算和/或信息处理的设备，其硬件包括但不限于微处理器、专用集成电路(ASIC)、现场可编程门阵列(FPGA)、数字信号处理器(DSP)、嵌入式设备等。所述电子设备还可包括网络设备和/或用户设备。其中，所述网络设备包括但不限于单个网络服务器、多个网络服务器组成的服务器组或基于云计算(Cloud Computing)的由大量主机或网络服务器构成的云。所述用户设备包括但不限于任何一种可与用户通过键盘、鼠标、遥控器、触摸板或声控设备等方式进行人机交互的电子产品，例如，个人计算机、平板电脑、智能手机、个人数字助理PDA等。

[0055] 请参见图1，图1是本发明公开的一种基于自然语言处理的数据搜索方法的较佳实施例的流程图。其中，根据不同的需求，该流程图中步骤的顺序可以改变，某些步骤可以省略。

[0056] S11、接收输入的政务搜索信息。

[0057] S12、获取所述政务搜索信息的信息类型。

[0058] 其中，所述信息类型可以为文字类型，也可以为图片类型，还可以是语音类型，本发明实施例不做限定。

[0059] S13、根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息。

[0060] 具体的，所述根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息包括：

[0061] 若所述信息类型为文本类型，判断所述政务搜索信息中是否存在错误文字；

[0062] 若所述政务搜索信息中存在错误文字，根据编辑距离算法从预设词库中查找与所述错误文字相似度高的第一文字；

[0063] 使用所述第一文字替换所述错误文字，并将替换后的政务搜索信息确定为预处理信息。

[0064] 在该可选的实施方式中，当用户输入的政务搜索信息出现单词拼写错误、同音字时，搜索系统可以自动识别纠错。汉字的纠错使用编辑距离算法实现，首先在预设词库(通常是同一行业领域的词库)中查找词相似度以及拼音相似度较高的词作为候选词，以缩小编辑距离计算范围，之后计算各个候选词与需要纠错词的编辑距离，取编辑距离最小的候选词，如果编辑距离值超过设置的纠错阀值，则作为结果返回。通过这种方式可以对政务搜索信息纠正，获得更加准确的搜索信息，有利于提高对用户意图识别的准确率。

[0065] 具体的，所述根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息包括：

[0066] 若所述信息类型为图片类型，采用图像去模糊算法，对所述政务搜索信息进行图片处理，获得第一图片；

[0067] 若所述第一图片存在边缘无关信息，删除所述边缘无关信息，并将删除后的第一图片确定为预处理信息。

[0068] 在该可选的实施方式中，当信息类型为图片类型时，通过对政务搜索信息进行图片处理以及删除边缘无关信息，不仅可以提高图片的清晰度，同时，也减少了图片的冗余信息，减少需要识别的信息量，从而可以提高对用户意图识别的准确率。

[0069] 具体的，所述根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信

息包括：

[0070] 若所述信息类型为语音类型，识别所述政务搜索信息所属的区域口音类型；

[0071] 基于所述区域口音类型，对所述政务搜索信息进行语音纠正处理，并将处理后的政务搜索信息确定为预处理信息。

[0072] 在该可选的实施方式中，不同区域的用户在输入语音时，会自带本区域的口音，由于口音的区域化，会导致识别困难，通过基于所述区域口音类型，对所述政务搜索信息进行语音纠正处理，可以获得符合标准的普通话，进而可以准确识别出用户的意图，从而可以提高对用户意图识别的准确率。

[0073] S14、通过自然语言处理NLP对所述预处理信息进行识别，确定所述预处理信息对应的问题类型。

[0074] 其中，NLP (Natural Language Processing，自然语言处理)可以对预处理信息进行识别，获得该预处理信息对应的问题类型。

[0075] 其中，可以将所述预处理信息归类为多组问题类型。

[0076] 问题类型1：“人物属性- 职称- 层级- 学校- 标签- 机构- 处室- 地理”，举例来说，XXX 机构有哪些部长，XXX 机构XXX 部门的部长是谁。

[0077] 问题类型2：“机构- 地理- 时间- 预算值”，举例来说，例如XXX 机构2018年的预算是多少，预算最多的机构是多少。

[0078] 可选的，所述通过自然语言处理NLP对所述预处理信息进行识别，确定所述预处理信息对应的问题类型之后，所述方法还包括：

[0079] 获取所述问题类型的机密级别，以及获取当前输入用户的用户级别；

[0080] 判断所述用户级别与所述机密级别是否匹配；

[0081] 若所述用户级别与所述机密级别匹配，基于所述问题类型，通过预先建立的政务知识图谱对所述预处理信息进行查找，获得第一答案。

[0082] 在该可选的实施方式中，由于用户是在搜索政务相关的信息，而这类信息通常是比较机密的，不是所有人均可以进行搜索的。通过将机密级别与用户级别进行匹配判断，可以对输入用户进行身份校验，如果匹配，表明输入用户属于合法用户，具备搜索的权限。通过这种用户身份的校验，可以确保信息的安全性。

[0083] S15、基于所述问题类型，通过预先建立的政务知识图谱对所述预处理信息进行查找，获得第一答案。

[0084] 可选的，所述方法还包括：

[0085] 通过网络爬虫技术，从各个政务网站获取政务信息；

[0086] 从所述政务信息中确定多个实体，并基于所述多个实体的标签，分析所述多个实体的关联关系；

[0087] 根据所述多个实体以及所述关联关系，建立政务知识图谱。

[0088] 在该可选的实施方式中，可以通过大量的规则和爬虫技术对全国各省市政府官网，财政厅官网，政府招投标等相关网站进行定时信息爬取。

[0089] 举例来说，假设A 市市长和B 市市长都曾经就读于北京大学，那么北京大学就是一个实体(entity)，其中一个关系(relation)是就读人员，这个关系关联到另一些实体就是A 市市长和B市市长。但由于政府领域的特殊性，领导班子可能换届，但岗位始终存在，所

以我们记录岗位以及对应的人。每个实体有特定的标签，比如机构、地理、职能、属性。以此类推，可以构建出大量实体和关系庞大的政务知识图谱。

[0090] S16、根据所述政务知识图谱的更新状态，判断所述第一答案是否有效。

[0091] 其中，如果所述政务知识图谱的更新状态表明所述政务知识图谱内的数据均是最新的数据，则可以确定所述第一答案是有效的，反正，如果所述政务知识图谱的更新状态表明所述政务知识图谱内的数据不是最新的数据，即所述政务知识图谱很久没有更新了，则可以确定所述第一答案是无效的。

[0092] 通过这种方式，既可以确保查询到的第一答案与所述预处理信息匹配，同时，还能确保第一答案对用户当前的查询来说是属于有效的答案。

[0093] S17、若所述第一答案有效，根据所述第一答案的答案类型，确定所述第一答案的多个关联维度，并基于所述多个关联维度，通过所述政务知识图谱对所述第一答案进行扩展，获得第二答案。

[0094] 其中，关联维度可以包括但不限于时间、地点、区域、人物等。

[0095] 举例来说，所述第一答案的答案类型为人物，则可以确定所述第一答案的多个关联维度可以为时间、区域，基于关联维度可以查找到这个人物的出生年月，这个人物出生在哪里，目前在哪里居住等等。

[0096] 又举例来说，所述第一答案的答案类型为某个区域，则可以确定所述第一答案的关联维度可以为时间，基于关联维度可以查找到这个区域近几年的预算情况。

[0097] 可选的，若所述第一答案有效，所述方法还包括：

[0098] 获取当前输入用户的用户信息；

[0099] 根据所述用户信息，判断所述输入用户是否具备扩展答案的权限；

[0100] 若所述输入用户具备扩展答案的权限，根据所述第一答案的答案类型，确定所述第一答案的多个关联维度，并基于所述多个关联维度，通过所述政务知识图谱对所述第一答案进行扩展，获得第二答案。

[0101] 在该可选的实施方式中，可以预先设置用户的扩展权限，限制有些用户搜索时只能获得单一的答案，有些用户搜索时不仅能获得答案还能获得相关的答案，通过扩展权限的设置，可以防止非法用户(即无权限用户)获取更深次的信息，避免信息的泄露。

[0102] S18、输出所述第一答案和所述第二答案。

[0103] 具体的，所述输出所述第一答案和所述第二答案包括：

[0104] 获取与所述信息类型匹配的输出模式；

[0105] 若所述输出模式为文字模式，获取所述政务搜索信息的文字属性，采用所述文字属性输出所述第一答案和所述第二答案；或

[0106] 若所述输出模式为图片模式，获取与所述答案类型匹配的图形模板，采用所述图形模板对所述第一答案和所述第二答案进行可视化展示；或

[0107] 若所述输出模式为语音模式，将所述第一答案和所述第二答案转换为与所述政务搜索信息的口音匹配的第二语音，并输出所述第二语音。

[0108] 在该实施方式中，如果信息类型是文字，则输出模式也是文字，如果信息类型是图片，则输出模式也是图片，如果信息类型是语音，则输出模式也是语音，通过这种输出方式，可以使得输出的信息更加符合用户的习惯，更加个性化，从而提高用户体验。

[0109] 在图1所描述的方法流程中，根据政务搜索信息的信息类型对政务搜索信息进行预处理，通过NLP进行语义识别，获得用户的真实意图，并基于政务知识图谱搜索用户想到的答案，同时，对答案进行关联搜索，可以给用户提供更多的关联信息，不仅可以提高搜索的效率和搜索的准确度，同时，还可以提高用户搜索的满意度，提高用户体验。

[0110] 以上所述，仅是本发明的具体实施方式，但本发明的保护范围并不局限于此，对于本领域的普通技术人员来说，在不脱离本发明创造构思的前提下，还可以做出改进，但这些均属于本发明的保护范围。

[0111] 请参见图2，图2是本发明公开的一种数据搜索装置的较佳实施例的功能模块图。

[0112] 在一些实施例中，所述数据搜索装置运行于电子设备中。所述数据搜索装置可以包括多个由程序代码段所组成的功能模块。所述数据搜索装置中的各个程序段的程序代码可以存储于存储器中，并由至少一个处理器所执行，以执行图1所描述的基于自然语言处理的数据搜索方法中的部分或全部步骤。

[0113] 本实施例中，所述数据搜索装置根据其所执行的功能，可以被划分为多个功能模块。所述功能模块可以包括：接收模块201、获取模块202、处理模块203、确定模块204、查找模块205、判断模块206及输出模块207。本发明所称的模块是指一种能够被至少一个处理器所执行并且能够完成固定功能的一系列计算机程序段，其存储在存储器中。

[0114] 接收模块201，用于接收输入的政务搜索信息。

[0115] 获取模块202，用于获取所述政务搜索信息的信息类型。

[0116] 其中，所述信息类型可以为文字类型，也可以为图片类型，还可以是语音类型，本发明实施例不做限定。

[0117] 处理模块203，用于根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息。

[0118] 具体的，所述根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息包括：

[0119] 若所述信息类型为文本类型，判断所述政务搜索信息中是否存在错误文字；

[0120] 若所述政务搜索信息中存在错误文字，根据编辑距离算法从预设词库中查找与所述错误文字相似度高的第一文字；

[0121] 使用所述第一文字替换所述错误文字，并将替换后的政务搜索信息确定为预处理信息。

[0122] 在该可选的实施方式中，当用户输入的政务搜索信息出现单词拼写错误、同音字时，搜索系统可以自动识别纠错。汉字的纠错使用编辑距离算法实现，首先在预设词库(通常是同一行业领域的词库)中查找词相似度以及拼音相似度较高的词作为候选词，以缩小编辑距离计算范围，之后计算各个候选词与需要纠错词的编辑距离，取编辑距离最小的候选词，如果编辑距离值超过设置的纠错阀值，则作为结果返回。通过这种方式可以对政务搜索信息纠正，获得更加准确的搜索信息，有利于提高对用户意图识别的准确率。

[0123] 具体的，所述根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息包括：

[0124] 若所述信息类型为图片类型，采用图像去模糊算法，对所述政务搜索信息进行图片处理，获得第一图片；

[0125] 若所述第一图片存在边缘无关信息，删除所述边缘无关信息，并将删除后的第一图片确定为预处理信息。

[0126] 在该可选的实施方式中，当信息类型为图片类型时，通过对政务搜索信息进行图片处理以及删除边缘无关信息，不仅可以提高图片的清晰度，同时，也减少了图片的冗余信息，减少需要识别的信息量，从而可以提高对用户意图识别的准确率。

[0127] 具体的，所述根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息包括：

[0128] 若所述信息类型为语音类型，识别所述政务搜索信息所属的区域口音类型；

[0129] 基于所述区域口音类型，对所述政务搜索信息进行语音纠正处理，并将处理后的政务搜索信息确定为预处理信息。

[0130] 在该可选的实施方式中，不同区域的用户在输入语音时，会自带本区域的口音，由于口音的区域化，会导致识别困难，通过基于所述区域口音类型，对所述政务搜索信息进行语音纠正处理，可以获得符合标准的普通话，进而可以准确识别出用户的意图，从而可以提高对用户意图识别的准确率。

[0131] 确定模块204，用于通过自然语言处理NLP对所述预处理信息进行识别，确定所述预处理信息对应的问题类型。

[0132] 其中，NLP (Natural Language Processing，自然语言处理)可以对预处理信息进行识别，获得该预处理信息对应的问题类型。

[0133] 其中，可以将所述预处理信息归类为多组问题类型。

[0134] 问题类型1：“人物属性- 职称- 层级- 学校- 标签- 机构- 处室- 地理”，举例来说，XXX 机构有哪些部长，XXX 机构XXX 部门的部长是谁。

[0135] 问题类型2：“机构- 地理- 时间- 预算值”，举例来说，例如XXX 机构2018年的预算是多少，预算最多的机构是多少。

[0136] 查找模块205，用于基于所述问题类型，通过预先建立的政务知识图谱对所述预处理信息进行查找，获得第一答案。

[0137] 判断模块206，用于根据所述政务知识图谱的更新状态，判断所述第一答案是否有效。

[0138] 其中，如果所述政务知识图谱的更新状态表明所述政务知识图谱内的数据均是最新的数据，则可以确定所述第一答案是有效的，反正，如果所述政务知识图谱的更新状态表明所述政务知识图谱内的数据不是最新的数据，即所述政务知识图谱很久没有更新了，则可以确定所述第一答案是无效的。

[0139] 通过这种方式，既可以确保查询到的第一答案与所述预处理信息匹配，同时，还能确保第一答案对用户当前的查询来说是属于有效的答案。

[0140] 所述确定模块204，还用于若所述第一答案有效，根据所述第一答案的答案类型，确定所述第一答案的多个关联维度，并基于所述多个关联维度，通过所述政务知识图谱对所述第一答案进行扩展，获得第二答案。

[0141] 其中，关联维度可以包括但不限于时间、地点、区域、人物等。

[0142] 举例来说，所述第一答案的答案类型为人物，则可以确定所述第一答案的多个关联维度可以为时间、区域，基于关联维度可以查找到这个人物的出生年月，这个人物出生在

哪里,目前在哪里居住等等。

[0143] 又举例来说,所述第一答案的答案类型为某个区域,则可以确定所述第一答案的关联维度可以为时间,基于关联维度可以查找到这个区域近几年的预算情况。

[0144] 输出模块,用于输出所述第一答案和所述第二答案。

[0145] 具体的,所述输出所述第一答案和所述第二答案包括:

[0146] 获取与所述信息类型匹配的输出模式;

[0147] 若所述输出模式为文字模式,获取所述政务搜索信息的文字属性,采用所述文字属性输出所述第一答案和所述第二答案;或

[0148] 若所述输出模式为图片模式,获取与所述答案类型匹配的图形模板,采用所述图形模板对所述第一答案和所述第二答案进行可视化展示;或

[0149] 若所述输出模式为语音模式,将所述第一答案和所述第二答案转换为与所述政务搜索信息的口音匹配的第二语音,并输出所述第二语音。

[0150] 在该实施方式中,如果信息类型是文字,则输出模式也是文字,如果信息类型是图片,则输出模式也是图片,如果信息类型是语音,则输出模式也是语音,通过这种输出方式,可以使得输出的信息更加符合用户的习惯,更加个性化,从而提高用户体验。

[0151] 可选的,所述获取模块202,还用于在所述确定模块204通过自然语言处理NLP对所述预处理信息进行识别,确定所述预处理信息对应的问题类型之后,获取所述问题类型的机密级别,以及获取当前输入用户的用户级别;

[0152] 所述判断模块206,还用于判断所述用户级别与所述机密级别是否匹配;

[0153] 若所述用户级别与所述机密级别匹配,基于所述问题类型,通过预先建立的政务知识图谱对所述预处理信息进行查找,获得第一答案。

[0154] 在该可选的实施方式中,由于用户是在搜索政务相关的信息,而这类信息通常是比较机密的,不是所有人均可以进行搜索的。通过将机密级别与用户级别进行匹配判断,可以对输入用户进行身份校验,如果匹配,表明输入用户属于合法用户,具备搜索的权限。通过这种用户身份的校验,可以确保信息的安全性。

[0155] 可选的,所述获取模块202,还用于通过网络爬虫技术,从各个政务网站获取政务信息;

[0156] 所述确定模块204,还用于从所述政务信息中确定多个实体,并基于所述多个实体的标签,分析所述多个实体的关联关系;

[0157] 所述数据搜索装置还包括:

[0158] 建立模块,用于根据所述多个实体以及所述关联关系,建立政务知识图谱。

[0159] 在该可选的实施方式中,可以通过大量的规则和爬虫技术对全国各省市政府官网,财政厅官网,政府招投标等相关网站进行定时信息爬取。

[0160] 举例来说,假设A 市市长和B 市市长都曾经就读于北京大学,那么北京大学就是一个实体(entity),其中一个关系(relation)是就读人员,这个关系关联到另一些实体就是A 市市长和B市市长。但由于政府领域的特殊性,领导班子可能换届,但岗位始终存在,所以我们记录岗位以及对应的人。每个实体有特定的标签,比如机构、地理、职能、属性。以此类推,可以构建出大量实体和关系庞大的政务知识图谱。

[0161] 可选的,若所述第一答案有效,所述获取模块202,还用于获取当前输入用户的用

户信息；

[0162] 所述判断模块206，还用于根据所述用户信息，判断所述输入用户是否具备扩展答案的权限；

[0163] 所述确定模块204，还用于若所述输入用户具备扩展答案的权限，根据所述第一答案的答案类型，确定所述第一答案的多个关联维度，并基于所述多个关联维度，通过所述政务知识图谱对所述第一答案进行扩展，获得第二答案。

[0164] 在该可选的实施方式中，可以预先设置用户的扩展权限，限制有些用户搜索时只能获得单一的答案，有些用户搜索时不仅能获得答案还能获得相关的答案，通过扩展权限的设置，可以防止非法用户(即无权限用户)获取更深次的信息，避免信息的泄露。

[0165] 在图2所描述数据搜索装置中，根据政务搜索信息的信息类型对政务搜索信息进行预处理，通过NLP进行语义识别，获得用户的真实意图，并基于政务知识图谱搜索用户想到的答案，同时，对答案进行关联搜索，可以给用户提供更多的关联信息，不仅可以提高搜索的效率和搜索的准确度，同时，还可以提高用户搜索的满意度，提高用户体验。

[0166] 如图3所示，图3是本发明实现基于自然语言处理的数据搜索方法的较佳实施例的电子设备的结构示意图。所述电子设备3包括存储器31、至少一个处理器32、存储在所述存储器31中并可在所述至少一个处理器32上运行的计算机程序33及至少一条通讯总线34。

[0167] 本领域技术人员可以理解，图3所示的示意图仅仅是所述电子设备3的示例，并不构成对所述电子设备3的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件，例如所述电子设备3还可以包括输入输出设备、网络接入设备等。

[0168] 所述至少一个处理器32可以是中央处理单元(Central Processing Unit，CPU)，还可以是其他通用处理器、数字信号处理器(Digital Signal Processor，DSP)、专用集成电路(Application Specific Integrated Circuit，ASIC)、现场可编程门阵列(Field-Programmable Gate Array，FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。该处理器32可以是微处理器或者该处理器32也可以是任何常规的处理器等，所述处理器32是所述电子设备3的控制中心，利用各种接口和线路连接整个电子设备3的各个部分。

[0169] 所述存储器31可用于存储所述计算机程序33和/或模块/单元，所述处理器32通过运行或执行存储在所述存储器31内的计算机程序和/或模块/单元，以及调用存储在存储器31内的数据，实现所述电子设备3的各种功能。所述存储器31可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序(比如声音播放功能、图像播放功能等)等；存储数据区可存储根据电子设备3的使用所创建的数据(比如音频数据)等。此外，存储器31可以包括非易失性存储器，例如硬盘、内存、插接式硬盘，智能存储卡(Smart Media Card，SMC)，安全数字(Secure Digital，SD)卡，闪存卡(Flash Card)、至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。

[0170] 结合图1，所述电子设备3中的所述存储器31存储多个指令以实现一种基于自然语言处理的数据搜索方法，所述处理器32可执行所述多个指令从而实现：

[0171] 接收输入的政务搜索信息；

[0172] 获取所述政务搜索信息的信息类型；

[0173] 根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息；

[0174] 通过自然语言处理NLP对所述预处理信息进行识别，确定所述预处理信息对应的问题类型；

[0175] 基于所述问题类型，通过预先建立的政务知识图谱对所述预处理信息进行查找，获得第一答案；

[0176] 根据所述政务知识图谱的更新状态，判断所述第一答案是否有效；

[0177] 若所述第一答案有效，根据所述第一答案的答案类型，确定所述第一答案的多个关联维度，并基于所述多个关联维度，通过所述政务知识图谱对所述第一答案进行扩展，获得第二答案；

[0178] 输出所述第一答案和所述第二答案。

[0179] 具体地，所述处理器32对上述指令的具体实现方法可参考图1对应实施例中相关步骤的描述，在此不赘述。

[0180] 在图3所描述的电子设备3中，根据政务搜索信息的信息类型对政务搜索信息进行预处理，通过NLP进行语义识别，获得用户的真实意图，并基于政务知识图谱搜索用户想到的答案，同时，对答案进行关联搜索，可以给用户提供更多的关联信息，不仅可以提高搜索的效率和搜索的准确度，同时，还可以提高用户搜索的满意度，提高用户体验。

[0181] 所述电子设备3集成的模块/单元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，本发明实现上述实施例方法中的全部或部分流程，也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计算机可读存储介质中，该计算机程序在被处理器执行时，可实现上述各个方法实施例的步骤。其中，所述计算机程序包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器以及只读存储器(ROM，Read-Only Memory)。

[0182] 在本发明所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块的划分，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。

[0183] 所述作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者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者全部模块来实现本实施例方案的目的。

[0184] 另外，在本发明各个实施例中的各功能模块可以集成在一个处理单元中，也可以是各个单元单独物理存在，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模块的形式实现。

[0185] 对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附关联图标记视为限制所涉及的权利要求。系统权利要求中陈述的多个单元或装置也可以通过软件或者硬件来实现。

[0186] 最后应说明的是,以上实施例仅用以说明本发明的技术方案而非限制,尽管参照较佳实施例对本发明进行了详细说明,本领域的普通技术人员应当理解,可以对本发明的技术方案进行修改或等同替换,而不脱离本发明技术方案的精神和范围。

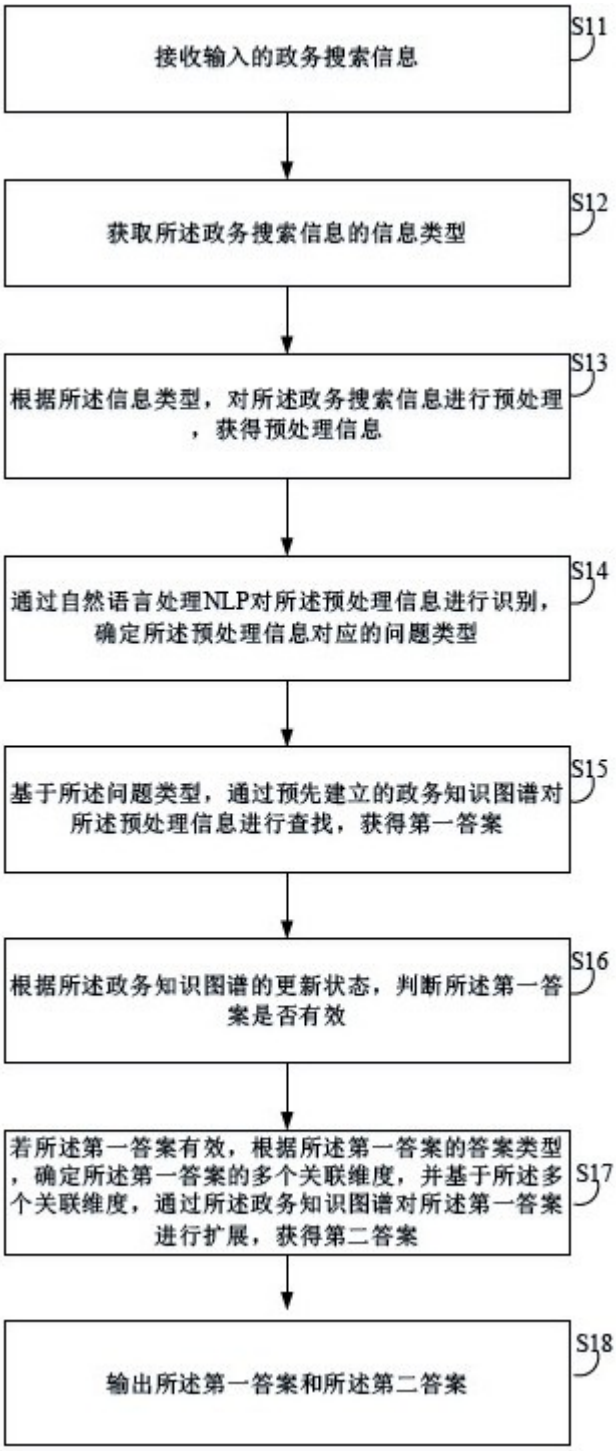

接收输入的政务搜索信息
S11
获取所述政务搜索信息的信息类型
S12
根据所述信息类型，对所述政务搜索信息进行预处理，获得预处理信息
S13
通过自然语言处理NLP对所述预处理信息进行识别，确定所述预处理信息对应的问题类型
S14
基于所述问题类型，通过预先建立的政务知识图谱对所述预处理信息进行查找，获得第一答案
S15
根据所述政务知识图谱的更新状态，判断所述第一答案是否有效
S16
若所述第一答案有效，根据所述第一答案的答案类型，确定所述第一答案的多个关联维度，并基于所述多个关联维度，通过所述政务知识图谱对所述第一答案进行扩展，获得第二答案
S17
输出所述第一答案和所述第二答案
S18

图1

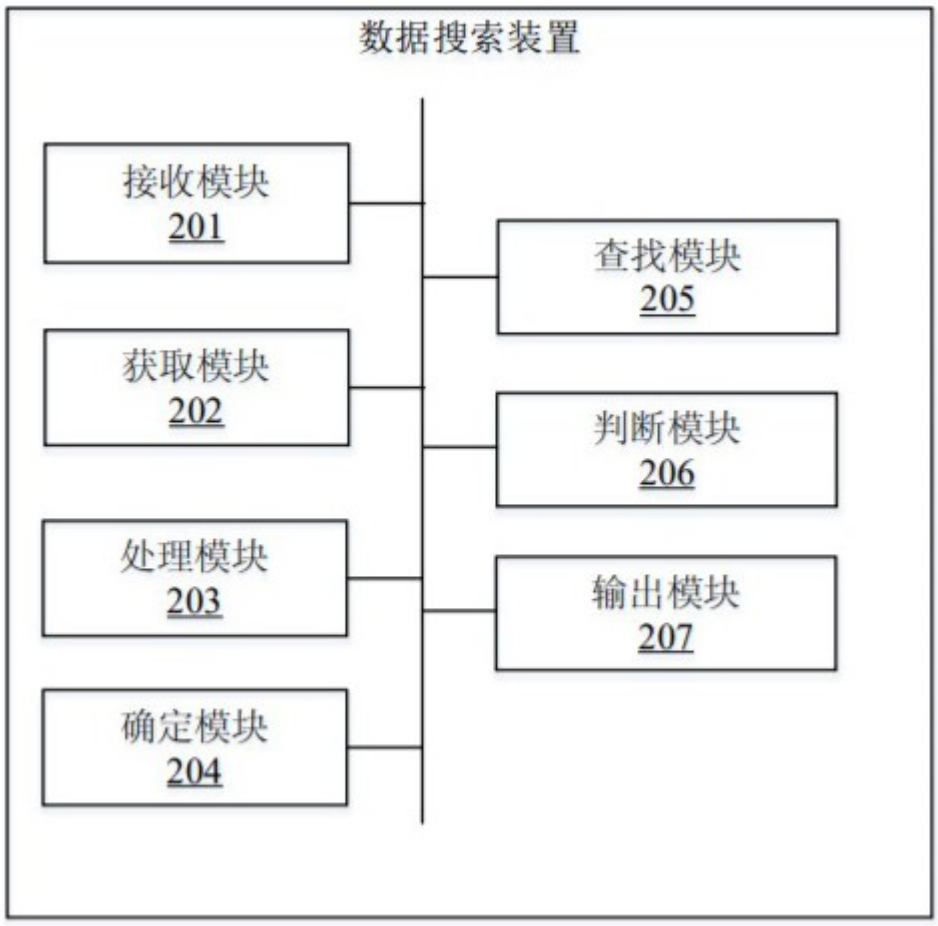

数据搜索装置
接收模块
201
获取模块
202
处理模块
203
确定模块
204
查找模块
205
判断模块
206
输出模块
207

图2

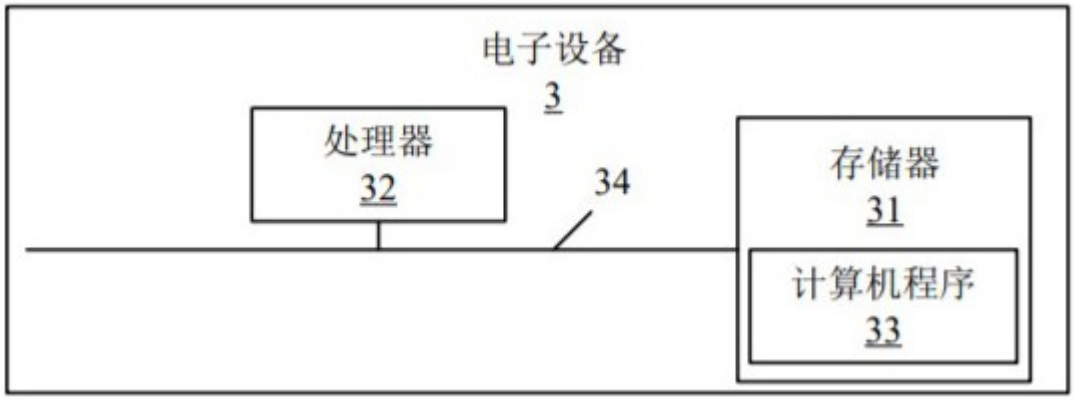

电子设备
3
处理器
32
34
存储器
31
计算机程序
33

图3