(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116108854 A
(43)申请公布日 2023.05.12

(21)申请号 202211717290.2

(22)申请日 2022.12.29

(71)申请人 江苏省未来网络创新研究院
地址 211111 江苏省南京市江宁区秣陵街道秣周东路12号悠谷2号楼

(72)发明人 冯立二 张发雨 王宁 党章 孟奥 杨正云 杜宇 袁扬

(74)专利代理机构 南京理工信达知识产权代理有限公司 32542
专利代理师 彭甲临

(51)Int.Cl.
G06F 40/30(2020.01)
G06F 40/242(2020.01)
G06F 16/31(2019.01)

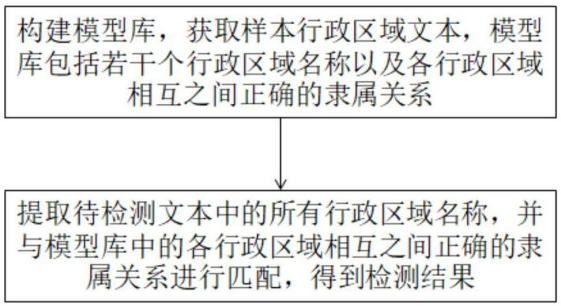
权利要求书1页 说明书4页 附图2页

(54)发明名称

一种检测行政区域名称表述错位的方法、设备及介质

(57)摘要

本发明涉及一种检测行政区域名称表述错位的方法、设备及介质，该方法包括构建模型库，获取样本行政区域文本，模型库包括若干个行政区域名称以及各行政区域相互之间正确的隶属关系；提取待检测文本中的所有行政区域名称，并与模型库中的各行政区域相互之间正确的隶属关系进行匹配，得到检测结果。本发明与现有技术相比，其显著优点是：通过采用双数组Trie树算法来检测文本，不需要人工校对，同时还提高了检测效率；支持多种样本语料存储、构建模型库，可应用于各大类型网站内容、新闻、媒体、国家机关等机构的文案中省(自治区)、市、区(县)表述错位的检测；同时，结合训练不同的样本语料能够实现优化模型库的目的，具有广泛的应用前景。



CN 116108854 A

1.一种检测行政区域名称表述错位的方法,其特征在于:该方法包括:

构建模型库,获取样本行政区域文本,模型库包括若干个行政区域名称以及各行政区域相互之间正确的隶属关系;

提取待检测文本中的所有行政区域名称,并与模型库中的各行政区域相互之间正确的隶属关系进行匹配,得到检测结果。

2.根据权利要求1所述的一种检测行政区域名称表述错位的方法,其特征在于:所述模型库的构建步骤为:

预置包括若干行政区域名称的行政区域字典,并按照各行政区域名称对应的行政区域等级定义前缀匹配规则;

从含有若干行政区域相互之间隶属关系表述的若干种样本语料中,获取所有行政区域名称以及行政区域相互之间的隶属关系,并计算每种行政区域相互之间的隶属关系出现的统计频率,当统计频率达到合格阈值时,确定当前行政区域名称相互之间的隶属关系为正确关系;

构建模型库,用于存储正确关系以及其对应的行政区域名称。

3.根据权利要求2所述的一种检测行政区域名称表述错位的方法,其特征在于:所述行政区域字典采用双数组Trie树,其中:将所述前缀匹配规则存入所述双数组Trie树的叶子节点中,所述前缀匹配规则表示各行政区域相互之间的隶属关系的正则表达式。

4.根据权利要求3所述的一种检测行政区域名称表述错位的方法,其特征在于:

利用所述双数组Trie树对每一种样本语料进行正向最大匹配,得到若干个行政区域名称,取出其对应的前缀匹配规则,同时获取其上下文;

将得到的若干个行政区域名称及其对应的前缀匹配规则,应用到获取的上下文中,得到若干个行政区域相互之间的隶属关系并累计得到统计频率。

5.根据权利要求4所述的一种检测行政区域名称表述错位的方法,其特征在于:所述检测结果的具体匹配流程为:

利用所述双数组Trie树,获得所述待检测文本中的行政区域名称,并取出其对应的前缀匹配规则,同时获取其上下文;

将获得的行政区域名称及其对应的前缀匹配规则,应用到获取的上下文中,得到各行政区域相互之间的归属关系;

将所述各行政区域相互之间的隶属关系输入到所述模型库中进行匹配,若匹配失败,则检测结果为所述待检测文本的当前行政区域名称相互之间的隶属关系表述错位;反之,则检测结果为表述正常。

6.一种检测行政区域名称表述错位的设备,其特征在于,包括:

存储器,用于存储计算机程序;

处理器,用于执行所述计算机程序时实现如权利要求1至5任一项所述一种检测行政区域名称表述错位的方法的步骤。

7.一种计算机可读存储介质,其特征在于:所述计算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如权利要求1至5任一项所述一种检测行政区域名称表述错位的方法的步骤。

一种检测行政区域名称表述错位的方法、设备及介质

技术领域

[0001] 本发明涉及语义检测技术领域，特别是一种检测行政区域名称表述错位的方法、设备及介质。

背景技术

[0002] 行政区域名称表述错位是一种常见的文字错误形式，例如以下表述形式：“安徽省南京市”、“南京市雨花区”、“南京句容市”，这些都是将行政区域的隶属关系表述错位了；实际正确的表述应该是“江苏省南京市”，“南京市雨花台区”，“镇江市句容市”。

[0003] 通常情况下解决方式包括：1、人工校对；2、使用通用的错别字校对软件校对；3、使用Macbert、Kenlm这样的自然语言预测模型进行预测分析。其中，方式1容易出现人工校对遗漏的问题，特别是在校对内容体量较大时，并且，人工校对耗费工作量较大成本较高。方式2是采用比较通用错别字的检测方案，但是检出率和正确率都不高；方式3的算法模型方法进一步改善了方式1和2的问题，但是仅适合数值计算体量较小的应用场景，适用范围较窄。

[0004] 文献1：中国授权发明专利CN114168705B公开了一种基于地址要素索引的中文地址匹配方法，利用余弦相似度计算方法对地址匹配结果集合进行筛选和排序，从而获得最优的匹配结果和对应的地址空间位置坐标；但是余弦相似度计算是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小的度量当向量空间的个体数量增多且计算量大，当有新地址名称加入时，就必须重新计算词的权值；不仅操作繁琐，而且影响匹配操作的稳定性，出现大量匹配报错的问题。

发明内容

[0005] 本发明的目的在于提供一种检测行政区域名称表述错位的方法、设备及介质，利用双数组Trie树，获得待检测文本中的行政区域名称，通过构建模型库，定义前缀匹配规则，将从待检测文本中得到的各行政区域相互之间的隶属关系输入到模型库中完成匹配。

[0006] 实现本发明目的的技术解决方案为：

[0007] 一种检测行政区域名称表述错位的方法，该方法包括：

[0008] 构建模型库，获取样本行政区域文本，模型库包括若干个行政区域名称以及各行政区域相互之间正确的隶属关系；

[0009] 提取待检测文本中的所有行政区域名称，并与模型库中的各行政区域相互之间正确的隶属关系进行匹配，得到检测结果。

[0010] 进一步的，模型库的构建步骤为：

[0011] 预置包括若干行政区域名称的行政区域字典，并按照各行政区域名称对应的行政区域等级定义前缀匹配规则；

[0012] 从含有若干行政区域相互之间隶属关系表述的若干种样本语料中，获取所有行政区域名称以及行政区域相互之间的隶属关系，并计算每种行政区域相互之间的隶属关系出

现的统计频率，当统计频率达到合格阈值时，确定当前行政区域名称相互之间的隶属关系为正确关系；

[0013] 构建模型库，用于存储正确关系以及其对应的行政区域名称。

[0014] 进一步的，行政区域字典采用双数组Trie树，其中：将前缀匹配规则存入双数组Trie树的叶子节点中，前缀匹配规则表示各行政区域相互之间的隶属关系的正则表达式。

[0015] 进一步的，利用双数组Trie树对每一种样本语料进行正向最大匹配，得到若干个行政区域名称，取出其对应的前缀匹配规则，同时获取其上下文；将得到的若干个行政区域名称及其对应的前缀匹配规则，应用到获取的上下文中，得到若干个行政区域相互之间的隶属关系并累计得到统计频率。

[0016] 进一步的，检测结果的具体匹配流程为：

[0017] 利用双数组Trie树，获得待检测文本中的行政区域名称，并取出其对应的前缀匹配规则，同时获取其上下文；

[0018] 将获得的行政区域名称及其对应的前缀匹配规则，应用到获取的上下文中，得到各行政区域相互之间的归属关系；

[0019] 将各行政区域相互之间的隶属关系输入到模型库中进行匹配，若匹配失败，则检测结果为待检测文本的当前行政区域名称相互之间的隶属关系表述错位；反之，则检测结果为表述正常。

[0020] 本发明还提供了一种检测行政区域名称表述错位的设备，该设备包括用于存储计算机程序的存储器，以及用于执行计算机程序时实现上述一种检测行政区域名称表述错位的方法的步骤的处理器。

[0021] 本发明还提供了一种计算机可读存储介质，该计算可读存储介质上存储有计算机程序，计算机程序被处理器执行时实现上述一种检测行政区域名称表述错位的方法的步骤。

[0022] 本发明与现有技术相比，其显著优点是：

[0023] 1、通过采用双数组Trie树算法来检测文本，不需要人工校对，降低了人工成本的同时还提高检测效率。

[0024] 2、支持多种样本语料存储、构建模型库，可应用于各大类型网站内容、新闻、媒体、国家机关等机构的文案中省(自治区)、市、区(县)表述错位的检测，同时，结合训练不同的样本语料能够实现优化模型库的目的，具有广泛的应用前景。

附图说明

[0025] 图1是本发明的一种检测行政区域名称表述错位的方法的流程示意图。

[0026] 图2是本发明的一种检测行政区域名称表述错位的方法的模型库构建流程示意图。

[0027] 图3是本发明的一种检测行政区域名称表述错位的方法的检测结果匹配流程示意图。

具体实施方式

[0028] 以下结合附图，详细说明本发明的实施方式。

[0029] 如图1所示,一种检测行政区域名称表述错位的方法,该方法包括:

[0030] 构建模型库,获取样本行政区域文本,模型库包括若干个行政区域名称以及各行政区域相互之间正确的隶属关系;

[0031] 提取待检测文本中的所有行政区域名称,并与模型库中的各行政区域相互之间正确的隶属关系进行匹配,得到检测结果。

[0032] 如图2所示,模型库的构建步骤为:

[0033] S10:构建省(自治区)、市、区(县)隶属关系的模型库:

[0034] S11:人工预置省(自治区)、市、区(县)的名称作为行政区域字典。

[0035] S12:人工标识行政区域字典中每个词的行政区域等级,便于通过程序自动生成对应的前缀匹配规则。例如:将“江苏省”标识为省级行政区域,将“南京市”标识为市级行政区域,将“雨花台区”标识位(县)级行政区域。

[0036] S13:根据步骤S11和步骤S12预置的行政区域名称和行政区域等级,生成对应的前缀匹配规则;例如:对于行政区域名称“南京市”,生成的正则表达式为“[\u4e00-\u9fa5]{2,5}[省|自治区]?南京”,该正则表达式用于匹配类似于“江苏省南京市”、“江苏南京”、“江苏的南京市”这种省市隶属关系;对于词条“雨花台区”,生成的正则表达式为“[\u4e00-\u9fa5]{2,5}[市]?雨花台区”,该正则表达式用于匹配类似于“南京市雨花台区”、“南京雨花台区”、“南京的雨花台区”这种市区隶属关系。

[0037] S14:准备大量含有行政区域隶属关系表述的样本语料。

[0038] S15:将行政区域字典构建为双数组Trie树,将前缀匹配规则也一并存入叶子节点。然后对输入的样本语料利用构建好的双数组Trie树进行匹配查找,即使用双数组Trie树算法对句子进行正向最大匹配,在样本语料中找出具体的省(自治区)、市、区(县)名称,取出前缀匹配规则,同时获取其上下文。

[0039] 其中,Trie树也称为字典树、前缀树,是一种常被用于词检索的树结构,其检索思想为:利用词的共同前缀以达到节省空间的目的,双数组树(Double-array Trie)结合了array查询效率高、list节省空间的优点,具体是通过两个数组来实现,可以用于字符串的快速检索。

[0040] S16:将步骤S15检出的每一个行政区域名称以及其对应的前缀匹配规则,应用到获取的上下文,匹配出行政区域隶属关系,并累加该行政区域隶属关系出现的次数,计算得到统计频率。

[0041] S17:将每一个行政区域隶属关系的最终统计频率与合格阈值进行比较,留下大于或等于合格阈值的行政区域隶属关系保存到数据库,我们称之为模型库。

[0042] S18:重复步骤S14、S15、S16、S17,不断地加入新的语料进行训练,优化模型库。

[0043] 如图3所示,检测结果的具体匹配流程为:

[0044] S20:检测实际文本:

[0045] S21:利用步骤S15构建的双数组Trie树对待检测文本进行匹配查找,找出具体的省(自治区)、市、区(县)名称,取出前缀匹配规则,同时获取其上下文。

[0046] S22:将步骤S21查找出的每个行政区域名称以及其对应的前缀匹配规则,应用到获取的上下文,匹配出该行政区域隶属关系。

[0047] S23:将步骤S22输出的行政区域隶属关系,去模型库中做匹配,匹配成功的为正常

表述，未匹配成功的为表述错位。

[0048] 本发明还提供了一种检测行政区域名称表述错位的设备，该设备包括用于存储计算机程序的存储器，以及用于执行计算机程序时实现上述一种检测行政区域名称表述错位的方法的步骤的处理器。

[0049] 本发明还提供了一种计算机可读存储介质，该计算可读存储介质上存储有计算机程序，计算机程序被处理器执行时实现上述一种检测行政区域名称表述错位的方法的步骤。

[0050] 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，计算机程序可存储于一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0051] 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

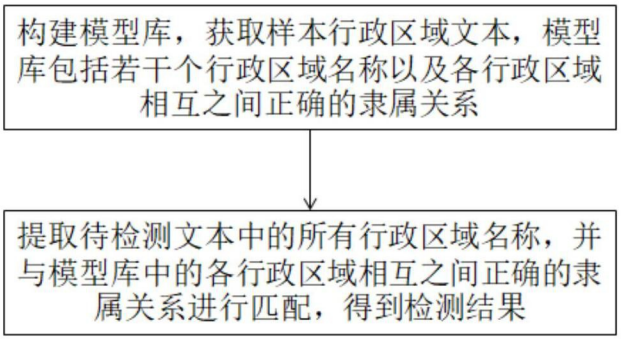

构建模型库，获取样本行政区域文本，模型库包括若干个行政区域名称以及各行政区域相互之间正确的隶属关系
提取待检测文本中的所有行政区域名称，并与模型库中的各行政区域相互之间正确的隶属关系进行匹配，得到检测结果

图1

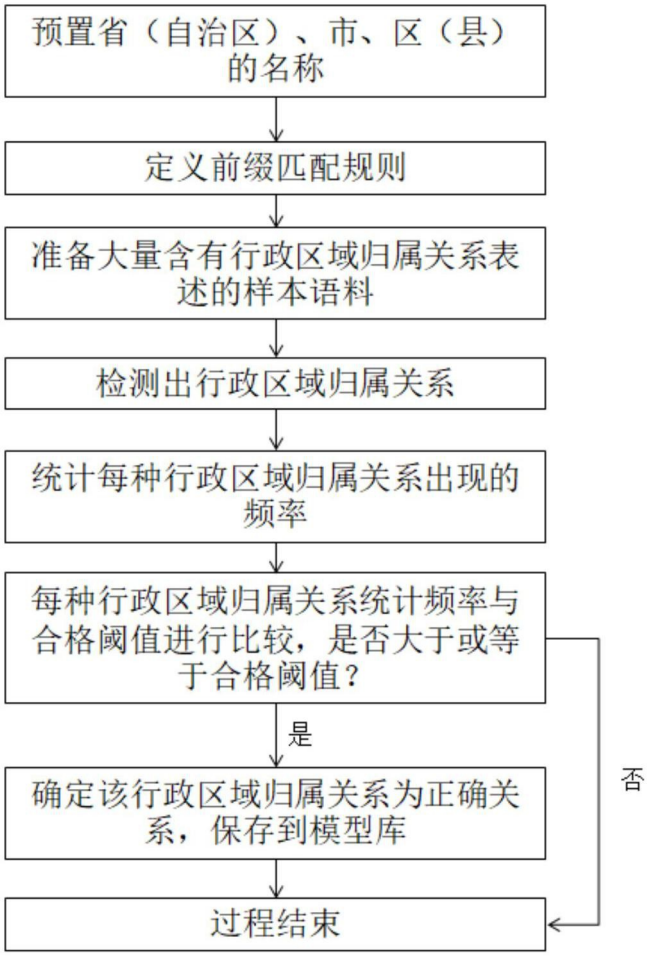

预置省（自治区）、市、区（县）的名称
定义前缀匹配规则
准备大量含有行政区域归属关系表述的样本语料
检测出行政区域归属关系
统计每种行政区域归属关系出现的频率
每种行政区域归属关系统计频率与合格阈值进行比较，是否大于或等于合格阈值？
是
否
确定该行政区域归属关系为正确关系，保存到模型库
过程结束

图2

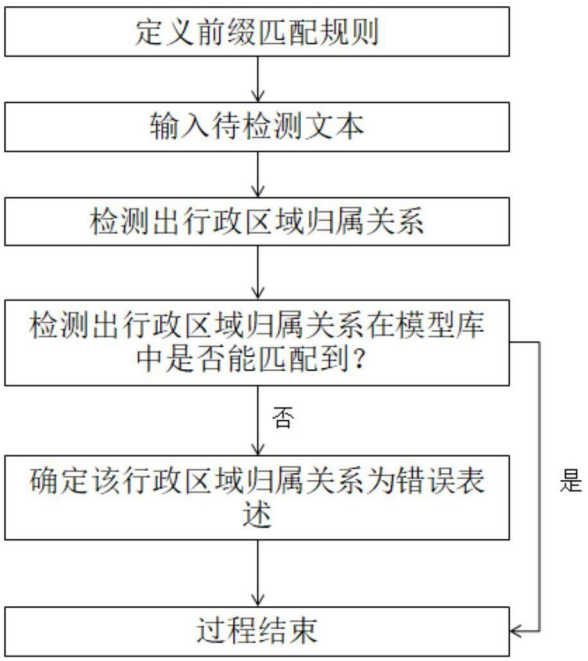

定义前缀匹配规则
输入待检测文本
检测出行政区域归属关系
检测出行政区域归属关系在模型库中是否能匹配到？
否
是
确定该行政区域归属关系为错误表述
过程结束

图3