



(12) 发明专利

(10) 授权公告号 CN 108628811 B

(45) 授权公告日 2022. 04. 12

(21) 申请号 201810315328.0

G06F 40/30 (2020.01)

(22) 申请日 2018.04.10

G06F 40/205 (2020.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 108628811 A

(56) 对比文件

CN 107239442 A, 2017.10.10

CN 107239442 A, 2017.10.10

CN 106096024 A, 2016.11.09

CN 104102667 A, 2014.10.15

CN 105677700 A, 2016.06.15

CN 105404686 A, 2016.03.16

CN 107239442 A, 2017.10.10

(43) 申请公布日 2018.10.09

(73) 专利权人 北京京东尚科信息技术有限公司

地址 100195 北京市海淀区杏石口路65号

西杉创意园四区11号楼东段1-4层西
段1-4层

专利权人 北京京东世纪贸易有限公司

李晓林 等. 非规范化中文地址的行政区划
提取算法.《计算机应用》.2017, 第37卷(第3期),
第876-882页.

(72) 发明人 吴帅 李晨 陈晓佳

孙存群 等. 基于分级地名库的中文地理编
码.《计算机应用》.2010, 第30卷(第7期), 第
1953-1958页.

(74) 专利代理机构 中原信达知识产权代理有限
责任公司 11219

代理人 张一军 张效荣

审查员 郑诗嘉

(51) Int. Cl.

G06F 40/194 (2020.01)

G06F 40/284 (2020.01)

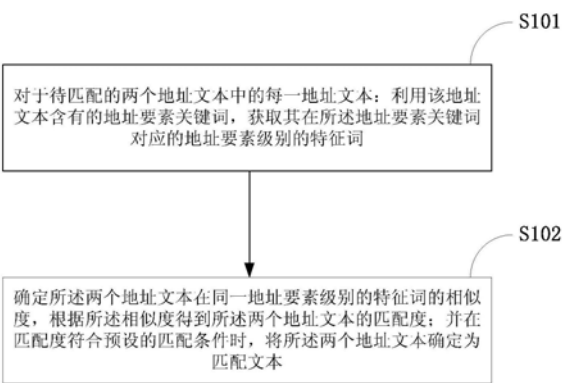
权利要求书4页 说明书15页 附图2页

(54) 发明名称

地址文本的匹配方法和装置

(57) 摘要

本发明公开了一种地址文本的匹配方法和装置, 涉及计算机技术领域。该方法的一具体实施方式包括: 对于待匹配的两个地址文本中的每一地址文本: 利用该地址文本含有的地址要素关键词, 获取其在所述地址要素关键词对应的地址要素级别的特征词; 确定所述两个地址文本在同一地址要素级别的特征词的相似度, 根据所述相似度得到所述两个地址文本的匹配度; 并在匹配度符合预设的匹配条件时, 将所述两个地址文本确定为匹配文本。该实施方式能够实现地址文本是否匹配的高精度判别。



1. 一种地址文本的匹配方法,其特征在于,包括:

对于待匹配的两个地址文本中的每一地址文本:利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词;

确定所述两个地址文本在同一地址要素级别的特征词的相似度,根据所述相似度得到所述两个地址文本的匹配度;并在匹配度符合预设的匹配条件时,将所述两个地址文本确定为匹配文本;

所述利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词具体包括:

以预设顺序依次处理每一地址文本对应的初选文本中的每一地址要素关键词:

对于该初选文本中的第一个地址要素关键词,将该初选文本在所述第一个地址要素关键词之前的部分确定为所述第一个地址要素关键词对应的地址要素级别的特征词;

对于该初选文本中在第一个地址要素关键词之后的任一地址要素关键词,将该初选文本在该任一地址要素关键词与其前一地址要素关键词之间的部分确定为该任一地址要素关键词对应的地址要素级别的特征词。

2. 根据权利要求1所述的方法,其特征在于,所述方法进一步包括:

在所述利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词之前:

利用预先建立的地址分词词典对每一地址文本进行分词,从分词结果中确定每一地址文本在至少一个行政区级别的特征词;其中,任一行政区级别高于任一地址要素级别;

当所述两个地址文本在同一行政区级别的特征词存在不符时,将所述两个地址文本确定为不匹配文本。

3. 根据权利要求2所述的方法,其特征在于,所述至少一个行政区级别为多个行政区级别,所述地址分词词典中包括所述多个行政区级别的行政区名称;以及,所述从分词结果中确定每一地址文本在至少一个行政区级别的特征词具体包括:

以预设顺序依次处理分词结果中的每一词:

对于分词结果中的第一词,按照行政区级别从高到低的顺序在所述地址分词词典中确定与其相同的一个行政区名称,并将所述第一词作为该地址文本在该行政区名称对应的行政区级别的特征词;

对于分词结果中位于所述第一词之后的任一词:从其前一词在所述地址分词词典中下辖的行政区名称中,按照行政区级别从高到低的顺序确定与所述任一词相同的一个行政区名称,并将所述任一词作为该地址文本在该行政区名称对应的行政区级别的特征词。

4. 根据权利要求3所述的方法,其特征在于,所述地址分词词典中的行政区名称包括行政区的非规范名称。

5. 根据权利要求3所述的方法,其特征在于,所述方法进一步包括:

在所述从分词结果中确定每一地址文本在至少一个行政区级别的特征词之后:当该地址文本缺少一个行政区级别的特征词、且含有该行政区级别的下一行政区级别的特征词时,将所述地址分词词典中下辖该特征词、且处于该地址文本缺少的行政区级别的行政区名称确定为该地址文本缺少的特征词。

6. 根据权利要求3所述的方法,其特征在于,所述方法进一步包括:

当所述两个地址文本在同一行政区级别的特征词不存在不符时,将每一地址文本在每一行政区级别的特征词去除,得到对应于每一地址文本的初选文本。

7. 根据权利要求6所述的方法,其特征在于,所述方法进一步包括:预先为至少一个地址要素级别设置前置依赖条件;以及,在所述将该初选文本在该任一地址要素关键词与其前一地址要素关键词之间的部分确定为该任一地址要素关键词对应的地址要素级别的特征词之后,所述方法进一步包括:

若该地址要素级别设置有前置依赖条件,则在该前一地址要素关键词不符合该前置依赖条件时,将该特征词舍弃。

8. 根据权利要求6所述的方法,其特征在于,所述方法进一步包括:

在以预设顺序依次处理每一初选文本中的每一地址要素关键词之后:若该初选文本中存在同一地址要素级别的多个特征词,则查询为该地址要素级别预设的限制策略是否支持多个特征词:若不支持,将所述多个特征词舍弃。

9. 根据权利要求6所述的方法,其特征在于,所述方法进一步包括:

在以预设顺序依次处理每一初选文本中的每一地址要素关键词之后,在该初选文本中不存在预设的匹配粒度级别的特征词时,将所述两个地址文本确定为不匹配文本。

10. 根据权利要求7所述的方法,其特征在于,

所述多个行政区级别为省级、地级、县级;

所述地址要素级别为以下一种或多种:镇级、村级、组级、路级、道路号级、栋级、单元级、层级、户级以及标志地名级;其中:

镇级的地址要素关键词为:镇、乡、街道;

村级的地址要素关键词为:村、居委会、社区;

组级的地址要素关键词为:组;

路级的地址要素关键词为:路、街、道、巷、弄、弄堂、胡同、大道、大路、大街;

道路号级的地址要素关键词为:号;

栋级的地址要素关键词为:栋、幢、座、楼、号楼;

单元级的地址要素关键词为:单元;

层级的地址要素关键词为:层、楼;

户级的地址要素关键词为:户、室、屋、房、号、门;

标志地名级的地址要素关键词为:广场、酒店、饭店、宾馆、中心、大厦、大楼、商城、公园、小区、花园、公司、寓所、会所、市场、社区、新村、苑、开发区、科技区、园区。

11. 根据权利要求10所述的方法,其特征在于,

组级的前置依赖条件为:其在初选文本中的前一地址要素关键词对应于村级;

层级的前置依赖条件为:其在初选文本的前一地址要素关键词对应于单元级或栋级。

12. 根据权利要求10所述的方法,其特征在于,所述方法进一步包括:

在以预设顺序依次处理每一初选文本中的每一地址要素关键词之后:若该初选文本中存在路级的两个特征词,则将二者按照预设排列顺序拼接为一个词,作为该初选文本在路级的特征词。

13. 根据权利要求10所述的方法,其特征在于,所述方法进一步包括:

在以预设顺序依次处理每一初选文本中的每一地址要素关键词时:

若路级的地址要素关键词的后一地址要素关键词为号,则确定该地址要素关键词对应于道路号级;

若路级的地址要素关键词的后一地址要素关键词不为号、且该路级的地址要素关键词在后邻接至少一位数字,则确定该数字为道路号级的特征词;

若栋级、单元级或层级的地址要素关键词的后一地址要素关键词为号,则确定该地址要素关键词对应于户级;

若栋级、单元级或层级的地址要素关键词在后邻接至少一位数字、且在后不存在地址要素关键词,则确定该数字为户级的特征词。

14. 根据权利要求10所述的方法,其特征在于,所述方法进一步包括:

在以预设顺序依次处理每一初选文本中的每一地址要素关键词时:

若道路号级或标志地名级的地址要素关键词的后一地址要素关键词为楼,则确定该地址要素关键词对应于栋级;

若栋级或单元级的地址要素关键词的后一地址要素关键词为楼,则确定该地址要素关键词对应于层级。

15. 根据权利要求1所述的方法,其特征在于,所述确定所述两个地址文本在同一地址要素级别的特征词的相似度,具体包括:

当所述两个地址文本在该地址要素级别均存在特征词时,所述相似度为两个特征词的编辑距离与两个特征词中字符串长度最大值之商;

当所述两个地址文本之一在该地址要素级别存在特征词、另一在该地址要素级别不存在特征词时,所述相似度为预设阈值;以及,

当所述两个地址文本在该地址要素级别均不存在特征词时,所述相似度为零。

16. 根据权利要求1所述的方法,其特征在于,所述方法进一步包括:为每一地址要素级别预先配置权重值;以及,所述根据所述相似度得到所述两个地址文本的匹配度包括:

将所述两个地址文本在每一地址要素级别的特征词的相似度的加权平均值作为所述两个地址文本的匹配度。

17. 根据权利要求15所述的方法,其特征在于,预设的匹配条件为:匹配度大于所述预设阈值。

18. 一种地址文本的匹配装置,其特征在于,包括:

特征词获取模块,用于针对待匹配的两个地址文本中的每一地址文本,利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词;

匹配模块,用于确定所述两个地址文本在同一地址要素级别的特征词的相似度,根据所述相似度得到所述两个地址文本的匹配度;并在匹配度符合预设的匹配条件时,将所述两个地址文本确定为匹配文本;

特征词获取模块进一步用于:以预设顺序依次处理每一地址文本对应的初选文本中的每一地址要素关键词:对于该初选文本中的第一个地址要素关键词,将该初选文本在所述第一个地址要素关键词之前的部分确定为所述第一个地址要素关键词对应的地址要素级别的特征词;对于该初选文本中在第一个地址要素关键词之后的任一地址要素关键词,将该初选文本在该任一地址要素关键词与其前一地址要素关键词之间的部分确定为该任一地址要素关键词对应的地址要素级别的特征词。

19. 一种电子设备,其特征在于,包括:

一个或多个处理器;

存储装置,用于存储一个或多个程序,

当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-17中任一所述的方法。

20. 一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述程序被处理器执行时实现如权利要求1-17中任一所述的方法。

地址文本的匹配方法和装置

技术领域

[0001] 本发明涉及计算机技术领域,尤其涉及一种地址文本的匹配方法和装置。

背景技术

[0002] 随着物流行业的迅速发展,服务方会积累到海量的用户地址数据,这些地址数据具有极大的挖掘价值和应用价值,而对任意地址数据进行匹配则是其中的关键。由于地址数据表述方式较多、用户书写不规范,所以任意地址数据的匹配具有较大难度。现有技术中,一般通过人工判定、模糊索引匹配、编辑距离匹配、经纬度匹配等方法实现地址数据的匹配。

[0003] 在实现本发明的过程中,发明人发现现有技术至少存在以下问题:

[0004] 1.在面對海量地址数据时,人工判定方法成本高,效率低,无法满足时效要求。

[0005] 2.模糊索引匹配一般只适用于特定环境且其精度较低。

[0006] 3.编辑距离匹配方法精度较低、对地址数据长度敏感,且易被地址数据中的无关信息、重复信息影响。

[0007] 4.经纬度匹配方法由于定位精度有限无法达到精确到户的水平,此外,经纬度为平面信息,该方法无法区分对应同一地面的不同楼层。

发明内容

[0008] 有鉴于此,本发明实施例提供一种地址文本的匹配方法和装置,能够通过提取两个地址文本在每一地址要素级别的特征词,并计算特征词的相似度从而判别二者是否匹配,该方法具有较高的判别精度与处理效率。

[0009] 为实现上述目的,根据本发明的一个方面,提供了一种地址文本的匹配方法。

[0010] 本发明实施例的地址文本的匹配方法包括:对于待匹配的两个地址文本中的每一地址文本:利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词;确定所述两个地址文本在同一地址要素级别的特征词的相似度,根据所述相似度得到所述两个地址文本的匹配度;并在匹配度符合预设的匹配条件时,将所述两个地址文本确定为匹配文本。

[0011] 可选地,所述方法进一步包括:在所述利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词之前:利用预先建立的地址分词词典对每一地址文本进行分词,从分词结果中确定每一地址文本在至少一个行政区级别的特征词;其中,任一行政区级别高于任一地址要素级别;当所述两个地址文本在同一行政区级别的特征词存在不符时,将所述两个地址文本确定为不匹配文本。

[0012] 可选地,所述至少一个行政区级别为多个行政区级别,所述地址分词词典中包括所述多个行政区级别的行政区名称;以及,所述从分词结果中确定每一地址文本在至少一个行政区级别的特征词具体包括:以预设顺序依次处理分词结果中的每一词:对于分词结果中的第一词,按照行政区级别从高到低的顺序在所述地址分词词典中确定与其相同的一

个行政区名称,并将所述第一词作为该地址文本在该行政区名称对应的行政区级别的特征词;对于分词结果中位于所述第一词之后的任一词:从其前一词在所述地址分词词典中下辖的行政区名称中,按照行政区级别从高到低的顺序确定与所述任一词相同的一个行政区名称,并将所述任一词作为该地址文本在该行政区名称对应的行政区级别的特征词。

[0013] 可选地,所述地址分词词典中的行政区名称包括行政区的非规范名称。

[0014] 可选地,所述方法进一步包括:在所述从分词结果中确定每一地址文本在至少一个行政区级别的特征词之后:当该地址文本缺少一个行政区级别的特征词、且含有该行政区级别的下一行政区级别的特征词时,将所述地址分词词典中下辖该特征词、且处于该地址文本缺少的行政区级别的行政区名称确定为该地址文本缺少的特征词。

[0015] 可选地,所述方法进一步包括:当所述两个地址文本在同一行政区级别的特征词不存在不符时,将每一地址文本在每一行政区级别的特征词去除,得到对应于每一地址文本的初选文本。

[0016] 可选地,所述利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词具体包括:以预设顺序依次处理每一初选文本中的每一地址要素关键词:对于该初选文本中的第一个地址要素关键词,将该初选文本在所述第一个地址要素关键词之前的部分确定为所述第一个地址要素关键词对应的地址要素级别的特征词;对于该初选文本中在第一个地址要素关键词之后的任一地址要素关键词,将该初选文本在该任一地址要素关键词与其前一地址要素关键词之间的部分确定为该任一地址要素关键词对应的地址要素级别的特征词。

[0017] 可选地,所述方法进一步包括:预先为至少一个地址要素级别设置前置依赖条件;以及,在所述将该初选文本在该任一地址要素关键词与其前一地址要素关键词之间的部分确定为该任一地址要素关键词对应的地址要素级别的特征词之后,所述方法进一步包括:若该地址要素级别设置有前置依赖条件,则在该前一地址要素关键词不符合该前置依赖条件时,将该特征词舍弃。

[0018] 可选地,所述方法进一步包括:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词之后:若该初选文本中存在同一地址要素级别的多个特征词,则查询为该地址要素级别预设的限制策略是否支持多个特征词:若不支持,将所述多个特征词舍弃。

[0019] 可选地,所述方法进一步包括:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词之后,在该初选文本中不存在预设的匹配粒度级别的特征词时,将所述两个地址文本确定为不匹配文本。

[0020] 可选地,所述多个行政区级别为省级、地级、县级;所述地址要素级别为以下一种或多种:镇级、村级、组级、路级、道路号级、栋级、单元级、层级、户级以及标志地名级;其中:镇级的地址要素关键词为:镇、乡、街道;村级的地址要素关键词为:村、居委会、社区;组级的地址要素关键词为:组;路级的地址要素关键词为:路、街、道、巷、弄、弄堂、胡同、大道、大路、大街;道路号级的地址要素关键词为:号;栋级的地址要素关键词为:栋、幢、座、楼、号楼;单元级的地址要素关键词为:单元;层级的地址要素关键词为:层、楼;户级的地址要素关键词为:户、室、屋、房、号、门;标志地名级的地址要素关键词为:广场、酒店、饭店、宾馆、中心、大厦、大楼、商城、公园、小区、花园、公司、寓所、会所、市场、社区、新村、苑、开发区、科技区、园区。

[0021] 可选地,组级的前置依赖条件为:其在初选文本中的前一地址要素关键词对应于村级;层级的前置依赖条件为:其在初选文本的前一地址要素关键词对应于单元级或栋级。

[0022] 可选地,所述方法进一步包括:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词之后:若该初选文本中存在路级的两个特征词,则将二者按照预设排列顺序拼接为一个词,作为该初选文本在路级的特征词。

[0023] 可选地,所述方法进一步包括:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词时:若路级的地址要素关键词的后一地址要素关键词为号,则确定该地址要素关键词对应于道路号级;若路级的地址要素关键词的后一地址要素关键词不为号、且该路级的地址要素关键词在后邻接至少一位数字,则确定该数字为道路号级的特征词;若栋级、单元级或层级的地址要素关键词的后一地址要素关键词为号,则确定该地址要素关键词对应于户级;若栋级、单元级或层级的地址要素关键词在后邻接至少一位数字、且在后不存在地址要素关键词,则确定该数字为户级的特征词。

[0024] 可选地,所述方法进一步包括:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词时:若道路号级或标志地名级的地址要素关键词的后一地址要素关键词为楼,则确定该地址要素关键词对应于栋级;若栋级或单元级的地址要素关键词的后一地址要素关键词为楼,则确定该地址要素关键词对应于层级。

[0025] 可选地,所述确定所述两个地址文本在同一地址要素级别的特征词的相似度,具体包括:当所述两个地址文本在该地址要素级别均存在特征词时,所述相似度为两个特征词的编辑距离与两个特征词中字符串长度最大值之商;当所述两个地址文本之一在该地址要素级别存在特征词、另一在该地址要素级别不存在特征词时,所述相似度为预设阈值;以及,当所述两个地址文本在该地址要素级别均不存在特征词时,所述相似度为零。

[0026] 可选地,所述方法进一步包括:为每一地址要素级别预先配置权重值;以及,所述根据所述相似度得到所述两个地址文本的匹配度包括:将所述两个地址文本在每一地址要素级别的特征词的相似度的加权平均值作为所述两个地址文本的匹配度。

[0027] 可选地,预设的匹配条件为:匹配度大于所述预设阈值。

[0028] 为实现上述目的,根据本发明的另一方面,提供了一种地址文本的匹配装置。

[0029] 本发明实施例的地址文本的匹配装置可包括:特征词获取模块,用于针对待匹配的两个地址文本中的每一地址文本,利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词;匹配模块,用于确定所述两个地址文本在同一地址要素级别的特征词的相似度,根据所述相似度得到所述两个地址文本的匹配度;并在匹配度符合预设的匹配条件时,将所述两个地址文本确定为匹配文本。

[0030] 可选地,所述装置可进一步包括:行政区处理模块,用于在所述利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词之前:利用预先建立的地址分词词典对每一地址文本进行分词,从分词结果中确定每一地址文本在至少一个行政区级别的特征词;其中,任一行政区级别高于任一地址要素级别;当所述两个地址文本在同一行政区级别的特征词存在不符时,将所述两个地址文本确定为不匹配文本。

[0031] 可选地,所述至少一个行政区级别为多个行政区级别,所述地址分词词典中包括所述多个行政区级别的行政区名称;以及,行政区处理模块可进一步用于:以预设顺序依次

处理分词结果中的每一词；对于分词结果中的第一词，按照行政区级别从高到低的顺序在所述地址分词词典中确定与其相同的一个行政区名称，并将所述第一词作为该地址文本在该行政区名称对应的行政区级别的特征词；对于分词结果中位于所述第一词之后的任一词：从其前一词在所述地址分词词典中下辖的行政区名称中，按照行政区级别从高到低的顺序确定与所述任一词相同的一个行政区名称，并将所述任一词作为该地址文本在该行政区名称对应的行政区级别的特征词。

[0032] 可选地，所述地址分词词典中的行政区名称包括行政区的非规范名称。

[0033] 可选地，行政区处理模块可进一步用于：在所述从分词结果中确定每一地址文本在至少一个行政区级别的特征词之后：当该地址文本缺少一个行政区级别的特征词、且含有该行政区级别的下一行政区级别的特征词时，将所述地址分词词典中下辖该特征词、且处于该地址文本缺少的行政区级别的行政区名称确定为该地址文本缺少的特征词。

[0034] 可选地，行政区处理模块可进一步用于：当所述两个地址文本在同一行政区级别的特征词不存在不符时，将每一地址文本在每一行政区级别的特征词去除，得到对应于每一地址文本的初选文本。

[0035] 可选地，特征词获取模块可进一步用于：以预设顺序依次处理每一初选文本中的每一地址要素关键词：对于该初选文本中的第一个地址要素关键词，将该初选文本在所述第一个地址要素关键词之前的部分确定为所述第一个地址要素关键词对应的地址要素级别的特征词；对于该初选文本中在第一个地址要素关键词之后的任一地址要素关键词，将该初选文本在该任一地址要素关键词与其前一地址要素关键词之间的部分确定为该任一地址要素关键词对应的地址要素级别的特征词。

[0036] 可选地，所述装置进一步包括依赖条件设置模块，用于预先为至少一个地址要素级别设置前置依赖条件；以及，特征词获取模块可进一步用于：在将该初选文本在该任一地址要素关键词与其前一地址要素关键词之间的部分确定为该任一地址要素关键词对应的地址要素级别的特征词之后，若该地址要素级别设置有前置依赖条件，则在该前一地址要素关键词不符合该前置依赖条件时，将该特征词舍弃。

[0037] 可选地，特征词获取模块可进一步用于：在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词之后：若该初选文本中存在同一地址要素级别的多个特征词，则查询为该地址要素级别预设的限制策略是否支持多个特征词：若不支持，将所述多个特征词舍弃。

[0038] 可选地，特征词获取模块可进一步用于：在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词之后，在该初选文本中不存在预设的匹配粒度级别的特征词时，将所述两个地址文本确定为不匹配文本。

[0039] 可选地，所述多个行政区级别为省级、地级、县级；所述地址要素级别为以下一种或多种：镇级、村级、组级、路级、道路号级、栋级、单元级、层级、户级以及标志地名级；其中：镇级的地址要素关键词为：镇、乡、街道；村级的地址要素关键词为：村、居委会、社区；组级的地址要素关键词为：组；路级的地址要素关键词为：路、街、道、巷、弄、弄堂、胡同、大道、大路、大街；道路号级的地址要素关键词为：号；栋级的地址要素关键词为：栋、幢、座、楼、号楼；单元级的地址要素关键词为：单元；层级的地址要素关键词为：层、楼；户级的地址要素关键词为：户、室、屋、房、号、门；标志地名级的地址要素关键词为：广场、酒店、饭店、宾馆、

中心、大厦、大楼、商城、公园、小区、花园、公司、寓所、会所、市场、社区、新村、苑、开发区、科技区、园区。

[0040] 可选地,组级的前置依赖条件为:其在初选文本中的前一地址要素关键词对应于村级;层级的前置依赖条件为:其在初选文本的前一地址要素关键词对应于单元级或栋级。

[0041] 可选地,特征词获取模块可进一步用于:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词之后:若该初选文本中存在路级的两个特征词,则将二者按照预设排列顺序拼接为一个词,作为该初选文本在路级的特征词。

[0042] 可选地,特征词获取模块可进一步用于:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词时:若路级的地址要素关键词的后一地址要素关键词为号,则确定该地址要素关键词对应于道路号级;若路级的地址要素关键词的后一地址要素关键词不为号、且该路级的地址要素关键词在后邻接至少一位数字,则确定该数字为道路号级的特征词;若栋级、单元级或层级的地址要素关键词的后一地址要素关键词为号,则确定该地址要素关键词对应于户级;若栋级、单元级或层级的地址要素关键词在后邻接至少一位数字、且在后不存在地址要素关键词,则确定该数字为户级的特征词。

[0043] 可选地,特征词获取模块可进一步用于:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词时:若道路号级或标志地名级的地址要素关键词的后一地址要素关键词为楼,则确定该地址要素关键词对应于栋级;若栋级或单元级的地址要素关键词的后一地址要素关键词为楼,则确定该地址要素关键词对应于层级。

[0044] 可选地,匹配模块可进一步用于:当所述两个地址文本在该地址要素级别均存在特征词时,将所述相似度确定为两个特征词的编辑距离与两个特征词中字符串长度最大值之商;当所述两个地址文本之一在该地址要素级别存在特征词、另一在该地址要素级别不存在特征词时,将所述相似度确定为预设阈值;以及,当所述两个地址文本在该地址要素级别均不存在特征词时,将所述相似度确定为零。

[0045] 可选地,所述装置可进一步包括权重配置模块,用于为每一地址要素级别预先配置权重值;以及,匹配模块可进一步用于:将所述两个地址文本在每一地址要素级别的特征词的相似度的加权平均值作为所述两个地址文本的匹配度。

[0046] 可选地,预设的匹配条件可为:匹配度大于所述预设阈值。

[0047] 为实现上述目的,根据本发明的又一方面,提供了一种电子设备。

[0048] 本发明的一种电子设备包括:一个或多个处理器;存储装置,用于存储一个或多个程序,当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现本发明所提供的地址文本的匹配方法。

[0049] 为实现上述目的,根据本发明的再一方面,提供了一种计算机可读存储介质。

[0050] 本发明的一种计算机可读存储介质,其上存储有计算机程序,所述程序被处理器执行时实现本发明所提供的地址文本的匹配方法。

[0051] 根据本发明的技术方案,上述发明中的一个实施例具有如下优点或有益效果:

[0052] 其一,通过提取两个地址文本在各地址要素级别的特征词、并计算特征词之间的相似度、进而利用相似度计算二者匹配度来实现地址文本的高精度匹配。此外,上述方法可由计算设备高速执行,适于海量数据处理。

[0053] 其二,在利用地址要素级别的特征词处理地址文本之前,可首先分析地址文本中

的行政区特征,将行政区特征不符的地址文本过滤。这样,可避免部分地址文本在地址要素阶段的冗余处理,进一步提升匹配效率。另外,在提取行政区特征时,采用包括行政区非规范名称的地址分词词典,可进一步提高匹配精度。

[0054] 其三,为了提升匹配方法的实用性与应用范围,设置三个行政区级别、十个地址要素级别,并考虑实际生产生活中每一级别的应用特点制定多种应用策略:前置依赖策略、(特征词)限制策略等。

[0055] 其四,本发明的匹配粒度可精确到户,同时受地址文本中噪声数据的影响较小,当地址文本中存在缩写、异名、冗余等情况时,可最大程度减弱其影响。

[0056] 上述的非惯用的可选方式所具有的进一步效果将在下文中结合具体实施方式加以说明。

附图说明

[0057] 附图用于更好地理解本发明,不构成对本发明的不当限定。其中:

[0058] 图1是根据本发明实施例中地址文本的匹配方法的主要步骤示意图;

[0059] 图2是根据本发明实施例中地址文本的匹配装置的组成部分示意图;

[0060] 图3是根据本发明实施例可以应用于其中的示例性系统架构图;

[0061] 图4是用来实现本发明实施例的地址文本的匹配方法的电子设备的结构示意图。

具体实施方式

[0062] 以下结合附图对本发明的示范性实施例做出说明,其中包括本发明实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本发明的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0063] 在本发明实施例的技术方案中,可首先分析地址文本中的行政区特征,将行政区特征不符的地址文本过滤;之后通过提取两个地址文本在各地址要素级别的特征词、并计算特征词之间的相似度、进而利用相似度计算二者匹配度来实现地址文本的高精度匹配,克服现有技术中数据处理效率较低、无法匹配到户、易受噪声数据影响等缺陷。

[0064] 需要指出的是,在不冲突的情况下,本发明的实施例以及实施例中的技术特征可以相互结合。

[0065] 图1是根据本发明实施例的地址文本的匹配方法的主要步骤示意图。

[0066] 如图1所示,本发明实施例的地址文本的匹配方法可按照以下步骤执行:

[0067] 步骤S101:对于待匹配的两个地址文本中的每一地址文本:利用该地址文本含有的地址要素关键词,获取其在地址要素关键词对应的地址要素级别的特征词。

[0068] 在本发明实施例中,匹配指的是判别两个地址文本是否指示同一地点,并对指示同一地点的地址文本进行标记的操作。地址文本可以是以任何一种自然语言书写的、包含地址信息的文本数据。以中文为例,绝大多数地址文本均围绕“镇”、“街”、“单元”等表征具体地域范围的字或词语书写,在本发明中这些字或词语通称为地址要素关键词。可以理解,每一个地址要素关键词均与一个地址要素级别对应。例如:地址要素关键词“镇”、“乡”对应的地址要素级别为镇级,“街”、“道”、“巷”对应的地址要素级别为路级。

[0069] 在本步骤中,特征词用于表示某一地址文本在某一地址要素级别的具体特征。例如,在地址文本“新阳街道西园路”中,“街道”为镇级的地址要素关键词,“新阳”即为“街道”对应的镇级的特征词;“路”为路级的地址要素关键词,“西园”即为“路”对应的路级的特征词。

[0070] 实际应用场景中,某些待匹配的地址文本在省、地、县级别存在差异,如此即可预先通过对省、地、县等较高行政区级别的判别实现地址文本的快速甄别。因此,在本发明实施例中,可在步骤S101之前执行以下步骤以实现基于较高行政区级别的地址文本数据处理:

[0071] 1.利用预先建立的地址分词词典对每一地址文本进行分词。

[0072] 其中,地址分词词典中可包括省级、地级、市级、镇级等各行政区名称。为了提高地址匹配精度,地址分词词典中还包括各行政区的非规范名称。例如:“新疆维吾尔自治区”(规范名称)的非规范名称“新疆”、“新疆省”、“新疆自治区”、“维吾尔自治区”。基于上述地址分词词典,本步骤可利用正向最大匹配算法进行分词。可以理解的是,本步骤也可利用其它适用算法如逆向最大匹配算法等进行分词,本发明对此不作任何限制。关于本步骤中的分词操作,可如下例所示:对地址文本“河北廊坊香河南京路10号西湖小区”分词可得到分词结果“河北、廊坊、香河、南京、西湖”

[0073] 2.从分词结果中确定每一地址文本在至少一个行政区级别的特征词。

[0074] 为了利用地址文本的高级别地址要素进行过滤,在本步骤中,任一行政区级别均高于步骤S101中的地址要素级别。例如,本步骤中的行政区级别可根据实际需求设置为省级、地级,也可以设置为省级、地级、县级,还可以是省级、地级、县级、镇级。地址文本在某行政区级别的特征词可表征该地址文本在该行政区级别的具体特征。以行政区级别是省级、地级、县级为例,在地址文本“吉林省长春市二道区”中,“吉林”为省级特征词,“长春”为地级特征词,“二道”为县级特征词。具体应用中,各级行政区级别以及各行政区名称的获取可以遵循行政区划相关法律、法规、标准及文件。

[0075] 具体而言,本步骤可以通过以下方式确定每一行政区级别的特征词:以预设顺序依次处理分词结果中的每一词,对于分词结果中的第一词,按照行政区级别从高到低的顺序在地址分词词典中确定与其相同的一个行政区名称,并将第一词作为该地址文本在该行政区名称对应的行政区级别的特征词。对于分词结果中位于第一词之后的任一词:从其前一词在地址分词词典中下辖的行政区名称中,按照行政区级别从高到低的顺序确定与上述任一词相同的一个行政区名称,并将上述任一词作为该地址文本在该行政区名称对应的行政区级别的特征词。

[0076] 也就是说,对于分词结果中的第一个词,首先从地址分词词典的省级行政区名称中寻找相同词,在未找到的情况下从地级行政区名称中寻找,仍未找到的情况下从县级行政区名称中寻找……直到最终找到或地址分词词典中不存在该词。对于分词结果中第一个词之后的词语,可在其前一词下辖的行政区名称中按照行政区级别从高到低的顺序寻找。其中,预设顺序一般为地址文本的书写或者阅读顺序,在多数情况下预设顺序是中文环境中最常见的从左到右顺序。可以理解,上述确定特征词的方法针对存在多个行政区级别的情形,只有一个行政区级别的情形可用类似方式处理。

[0077] 例如:在从分词结果“河北、廊坊、香河、南京、西湖”中确定上述特征词时,首先判

别第一个词“河北”，以行政区级别从高到低的顺序进行比较，发现地址分词词典中存在与该词相同的省级行政区名称“河北”，于是将第一个词“河北”确定为地址文本在省级的特征词。之后判别第二个词“廊坊”，从其前一词“河北”下辖的行政区中以行政区级别从高到低的顺序进行比较，发现地址分词词典中“河北”下辖的地级行政区名称“廊坊”与该词相同，将该词确定为地址文本在地级的特征词。之后以类似方式将“香河”确定为地址文本在县级的特征词。

[0078] 作为一个优选方案，在确定地址文本在每一行政区级别的特征词之后，当地址文本缺少某一个行政区级别的特征词、同时含有该行政区级别的下一行政区级别的特征词时，将地址分词词典中下辖该特征词、且处于地址文本缺少的行政区级别的行政区名称确定为地址文本缺少的特征词。也就是说，在地址文本含有下级特征词、缺少上级特征词时，将下辖下级特征词的对应行政区名称作为地址文本缺少的特征词。

[0079] 例如，在确定地址文本“河北香河南京路10号西湖小区”的分词结果“河北、香河、南京、西湖”的省级特征词为“河北”、县级行政区为“香河”之后，该地址文本缺少地级关键词，于是将地址分词词典中下辖“香河”的地级行政区名称“廊坊”补充为该地址文本的地级关键词。

[0080] 3. 当两个地址文本在同一行政区级别的特征词存在不符时，将二者确定为不匹配文本。

[0081] 在本步骤中，两个地址文本在同一行政区级别的特征词是否不符是根据预设规则确定的。在一个实施例中，两个地址文本在同一行政区级别的特征词存在不符包括以下两种情形：

[0082] (1) 当两个地址文本均存在省级、地级、县级特征词时，任一级别的特征词不对应即为不符。例如：“河北省、廊坊市、香河县”与“河北省、廊坊市、三河县”不符。

[0083] (2) 当两个地址文本中的任一地址文本中缺少地级特征词时，二者特征词不符。例如：“河北省、廊坊市、香河县”与“河北省”不符。

[0084] 也就是说，根据上述预设规则，两个地址文本中的任一地址文本中缺少县级特征词时，二者特征词是相符的。例如“河北省、廊坊市、香河县”与“河北省、廊坊市”相符，这种处理规则可避免将潜在匹配的地址文本不当去除，有助于提高系统精度。

[0085] 这样，预先从行政区角度处理地址文本之后，可在未确定为不匹配文本的两个地址文本中去除每一行政区级别的特征词，得到对应于每一地址文本的初选文本用于后续匹配。例如：地址文本“河北香河南京路10号西湖小区”去除行政区级别的特征词之后的初选文本为“南京路10号西湖小区”。

[0086] 在步骤S101中，可以对初选文本进行进一步处理，从而获得待匹配的两个地址文本在各地址要素级别的特征词。具体而言，可按照预设顺序依次处理每一初选文本中的每一地址要素关键词：对于该初选文本中的第一个地址要素关键词，将该初选文本在第一个地址要素关键词之前的部分确定为第一个地址要素关键词对应的地址要素级别的特征词；对于该初选文本中在第一个地址要素关键词之后的任一地址要素关键词，将该初选文本在该任一地址要素关键词与其前一地址要素关键词之间的部分确定为该任一地址要素关键词对应的地址要素级别的特征词。可以理解的是，通过上述方式提取到的特征词即为初选文本对应的地址文本在相应地址要素级别的特征词。

[0087] 其中,地址要素级别可以是以下一种或多种:镇级、村级、组级、路级、道路号级、栋级、单元级、层级、户级以及标志地名级。镇级的地址要素关键词为:镇、乡、街道;村级的地址要素关键词为:村、居委会、社区;组级的地址要素关键词为:组;路级的地址要素关键词为:路、街、道、巷、弄、弄堂、胡同、大道、大路、大街;道路号级的地址要素关键词为:号;栋级的地址要素关键词为:栋、幢、座、楼、号楼;单元级的地址要素关键词为:单元;层级的地址要素关键词为:层、楼;户级的地址要素关键词为:户、室、屋、房、号、门;标志地名级的地址要素关键词为:广场、酒店、饭店、宾馆、中心、大厦、大楼、商城、公园、小区、花园、公司、寓所、会所、市场、社区、新村、苑、开发区、科技区、园区。

[0088] 此外,上述在第一个地址要素关键词之前的部分不包括该地址要素关键词,上述在该任一地址要素关键词与其前一地址要素关键词之间的部分不包括这两个地址要素关键词,通过这种处理能够优化特征词提取准确性。例如,对于两个初选文本“平安大道”与“平安道”(二者事实上指示同一街道,只是书写方式存在差异),如果将地址要素关键词作为特征词的一部分进行提取,则可提取到二者在路级的特征词分别为存在差异的“平安大道”与“平安道”,这会使系统以一定概率判定为不同地址。如果在提取特征词时不考虑地址要素关键词,则可提取到二者在路级的特征词皆为“平安”,如此即可解决上述问题。

[0089] 具体应用中,还可预先为至少一个地址要素级别设置前置依赖条件以过滤自身存在缺陷的地址文本。例如,可设置组级的前置依赖条件为:其在初选文本中的前一地址要素关键词对应于村级;层级的前置依赖条件为:其在初选文本的前一地址要素关键词对应于单元级或栋级。这样,当某地址要素级别的前一地址要素关键词不符合该地址要素级别的前置依赖条件时,将该地址要素级别的特征词舍弃。例如,对于初选文本“界炮镇4组”,由于组级的前一地址要素关键词“镇”不属于村级,不满足组级的前置依赖条件,因此将组级的特征词“4”去除。

[0090] 具体应用场景中,某一地址要素级别下可能允许出现多个特征词,针对这种情况,可为每一地址要素级别预先设置限制策略,用于表征该地址要素级别是否支持多个特征词。在各地址要素级别的特征词提取完成之后,若提取到某地址要素级别的多个特征词,而该地址要素级别的限制策略对此并不支持时,将多个特征词去除。

[0091] 特别地,路级的限制策略支持多个特征词,若某初选文本中存在路级的两个特征词,则将二者按照预设排列顺序拼接为一个词,作为该初选文本在路级的特征词。例如:初选文本“北辰西路与大屯路交叉口”中路级的特征词为“北辰西,大屯”,可按照拼音排列顺序将二者拼接为“北辰西大屯”作为最终的路级特征词。这样,当该初选文本面对另一初选文本“大屯路与北辰西路交叉口”时,上述另一初选文本拼接的路级特征词也是“北辰西大屯”,从而防止系统识别出错,提高识别精度。可以理解的是,上述“交叉口”的情形同样适用于“交叉”、“交汇”、“交汇处”、“交叉处”等,这些词语可在路级特征词提取之后删除。

[0092] 在前述的各地址要素级别的关键词中,道路号级与户级存在相同的地址要素关键词“号”,且道路号级的“号”往往被用户省略。为了准确区分“号”实际对应的地址要素级别以及确定道路号级的特征词,本发明在提取地址要素关键词时执行如下步骤:

[0093] 若路级的地址要素关键词的后一地址要素关键词为号,则确定该地址要素关键词对应于道路号级;若路级的地址要素关键词的后一地址要素关键词不为号、且该路级的地址要素关键词在后邻接至少一位数字,则确定该数字为道路号级的特征词;

[0094] 若栋级、单元级或层级的地址要素关键词的后一地址要素关键词为号,则确定该地址要素关键词对应于户级;若栋级、单元级或层级的地址要素关键词在后邻接至少一位数字、且在后不存在地址要素关键词,则确定该数字为户级的特征词。

[0095] 例如:在初选文本“平安大道100号”中,“号”的前一地址要素关键词“大道”对应于路级,则“号”对应于道路号级。在初选文本“平安大道100花园小区5栋”中,路级的地址要素关键词“大道”的后一地址要素关键词“小区”不是“号”,而且“大道”后邻接数字“100”,则判定“100”为道路号级的特征词。

[0096] 再如:在初选文本“花园小区4栋5单元602号”中,“号”的前一地址要素关键词“单元”对应于单元级,则“号”对应于户级。在初选文本“花园小区4栋5单元602”中,单元级的地址要素关键词“单元”在后邻接数字“602”,且“单元”后没有其它地址要素关键词,则将“602”确定为户级的特征词。

[0097] 在前述的各地址要素级别的关键词中,除了以上情况,栋级与层级也存在相同的关键词“楼”,可在提取特征词时以类似方式处理:若道路号级或标志地名级的地址要素关键词的后一地址要素关键词为楼,则确定该地址要素关键词对应于栋级;若栋级或单元级的地址要素关键词的后一地址要素关键词为楼,则确定该地址要素关键词对应于层级。

[0098] 在本发明实施例中,还可根据需求预先设置匹配粒度级别,如将匹配粒度级别设置为单元级或户级。这样,可根据匹配粒度级别对初选文本进行筛选,将不具有匹配粒度级别特征词的初选文本去除。具体来说,提取到两个初选文本在每一地址要素级别的特征词之后,若某初选文本中不存在匹配粒度级别的特征词,则将二者对应的地址文本确定为不匹配文本。

[0099] 步骤S102:确定两个地址文本在同一地址要素级别的特征词的相似度,根据相似度得到两个地址文本的匹配度;并在匹配度符合预设的匹配条件时,将两个地址文本确定为匹配文本。

[0100] 在本步骤中,可计算步骤S101提取的两个地址文本在各地址要素级别的特征词之间的相似度,进而得到两个地址文本的匹配度。

[0101] 具体地,当两个地址文本在该地址要素级别均存在特征词时,相似度为两个特征词的编辑距离与两个特征词中字符串长度最大值之商,即:

[0102]
$$\text{score} = \text{ed}(x, y) / \max[\text{len}(x), \text{len}(y)]$$

[0103] 其中, score为相似度, x、y分别为两个地址文本在同一地址要素级别的特征词, ed(x, y)为x、y的编辑距离, len(x)、len(y)分别为x、y的字符串长度。

[0104] 可以理解的是,除上述公式之外,相似度也可以直接采用编辑距离、利用余弦相似度或杰卡德相似度(Jaccard similarity coefficient)计算,本发明对此不作限制。

[0105] 当两个地址文本之一在该地址要素级别存在特征词、另一在该地址要素级别不存在特征词时,相似度为预设阈值。其中,上述预设阈值为后续即将介绍的衡量两个地址文本匹配度的阈值。这种设置的原因为:两个地址文本之一在某地址要素级别不存在特征词时,将相似度确定为预设阈值可使此地址要素级别不影响地址文本匹配度的计算,从而提高匹配度计算精度。此外,当两个地址文本在该地址要素级别均不存在特征词时,相似度为零。

[0106] 为了反映不同地址要素级别在地址匹配中的不同贡献,可预先为每一地址要素级别配置权重值。一般地,较高的地址要素级别以及匹配粒度级别具有较高权重,权重值为非

负数。这样,两个地址文本的匹配度为两个地址文本在每一地址要素级别的特征词的相似度的(基于上述权重值的)加权平均值。可以理解的是,实际应用中,也可直接将两个地址文本在每一地址要素级别的特征词的相似度的算术平均值作为两个地址文本的匹配度。

[0107] 最后,可根据预设的匹配条件对两个地址文本进行判别。具体应用中,在两个地址文本的匹配度大于预设阈值时,可将二者确定为匹配文本。这样即实现了待匹配地址文本的最终判别。

[0108] 需要说明的是,虽然以上示例均为中文环境下的地址文本,但本发明的地址文本匹配方法并不局限于中文环境。事实上,一切自然语言下的地址文本均适于利用本发明进行匹配与判别。

[0109] 根据本发明实施例的方法可以看出,本发明首先分析地址文本中的行政区特征,将行政区特征不符的地址文本过滤;之后通过提取两个地址文本在各地址要素级别的特征词、并计算特征词之间的相似度、进而利用相似度计算二者匹配度来实现地址文本的高精度匹配,克服现有技术中数据处理效率较低、无法匹配到户、易受噪声数据影响等缺陷。

[0110] 图2是本发明实施例的地址文本的匹配装置的组成部分示意图。

[0111] 如图2所示,本发明实施例的地址文本的匹配装置200可包括:特征词获取模块201和匹配模块202。其中:

[0112] 特征词获取模块201可用于针对待匹配的两个地址文本中的每一地址文本,利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词;

[0113] 匹配模块202可用于确定所述两个地址文本在同一地址要素级别的特征词的相似度,根据所述相似度得到所述两个地址文本的匹配度;并在匹配度符合预设的匹配条件时,将所述两个地址文本确定为匹配文本。

[0114] 在本发明实施例中,所述装置200可进一步包括:行政区处理模块,其用于在所述利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词之前:利用预先建立的地址分词词典对每一地址文本进行分词,从分词结果中确定每一地址文本在至少一个行政区级别的特征词;其中,任一行政区级别高于任一地址要素级别;当所述两个地址文本在同一行政区级别的特征词存在不符时,将所述两个地址文本确定为不匹配文本。

[0115] 具体应用中,所述至少一个行政区级别为多个行政区级别,所述地址分词词典中包括所述多个行政区级别的行政区名称;以及,行政区处理模块可进一步用于:以预设顺序依次处理分词结果中的每一词:对于分词结果中的第一词,按照行政区级别从高到低的顺序在所述地址分词词典中确定与其相同的一个行政区名称,并将所述第一词作为该地址文本在该行政区名称对应的行政区级别的特征词;对于分词结果中位于所述第一词之后的任一词:从其前一词在所述地址分词词典中下辖的行政区名称中,按照行政区级别从高到低的顺序确定与所述任一词相同的一个行政区名称,并将所述任一词作为该地址文本在该行政区名称对应的行政区级别的特征词。

[0116] 实际应用中,所述地址分词词典中的行政区名称包括行政区的非规范名称。

[0117] 在一个实施例中,行政区处理模块可进一步用于:在所述从分词结果中确定每一地址文本在至少一个行政区级别的特征词之后:当该地址文本缺少一个行政区级别的特征

词、且含有该行政区级别的下一行政区级别的特征词时,将所述地址分词词典中下辖该特征词、且处于该地址文本缺少的行政区级别的行政区名称确定为该地址文本缺少的特征词。

[0118] 在一可选实现方式中,行政区处理模块可进一步用于:当所述两个地址文本在同一行政区级别的特征词不存在不符时,将每一地址文本在每一行政区级别的特征词去除,得到对应于每一地址文本的初选文本。

[0119] 具体应用场景中,特征词获取模块201可进一步用于:以预设顺序依次处理每一初选文本中的每一地址要素关键词:对于该初选文本中的第一个地址要素关键词,将该初选文本在所述第一个地址要素关键词之前的部分确定为所述第一个地址要素关键词对应的地址要素级别的特征词;对于该初选文本中在第一个地址要素关键词之后的任一地址要素关键词,将该初选文本在该任一地址要素关键词与其前一地址要素关键词之间的部分确定为该任一地址要素关键词对应的地址要素级别的特征词。

[0120] 实际应用场景中,所述装置200可进一步包括依赖条件设置模块,用于预先为至少一个地址要素级别设置前置依赖条件;以及,特征词获取模块201可进一步用于:在将该初选文本在该任一地址要素关键词与其前一地址要素关键词之间的部分确定为该任一地址要素关键词对应的地址要素级别的特征词之后,若该地址要素级别设置有前置依赖条件,则在该任一地址要素关键词不符合该前置依赖条件时,将该特征词舍弃。

[0121] 作为一个优选方案,特征词获取模块201可进一步用于:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词之后:若该初选文本中存在同一地址要素级别的多个特征词,则查询为该地址要素级别预设的限制策略是否支持多个特征词:若不支持,将所述多个特征词舍弃。

[0122] 较佳地,特征词获取模块201可进一步用于:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词之后,在该初选文本中不存在预设的匹配粒度级别的特征词时,将所述两个地址文本确定为不匹配文本。

[0123] 特别地,在本发明实施例中,所述多个行政区级别为省级、地级、县级;所述地址要素级别为以下一种或多种:镇级、村级、组级、路级、道路号级、栋级、单元级、层级、户级以及标志地名级;其中:镇级的地址要素关键词为:镇、乡、街道;村级的地址要素关键词为:村、居委会、社区;组级的地址要素关键词为:组;路级的地址要素关键词为:路、街、道、巷、弄、弄堂、胡同、大道、大路、大街;道路号级的地址要素关键词为:号;栋级的地址要素关键词为:栋、幢、座、楼、号楼;单元级的地址要素关键词为:单元;层级的地址要素关键词为:层、楼;户级的地址要素关键词为:户、室、屋、房、号、门;标志地名级的地址要素关键词为:广场、酒店、饭店、宾馆、中心、大厦、大楼、商城、公园、小区、花园、公司、寓所、会所、市场、社区、新村、苑、开发区、科技区、园区。

[0124] 在一个实施例中,组级的前置依赖条件为:其在初选文本中的前一地址要素关键词对应于村级;层级的前置依赖条件为:其在初选文本的前一地址要素关键词对应于单元级或栋级。

[0125] 在可选的实现方式中,特征词获取模块201可进一步用于:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词之后:若该初选文本中存在路级的两个特征词,则将二者按照预设排列顺序拼接为一个词,作为该初选文本在路级的特征词。

[0126] 较佳地,在本发明实施例中,特征词获取模块201可进一步用于:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词时:若路级的地址要素关键词的后一地址要素关键词为号,则确定该地址要素关键词对应于道路号级;若路级的地址要素关键词的后一地址要素关键词不为号、且该路级的地址要素关键词在后邻接至少一位数字,则确定该数字为道路号级的特征词;若栋级、单元级或层级的地址要素关键词的后一地址要素关键词为号,则确定该地址要素关键词对应于户级;若栋级、单元级或层级的地址要素关键词在后邻接至少一位数字、且在后不存在地址要素关键词,则确定该数字为户级的特征词。

[0127] 在本发明实施例中,特征词获取模块201可进一步用于:在所述以预设顺序依次处理每一初选文本中的每一地址要素关键词时:若道路号级或标志地名级的地址要素关键词的后一地址要素关键词为楼,则确定该地址要素关键词对应于栋级;若栋级或单元级的地址要素关键词的后一地址要素关键词为楼,则确定该地址要素关键词对应于层级。

[0128] 实际应用中,匹配模块202可进一步用于:当所述两个地址文本在该地址要素级别均存在特征词时,将所述相似度确定为两个特征词的编辑距离与两个特征词中字符串长度最大值之商;当所述两个地址文本之一在该地址要素级别存在特征词、另一在该地址要素级别不存在特征词时,将所述相似度确定为预设阈值;以及,当所述两个地址文本在该地址要素级别均不存在特征词时,将所述相似度确定为零。

[0129] 具体应用场景中,所述装置200可进一步包括权重配置模块,用于为每一地址要素级别预先配置权重值;以及,匹配模块202可进一步用于:将所述两个地址文本在每一地址要素级别的特征词的相似度的加权平均值作为所述两个地址文本的匹配度。

[0130] 此外,在本发明实施例中,预设的匹配条件为:匹配度大于所述预设阈值。

[0131] 在本发明实施例的技术方案中,可首先分析地址文本中的行政区特征,将行政区特征不符的地址文本过滤;之后通过提取两个地址文本在各地址要素级别的特征词、并计算特征词之间的相似度、进而利用相似度计算二者匹配度来实现地址文本的高精度匹配,克服现有技术中数据处理效率较低、无法匹配到户、易受噪声数据影响等缺陷。

[0132] 图3示出了可以应用本发明实施例的地址文本的匹配方法或地址文本的匹配装置的示例性系统架构300。

[0133] 如图3所示,系统架构300可以包括终端设备301、302、303,网络304和服务器305(此架构仅仅是示例,具体架构中包含的组件可以根据申请具体情况调整)。网络304用以在终端设备301、302、303和服务器305之间提供通信链路的介质。网络304可以包括各种连接类型,例如有线、无线通信链路或者光纤电缆等等。

[0134] 用户可以使用终端设备301、302、303通过网络304与服务器305交互,以接收或发送消息等。终端设备301、302、303上可以安装有各种通讯客户端应用,例如购物类应用、网页浏览器应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等(仅为示例)。

[0135] 终端设备301、302、303可以是具有显示屏并且支持网页浏览的各种电子设备,包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。

[0136] 服务器305可以是提供各种服务的服务器,例如对用户利用终端设备301、302、303所浏览的购物类网站提供支持的后台管理服务器(仅为示例)。后台管理服务器可以对接收到的产品信息查询请求等数据进行分析等处理,并将处理结果(例如目标推送信息、产品信息--仅为示例)反馈给终端设备。

[0137] 需要说明的是,本发明实施例所提供的地址文本的匹配方法一般由服务器305执行,相应地,地址文本的匹配装置一般设置于服务器305中。

[0138] 应该理解,图3中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器。

[0139] 本发明还提供了一种电子设备。本发明实施例的电子设备包括:一个或多个处理器;存储装置,用于存储一个或多个程序,当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现本发明所提供的地址文本的匹配方法。

[0140] 下面参考图4,其示出了适于用来实现本发明实施例的电子设备的计算机系统400的结构示意图。图4示出的电子设备仅仅是一个示例,不应对本发明实施例的功能和使用范围带来任何限制。

[0141] 如图4所示,计算机系统400包括中央处理单元(CPU) 401,其可以根据存储在只读存储器(ROM) 402中的程序或者从存储部分408加载到随机访问存储器(RAM) 403中的程序而执行各种适当的动作和处理。在RAM403中,还存储有计算机系统400操作所需的各种程序和数据。CPU401、ROM 402以及RAM 403通过总线404彼此相连。输入/输出(I/O) 接口405也连接至总线404。

[0142] 以下部件连接至I/O接口405:包括键盘、鼠标等的输入部分406;包括诸如阴极射线管(CRT)、液晶显示器(LCD) 等以及扬声器等的输出部分407;包括硬盘等的存储部分408;以及包括诸如LAN卡、调制解调器等网络接口卡的通信部分409。通信部分409经由诸如因特网的网络执行通信处理。驱动器410也根据需要连接至I/O接口405。可拆卸介质411,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器410上,以便从其上读出的计算机程序根据需要被安装入存储部分408。

[0143] 特别地,根据本发明公开的实施例,上文的主要步骤图描述的过程可以被实现为计算机软件程序。例如,本发明实施例包括一种计算机程序产品,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行主要步骤图所示的方法的程序代码。在上述实施例中,该计算机程序可以通过通信部分409从网络上被下载和安装,和/或从可拆卸介质411被安装。在该计算机程序被中央处理单元401执行时,执行本发明的系统中限定的上述功能。

[0144] 需要说明的是,本发明所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本发明中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。在本发明中,计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执

行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:无线、电线、光缆、RF等等,或者上述的任意合适的组合。

[0145] 附图中的流程图和框图,图示了按照本发明各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这根据所涉及的功能而定。也要注意,框图或流程图中的每个方框、以及框图或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0146] 描述于本发明实施例中所涉及到的模块可以通过软件的方式实现,也可以通过硬件的方式来实现。所描述的模块也可以设置在处理器中,例如,可以描述为:一种处理器包括特征词获取模块和匹配模块。其中,这些模块的名称在某种情况下并不构成对该模块本身的限定,例如,特征词获取模块还可以被描述为“向匹配模块发送特征词的模块”。

[0147] 作为另一方面,本发明还提供了一种计算机可读介质,该计算机可读介质可以是上述实施例中描述的设备中所包含的;也可以是单独存在,而未装配入该设备中的。上述计算机可读介质承载有一个或者多个程序,当上述一个或者多个程序被该设备执行时,使得该设备执行的步骤包括:对于待匹配的两个地址文本中的每一地址文本:利用该地址文本含有的地址要素关键词,获取其在所述地址要素关键词对应的地址要素级别的特征词;确定所述两个地址文本在同一地址要素级别的特征词的相似度,根据所述相似度得到所述两个地址文本的匹配度;并在匹配度符合预设的匹配条件时,将所述两个地址文本确定为匹配文本。

[0148] 根据本发明实施例的方法可以看出,本发明首先分析地址文本中的行政区特征,将行政区特征不符的地址文本过滤;之后通过提取两个地址文本在各地址要素级别的特征词、并计算特征词之间的相似度、进而利用相似度计算二者匹配度来实现地址文本的高精度匹配,克服现有技术中数据处理效率较低、无法匹配到户、易受噪声数据影响等缺陷。

[0149] 上述具体实施方式,并不构成对本发明保护范围的限制。本领域技术人员应该明白的是,取决于设计要求和因素,可以发生各种各样的修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等,均应包含在本发明保护范围之内。

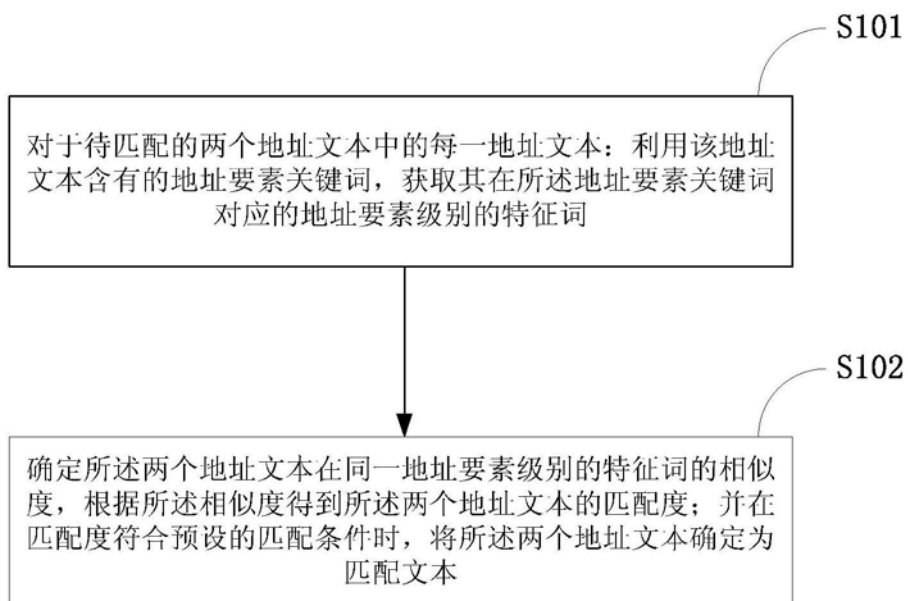


图1

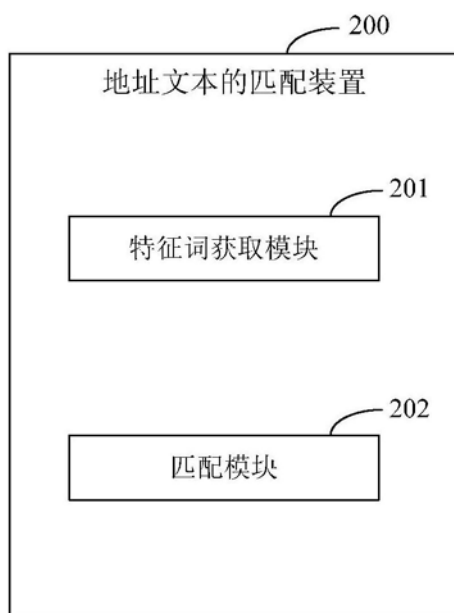


图2

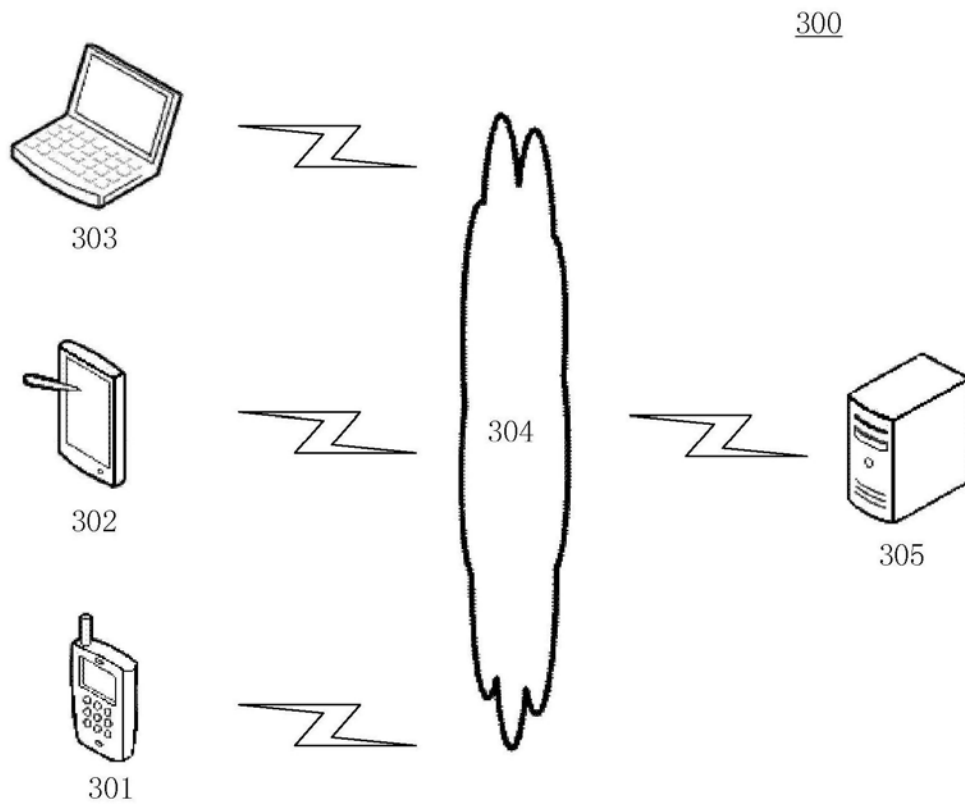


图3

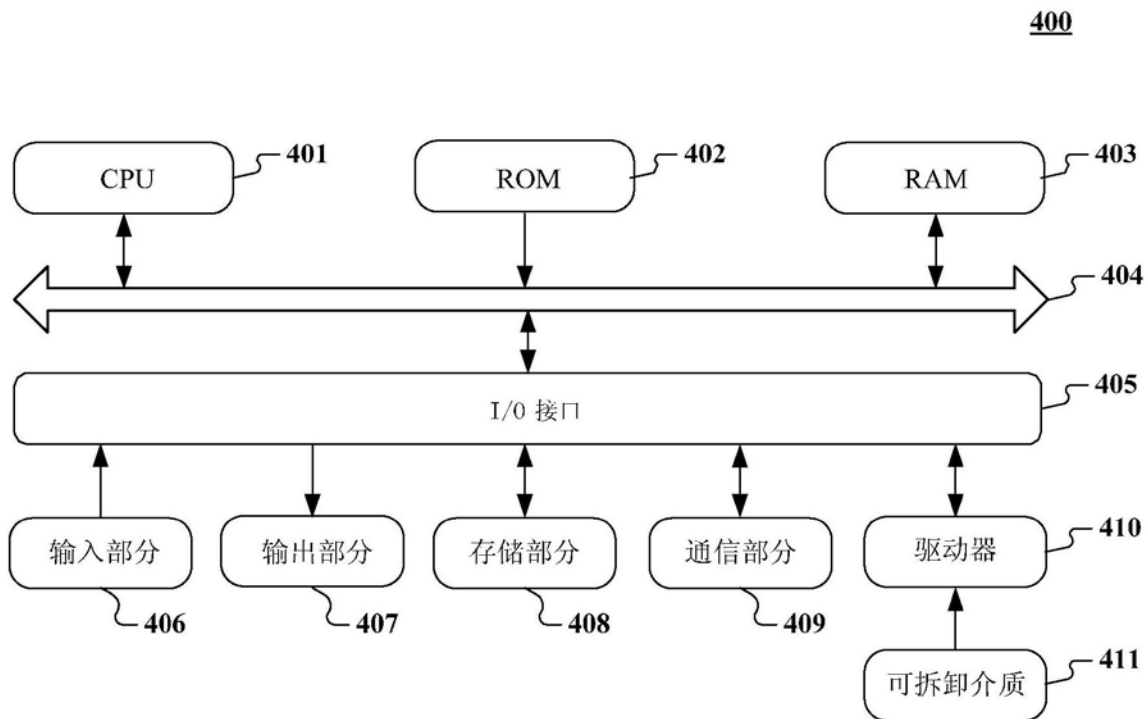


图4