

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2016 年 10 月 20 日 (20.10.2016)



(10) 国际公布号
WO 2016/165538 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2016/077297
- (22) 国际申请日: 2016 年 3 月 25 日 (25.03.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510172985.0 2015 年 4 月 13 日 (13.04.2015) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 英属开曼群岛大开曼乔治城资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 及
- (71) 申请人 (仅对美国): 吴保华 (WU, Baohua) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。
- (74) 代理人: 北京国昊天诚知识产权代理有限公司 (CO-HORIZON INTELLECTUAL PROPERTY INC.); 中国北京市朝阳区小关北里甲 2 号渔阳置业大厦 B 座 605, Beijing 100029 (CN)。

- (81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。
- (84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

(54) Title: ADDRESS DATA MANAGEMENT METHOD AND DEVICE

(54) 发明名称: 一种地址数据的管理方法和装置

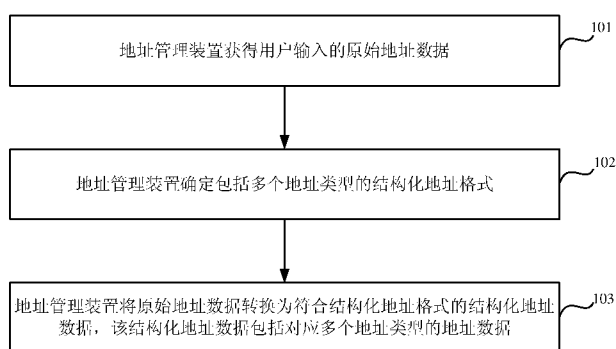


图 1

- 101 An address management device acquires original address data input by a user
- 102 The address management device determines a structured address format comprising multiple address types
- 103 The address management device converts the original address data into structured address data satisfying the structured address format, the structured address data comprises address data corresponding to the multiple address types

(57) Abstract: Disclosed are an address data management method and device. The method comprises: an address management device acquires original address data input by a user; the address management device determines a structured address format comprising multiple address types; and the address management device converts the original address data into structured address data satisfying the structured address format, the structured address data comprises address data corresponding to the multiple address types. In embodiments of the present application, by setting a structured address format comprising multiple address types and generating structured address data satisfying the structured address format, normalized and standardized address data is generated, the problem of failing to standardize text addresses is solved, similarities and differences among different text addresses can be determined, and the related homing of the text addresses can be identified.

(57) 摘要: 本申请公开了一种地址数据的管理方法和装置, 该方法包括: 地址管理装置获得用户输入的原始地址数据; 所述地址管理装置确定包括多个地址类型的结构化

化地址格式; 所述地址管理装置将所述原始地址数据转换为符合所述结构化地址格式的结构化地址数据, 所述结构化地址数据包括对应多个地址类型的地址数据。本申请实施例中, 通过设置包括多个地址类型的结构化地址格式, 并生成符合结构化地址格式的结构化地址数据, 从而生成规范化、标准化的地址数据, 解决无法对文本地址进行规范化的问题, 并能够判断不同文本地址间的异同性, 能够识别文本地址的相关归属。

一种地址数据的管理方法和装置

技术领域

5 本申请涉及通信技术领域，尤其涉及一种地址数据的管理方法和装置。

背景技术

在电子商务网站和物流系统中产生了大量文本地址，这些文本地址的输入格式和地址元素因用户而不同。例如，用户A输入的文本地址只包括门牌号
10 信息，用户B输入的文本地址只包括POI（Point of Interest，兴趣点）信息，用户C输入的文本地址包括错误的区县或门牌号信息。这些文本地址缺乏规范化、标准化，无法判断不同文本地址间的异同性，无法识别文本地址的相关归属。其中，地址元素是指文本地址中的各级元素，如省、市、区、开发区、镇、路、POI等。POI可以是一栋房子、一个商铺、一个邮筒、一个公交站等。
15

发明内容

本申请实施例提供一种地址数据的管理方法和装置，以生成规范化、标准化的地址数据，从而解决无法对文本地址进行规范化的问题。

本申请实施例提供一种地址数据的管理方法，所述方法包括以下步骤：
20 地址管理装置获得用户输入的原始地址数据；

所述地址管理装置确定包括多个地址类型的结构化地址格式；

所述地址管理装置将所述原始地址数据转换为符合所述结构化地址格式的结构化地址数据，所述结构化地址数据包括对应多个地址类型的地址数据。

所述地址管理装置将所述原始地址数据转换为符合所述结构化地址格式
25 的结构化地址数据，具体包括：

所述地址管理装置基于多个地址类型对原始地址数据进行预处理；

所述地址管理装置基于多个地址类型对预处理后的地址数据进行切分；
所述地址管理装置基于多个地址类型对切分后地址数据进行补全校验；
所述地址管理装置对补全校验后的地址数据进行规范化处理，以得到符合所述结构化地址格式的结构化地址数据。

5 所述地址管理装置基于多个地址类型对原始地址数据进行预处理的过程，具体包括：

所述地址管理装置从所述原始地址数据中筛选出未对应所述多个地址类型的地址数据，从所述原始地址数据中删除当前筛选的地址数据，并将所述原始地址数据中存在的非规范格式的地址数据转换为规范格式的地址数据。

10 所述地址管理装置基于多个地址类型对预处理后的地址数据进行切分的过程，具体包括：

所述地址管理装置获得所述多个地址类型对应的分词器词典，利用所述多个地址类型对应的分词器词典切分出对应所述多个地址类型的地址数据。

15 所述地址管理装置基于多个地址类型对切分后地址数据进行补全校验的过程，具体包括：

所述地址管理装置校验切分后地址数据是否已经包含对应所述多个地址类型的地址数据；如果否，则所述地址管理装置确定切分后地址数据中不包含的地址类型，并基于历史数据补全所述地址类型的地址数据。

20 所述地址管理装置对补全校验后的地址数据进行规范化处理的过程，具体包括：所述地址管理装置利用拼音相似度算法对补全校验后的地址数据进行规范化处理；和/或，所述地址管理装置利用基于概率检索模型的兴趣点 POI 规范化算法对补全校验后的地址数据进行规范化处理。

本申请实施例提供一种地址管理装置，所述地址管理装置具体包括：

获得模块，用于获得用户输入的原始地址数据；

25 确定模块，用于确定包括多个地址类型的结构化地址格式；

处理模块，用于将所述原始地址数据转换为符合所述结构化地址格式的结构化地址数据，所述结构化地址数据包括对应多个地址类型的地址数据。

所述处理模块包括：预处理子模块，用于基于多个地址类型对原始地址数据进行预处理；切分子模块，用于基于多个地址类型对预处理后的地址数据
5 据进行切分；补全子模块，用于基于多个地址类型对切分后地址数据进行补全校验；规范化子模块，用于对补全校验后的地址数据进行规范化处理，以得到符合所述结构化地址格式的结构化地址数据。

所述预处理子模块，具体用于从原始地址数据中筛选出未对应所述多个地址类型的地址数据，从原始地址数据中删除当前筛选的地址数据，并将原
10 始地址数据中存在的非规范格式的地址数据转换为规范格式的地址数据。

所述切分子模块，具体用于获得多个地址类型对应的分词器词典，利用多个地址类型对应的分词器词典切分出对应所述多个地址类型的地址数据。

所述补全子模块，具体用于校验切分后的地址数据是否已经包含对应所述多个地址类型的地址数据；如果否，则确定切分后的地址数据中不包含的
15 地址类型，并基于历史数据补全所述地址类型的地址数据。

所述规范化子模块，具体用于利用拼音相似度算法对补全校验后的地址数据进行规范化处理；和/或，利用基于概率检索模型的兴趣点 POI 规范化算法对补全校验后的地址数据进行规范化处理。

与现有技术相比，本申请实施例至少具有以下优点：本申请实施例中，
20 通过设置包括多个地址类型的结构化地址格式，并生成符合结构化地址格式的结构化地址数据，从而生成规范化、标准化的地址数据，解决无法对文本地址进行规范化的问题，并能够判断不同文本地址间的异同性，能够识别文本地址的相关归属。具体的，通过对海量历史文本地址中的地址数据进行识别和提取，通过学习的方式从中学习出地址数据之间的知识和规则，并将学
25 习的知识和规则对漏写地址数据进行补全、对错误地址数据进行校验，对非

规范地址数据进行规范化处理，重新生成一条分级的结构化地址数据。

附图说明

为了更加清楚地说明本申请实施例的技术方案，下面将对本申请实施例
5 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申请的一些实施例，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，还可以根据本申请实施例的这些附图获得其他的附图。

图 1 是本申请实施例一提供的一种地址数据的管理方法流程示意图；

图 2 是本申请实施例二提供的一种地址管理装置的结构示意图。

10

具体实施方式

下面将结合本申请实施例中的附图，对本申请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申请的一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有
15 做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

实施例一

针对现有技术中存在的问题，本申请实施例一提供一种地址数据的管理方法，如图 1 所示，该地址数据的管理方法具体可以包括以下步骤：

步骤 101，地址管理装置获得用户输入的原始地址数据。

20 本申请实施例中，地址管理装置内可以配置整合模块，整合模块用于将各方地址数据源进行整合，生成唯一的 key（密钥），并装入文本地址库。其中，文本地址库中的针对一个 key 的地址数据，即用户输入的原始地址数据。

步骤 102，地址管理装置确定包括多个地址类型的结构化地址格式。

其中，结构化地址格式中包括的多个地址类型具体包括但不限于以下之
25 一或者任意组合：省、市、区县、乡镇（街道办）、开发区、主路、主路门牌号、支路、支路门牌号、标志性 POI（楼盘等）、幢、单元（楼层）、房间号等。

步骤 103, 地址管理装置将原始地址数据转换为符合结构化地址格式的结构化地址数据, 该结构化地址数据包括对应多个地址类型的地址数据。

例如, 在地址管理装置生成的符合结构化地址格式的结构化地址数据中, 可以包括对应于省的地址数据、对应于市的地址数据、对应于区县的地址数据、对应于乡镇(街道办)、对应于开发区的地址数据、对应于主路的地址数据、对应于主路门牌号的地址数据、对应于支路的地址数据、对应于支路门牌号的地址数据、对应于标志性 POI(楼盘等)的地址数据、对应于幢的地址数据、对应于单元(楼层)的地址数据、对应于房间号的地址数据等。

本申请实施例中, 地址管理装置将原始地址数据转换为符合结构化地址格式的结构化地址数据的过程, 具体包括但不限于: 地址管理装置基于多个地址类型对原始地址数据进行预处理; 之后, 地址管理装置基于多个地址类型对预处理后的地址数据进行切分; 之后, 地址管理装置基于多个地址类型对切分后地址数据进行补全校验; 之后, 地址管理装置对补全校验后的地址数据进行规范化处理, 以得到符合结构化地址格式的结构化地址数据。

本申请实施例中, 地址管理装置基于多个地址类型对原始地址数据进行预处理的过程, 具体包括: 地址管理装置从原始地址数据中筛选出未对应多个地址类型的地址数据, 从原始地址数据中删除当前筛选的地址数据, 并将原始地址数据中存在的非规范格式的地址数据转换为规范格式的地址数据。

本申请实施例中, 地址管理装置内可以配置预处理模块, 由该预处理模块从原始地址数据中筛选出未对应多个地址类型的地址数据, 并从原始地址数据中删除当前筛选的地址数据。进一步的, 由该预处理模块将原始地址数据中存在的非规范格式的地址数据转换为规范格式的地址数据。

其中, 由于用户输入的原始地址数据是用户填写的, 具有随意性, 因此原始地址数据中会包含对应多个地址类型的地址数据, 如河北省、保定市等地址数据, 原始地址数据中也会包含未对应多个地址类型的地址数据, 如话

费充值信息、虚拟游戏点卡信息等，这些未对应多个地址类型的地址数据是需要进行数据清洗的。基于此，预处理模块从原始地址数据中筛选出未对应多个地址类型的地址数据，并从原始地址数据中删除当前筛选的地址数据。

其中，由于用户输入的原始地址数据是用户填写的，具有随意性，因此
5 原始地址数据中会存在非规范格式的地址数据。如英文；数字写为全角；非香港、澳门、台湾地区的地址存在繁体地址现象；香港、澳门、台湾地区的地址存在简体地址现象；门牌号的地址存在中文现象（如二十号）；以数字命名的道路名出现数字现象（如文 2 路）等。基于此，预处理模块将原始地址数据中存在的非规范格式的地址数据转换为规范格式的地址数据。其中，规范格式的地址数据包括但不限于：英文、数字的全角更改为半角；大陆地址
10 一律规范格式为简体中文；港澳台地区的地址一律规范格式为繁体中文；道路名一律规范格式为中文；门牌号、房间号等一律规范格式为数字。

本申请实施例中，地址管理装置基于多个地址类型对预处理后的地址数据进行切分的过程，具体包括但不限于如下方式：地址管理装置获得多个地
15 址类型对应的分词器词典，并利用该多个地址类型对应的分词器词典将预处理后的地址数据切分出对应于这多个地址类型的地址数据。例如，基于多个地址类型对应的分词器词典，地址管理装置可以将预处理后的地址数据切分出对应于省的地址数据、对应于市的地址数据、对应于区县的地址数据、对应于乡镇（街道办）、对应于开发区的地址数据、对应于主路的地址数据、对应于主路门牌号的地址数据、对应于支路的地址数据、对应于支路门牌号的
20 地址数据、对应于标志性 POI（楼盘等）的地址数据、对应于幢的地址数据、对应于单元（楼层）的地址数据、对应于房间号的地址数据等。

本申请实施例中，地址管理装置内可以配置切分模块，由该切分模块获得多个地址类型对应的分词器词典，并利用该多个地址类型对应的分词器词典将预处理后的地址数据切分出对应于这多个地址类型的地址数据。
25

其中，分词器词典包括但不限于：省、市、区县词典；乡镇词典；工业区词典；村庄词典；街道词典；高校词典；社区标准词典；社区自学习词典。

其中，在切分模块利用分词器词典将预处理后的地址数据切分出对应于多个地址类型的地址数据的过程中，则相应的切分算法具体包括：前向有限状态最大匹配算法，其切分规则包括：基于关键字切分，如：镇，街，路，公司，大厦，中学，门牌号，社区详细地址（幢、单元、房间号）等。进一步的，相应的切分流程具体包括：省、市、区切分：采用基于省市区词典初始化的分词器切割详细地址，若切分后的省、市、区与原始的省、市、区字段不同，则替换，并减少后续切分误差，保留剩余地址。乡镇（工业区）切分：采用基于乡镇（工业区）词典初始化分词器（以市为单元共 362 个）切分上一步的剩余地址；若分词器切分失败则切分详细地址；若仍切分失败则采用乡镇规则切分，并标记后续处理。道路切分：与乡镇（工业区）切分流程类似，只是采用乡镇词典初始化 362 个道路分词器。门牌号切分：采用相应的切分规则进行切分。社区（楼盘）切分：采用社区词典初始化社区分词器（以市为单元共 362 个），切分上一步的剩余地址；若分词器切分失败则切分详细地址；若切分出两个社区元素，则字符串长度最大的作为社区元素；若仍切分失败则采用自学习词典的分词器切分详细地址；若仍切分失败则采用社区规则切分，并将采用自学习词典或社区规则切分的社区标记后续处理。社区内详细地址切分（幢、单元、房间号）：采用相应的切分规则进行切分。

本申请实施例中，地址管理装置基于多个地址类型对切分后地址数据进行补全校验的过程，具体包括但不限于：地址管理装置校验切分后地址数据是否已经包含对应所有多个地址类型的地址数据；如果否，则地址管理装置确定切分后的地址数据中不包含的地址类型，并基于历史数据补全该地址类型的地址数据；如果是，则地址管理装置不需要补全相应的地址数据。

例如，当地址管理装置基于多个地址类型切分出对应于省的地址数据、

对应于区县的地址数据、对应于开发区的地址数据、对应于主路的地址数据、对应于主路门牌号的地址数据、对应于支路的地址数据、对应于支路门牌号的地址数据、对应于单元（楼层）的地址数据时，则：地址管理装置校验出切分后的地址数据未包含对应所有多个地址类型的地址数据，并基于历史数据补全对应于市的地址数据、对应于乡镇（街道办）、对应于标志性 POI（楼盘等）的地址数据、对应于幢的地址数据、对应于房间号的地址数据。

本申请实施例中，地址管理装置内可以配置补全校验模块，由该补全校验模块校验切分后地址数据是否已经包含对应所有多个地址类型的地址数据；如果否，则确定切分后的地址数据中不包含的地址类型，并基于历史数据补全该地址类型的地址数据；如果是，则不需要补全相应的地址数据。

其中，地址数据中存在大量非正确的地址数据，如正确地址数据：杭州市文二路 391 号西湖国际科技大厦 B 座 2 楼小邮局，而用户填写如下非标准或不正确的地址数据：杭州市文二路 391 号 2 楼小邮局；杭州市文二路西湖国际科技大厦 B 座 2 楼小邮局；杭州市文二路 380 号西湖国际科技大厦 B 座 2 楼小邮局。基于上述情况，补全校验模块在地址数据处理过程中，对上述情况进行处理，在切分后的地址数据的门牌号或社区字段进行补全与校正。

其中，基于结构地址标准库，则可以将结构地址标准库中的每条地址数据采用相应的切分算法进行结构化为：市+区县+道路+门牌号+社区。统计以上 5 个字段都完全的地址频次。筛选地址频次大于 3 的地址。统计市+区县+道路+门牌号下每个社区的使用频次，并保留频次最大的市+区县+道路+门牌号+社区，并将其加入结构地址标准库中。或者，基于结构地址标准库，则可以将结构地址标准库中的每条地址数据采用相应的切分算法进行结构化为：市+道路+门牌号+社区。统计以上 4 个字段都完全的地址频次。筛选地址频次大于等于 1 的地址。统计市+道路+门牌号下每个社区的使用频次，并保留频次最大的市+道路+门牌号+社区，并将其加入结构地址标准库中。

基于结构地址标准库，则在地址数据的补全与校正过程中，假设市+区县+道路+门牌号下仅有一个社区，针对每一条已结构化的地址数据，如果社区字段为 null（空）或者为规则切分或者为自学习词典分词器切分，则可以从结构地址标准库中查询市+区县+道路+门牌号为 key 的社区，并补全或者校正社区

5 区字段。进一步的，基于结构地址标准库，假设市+区县+道路+社区下仅有一个门牌号，针对每一条已结构化的地址数据，如果门牌号为 null 或者为规则切分或者为自学习词典分词器切分，则可以从结构地址标准库中查询市+区县+道路+社区为 key 的门牌号，并补全或者校正门牌号字段。

本申请实施例中，地址管理装置对补全校验后的地址数据进行规范化处理的过程，具体包括但不限于如下方式：地址管理装置利用拼音相似度算法对补全校验后的地址数据进行规范化处理；和/或，地址管理装置利用基于概率检索模型的 POI 规范化算法对补全校验后的地址数据进行规范化处理。

10

本申请实施例中，地址管理装置内可以配置规范化模块，规范化模块利用拼音相似度算法对补全校验后的地址数据进行规范化处理；和/或，利用基于概率检索模型的 POI 规范化算法对补全校验后的地址数据进行规范化处理。

15

其中，用户填写的地址数据中存在大量的地址数据的简称、缩写、错别字、谐音等非规范现象。如标准地址数据为西湖国际科技大厦，非规范化的地址数据为西湖国际（缩写）；标准地址数据为浙江大学第一附属医院，非规范化的地址数据为浙大一附院（简称）；标准地址数据为古墩路，非规范化的地址数据为古吨路（谐音）；标准地址数据为保淑路，非规范化的地址数据为保椒路（错别字）。虽然在地址结构化过程中能够将这些地址数据切分出来，但由于多名称现象在地址坐标标注及后续的地址数据分析中存在很大的困难和弊端，因此，规范化模块需要对非规范化的地址数据进行规范化处理。

20

进一步的，规范化模块对非规范化的地址数据进行规范化处理的算法包括但不限于：拼音相似度算法、基于概率检索模型的 POI 规范化算法。

25

针对拼音相似度算法：规范化模块将非规范化的地址数据和规范化的地址数据转换为拼音，计算相似距离（如最小编辑距离），并将高于阈值且相似度最高的规范化的地址数据作为非规范化的地址数据的标准化地址数据。

5 针对基于概率检索模型的 POI 规范化算法，规范化模块将识别出来的类 POI 进行 bigram（二元语法）切分，然后对于同时出现在类 POI 和候选标准 POI 中的 bigram，累加每个 bigram 的估值，各 bigram 的估值的和就是候选标准 POI 与类 POI 的相关性度量。进一步的，计算出候选 POI 的相关性得分，并对这些 POI 得分进行从大到小的排序，筛选出 POI 类型、POI 的区县与地址类型以及地址对应的区县相符的且得分最大的 POI，即为规范 POI。

10 为了实现上述过程，可以采用如下的 BM25（二元独立模型）计算公式：

$$S = \sum_{i \in Q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} * \frac{k + 1}{K + 1} * \frac{w}{I}$$

$$K = k \left((1 - b) + b * \frac{dl}{avdl} \right)$$

$$I = (1 - b) + b * \frac{index_i}{avg_index_i}$$

$$w = \begin{cases} 0.35 & index_i < 0.333 \\ 0.5 & 0.333 \leq index_i < 0.66 \\ 0.15 & index_i \geq 0.66 \end{cases}$$

15 其中，上述四个公式的相关参数说明如下所示：

	相关 POI	不相关 POI	POI 数量
$b_i=1$	r_i	$n_i - r_i$	n_i
$b_i=0$	$R - r_i$	$(N - R) - (n_i - r_i)$	$N - n_i$
POI 个数	R	$N - R$	N

进一步的，S：候选 POI 的相关性得分；N：一个城市或者区县的 POI 数

量; R: 与类 POI 具有两个相同的 bigram 且 jaccard (相似性系数) 相似度大于 0.4 的相关 POI 数量; n_i : 为包含 bigram b_i 的 POI 数量; dl: 当前候选标准 POI 中的 bigram 个数; avdl: 平均每个候选标准 POI 包含的 bigram 个数; r_i : 为 n_i 中的相关 POI 数量; $index_i$: b_i 在当前 POI 中出现的位置次序; $avgindex_i$: b_i 在包含其的 POI 中出现的平均位置次序; k, b: 为自由调节参数, 根据经验 k 设置为: 1.2, b 设置为 0.75; K, I: 为公式中的临时变量。

与现有技术相比, 本申请实施例至少具有以下优点: 本申请实施例中, 通过设置包括多个地址类型的结构化地址格式, 并生成符合结构化地址格式的结构化地址数据, 从而生成规范化、标准化的地址数据, 解决无法对文本地址进行规范化的问题, 并能够判断不同文本地址间的异同性, 能够识别文本地址的相关归属。具体的, 通过对海量历史文本地址中的地址数据进行识别和提取, 通过学习的方式从中学习出地址数据之间的知识和规则, 并将学习的知识和规则对漏写地址数据进行补充、对错误地址数据进行校验, 对非规范地址数据进行规范化处理, 重新生成一条分级的结构化地址数据。

基于与上述方法同样的申请构思, 本申请实施例中还提供了一种地址管理装置, 如图 2 所示, 所述地址管理装置具体包括:

获得模块 11, 用于获得用户输入的原始地址数据;

确定模块 12, 用于确定包括多个地址类型的结构化地址格式;

处理模块 13, 用于将所述原始地址数据转换为符合所述结构化地址格式的结构化地址数据, 所述结构化地址数据包括对应多个地址类型的地址数据。

其中, 所述处理模块 13 具体包括: 预处理子模块 131, 用于基于多个地址类型对原始地址数据进行预处理; 切分子模块 132, 用于基于多个地址类型对预处理后的地址数据进行切分; 补全子模块 133, 用于基于多个地址类型对切分后地址数据进行补全校验; 规范化子模块 134, 用于对补全校验后的地址数据进行规范化处理, 以得到符合所述结构化地址格式的结构化地址数据。

所述预处理子模块 131, 具体用于从原始地址数据中筛选出未对应所述多个地址类型的地址数据, 从原始地址数据中删除当前筛选的地址数据, 并将原始地址数据中存在的非规范格式的地址数据转换为规范格式的地址数据。

5 所述切分子模块 132, 具体用于获得多个地址类型对应的分词器词典, 利用多个地址类型对应的分词器词典切分出对应多个地址类型的地址数据。

所述补全子模块 133, 具体用于校验切分后的地址数据是否已经包含对应所述多个地址类型的地址数据; 如果否, 则确定切分后的地址数据中不包含的地址类型, 并基于历史数据补全所述地址类型的地址数据。

10 所述规范化子模块 134, 具体用于利用拼音相似度算法对补全校验后的地址数据进行规范化处理; 和/或, 利用基于概率检索模型的兴趣点 POI 规范化算法对补全校验后的地址数据进行规范化处理。

其中, 本申请装置的各个模块可以集成于一体, 也可以分离部署。上述模块可以合并为一个模块, 也可以进一步拆分成多个子模块。

15 通过以上的实施方式的描述, 本领域的技术人员可以清楚地了解到本申请可借助软件加必需的通用硬件平台的方式来实现, 当然也可以通过硬件, 但很多情况下前者是更佳的实施方式。基于这样的理解, 本申请的技术方案本质上或者说对现有技术做出贡献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
20 本申请各个实施例所述的方法。本领域技术人员可以理解附图只是一个优选实施例的示意图, 附图中的模块或流程并不一定是实施本申请所必须的。本领域技术人员可以理解实施例中的装置中的模块可以按照实施例描述进行分布于实施例的装置中, 也可以进行相应变化位于不同于本实施例的一个或多个装置中。上述实施例的模块可以合并为一个模块, 也可以进一步
25 拆分成多个子模块。上述本申请实施例序号仅仅为了描述, 不代表实施例

的优劣。以上公开的仅为本申请的几个具体实施例，但是，本申请并非局限于此，任何本领域的技术人员能思之的变化都应落入本申请的保护范围。

权 利 要 求 书

1、一种地址数据的管理方法，其特征在于，所述方法包括以下步骤：

地址管理装置获得用户输入的原始地址数据；

所述地址管理装置确定包括多个地址类型的结构化地址格式；

5 所述地址管理装置将所述原始地址数据转换为符合所述结构化地址格式的结构化地址数据，所述结构化地址数据包括对应多个地址类型的地址数据。

2、如权利要求 1 所述的方法，其特征在于，所述地址管理装置将所述原始地址数据转换为符合所述结构化地址格式的结构化地址数据，具体包括：

所述地址管理装置基于多个地址类型对原始地址数据进行预处理；

10 所述地址管理装置基于多个地址类型对预处理后的地址数据进行切分；

所述地址管理装置基于多个地址类型对切分后地址数据进行补全校验；

所述地址管理装置对补全校验后的地址数据进行规范化处理，以得到符合所述结构化地址格式的结构化地址数据。

3、如权利要求 2 所述的方法，其特征在于，所述地址管理装置基于多个地址类型对原始地址数据进行预处理的过程，具体包括：

所述地址管理装置从所述原始地址数据中筛选出未对应所述多个地址类型的地址数据，从所述原始地址数据中删除当前筛选的地址数据，并将所述原始地址数据中存在的非规范格式的地址数据转换为规范格式的地址数据。

4、如权利要求 2 所述的方法，其特征在于，所述地址管理装置基于多个地址类型对预处理后的地址数据进行切分的过程，具体包括：

所述地址管理装置获得所述多个地址类型对应的分词器词典，利用所述多个地址类型对应的分词器词典切分出对应所述多个地址类型的地址数据。

5、如权利要求 2 所述的方法，其特征在于，所述地址管理装置基于多个地址类型对切分后地址数据进行补全校验的过程，具体包括：

25 所述地址管理装置校验切分后地址数据是否已经包含对应所述多个地址

类型的地址数据；如果否，则所述地址管理装置确定切分后地址数据中不包含的地址类型，并基于历史数据补全所述地址类型的地址数据。

6、如权利要求 2 所述的方法，其特征在于，所述地址管理装置对补全校验后的地址数据进行规范化处理的过程，具体包括：

5 所述地址管理装置利用拼音相似度算法对补全校验后的地址数据进行规范化处理；和/或，所述地址管理装置利用基于概率检索模型的兴趣点 POI 规范化算法对补全校验后的地址数据进行规范化处理。

7、一种地址管理装置，其特征在于，所述地址管理装置具体包括：

获得模块，用于获得用户输入的原始地址数据；

10 确定模块，用于确定包括多个地址类型的结构化地址格式；

处理模块，用于将所述原始地址数据转换为符合所述结构化地址格式的结构化地址数据，所述结构化地址数据包括对应多个地址类型的地址数据。

8、如权利要求 7 所述的地址管理装置，其特征在于，所述处理模块包括：

预处理子模块，用于基于多个地址类型对原始地址数据进行预处理；

15 切分子模块，用于基于多个地址类型对预处理后的地址数据进行切分；

补全子模块，用于基于多个地址类型对切分后地址数据进行补全校验；

规范化子模块，用于对补全校验后的地址数据进行规范化处理，以得到符合所述结构化地址格式的结构化地址数据。

9、如权利要求 8 所述的地址管理装置，其特征在于，

20 所述预处理子模块，具体用于从原始地址数据中筛选出未对应所述多个地址类型的地址数据，从原始地址数据中删除当前筛选的地址数据，并将原始地址数据中存在的非规范格式的地址数据转换为规范格式的地址数据。

10、如权利要求 8 所述的地址管理装置，其特征在于，

所述切分子模块，具体用于获得多个地址类型对应的分词器词典，利用
25 多个地址类型对应的分词器词典切分出对应所述多个地址类型的地址数据。

11、如权利要求 8 所述的地址管理装置，其特征在于，

所述补全子模块，具体用于校验切分后的地址数据是否已经包含对应所述多个地址类型的地址数据；如果否，则确定切分后的地址数据中不包含的地址类型，并基于历史数据补全所述地址类型的地址数据。

5 12、如权利要求 8 所述的地址管理装置，其特征在于，

所述规范化子模块，具体用于利用拼音相似度算法对补全校验后的地址数据进行规范化处理；和/或，利用基于概率检索模型的兴趣点 POI 规范化算法对补全校验后的地址数据进行规范化处理。

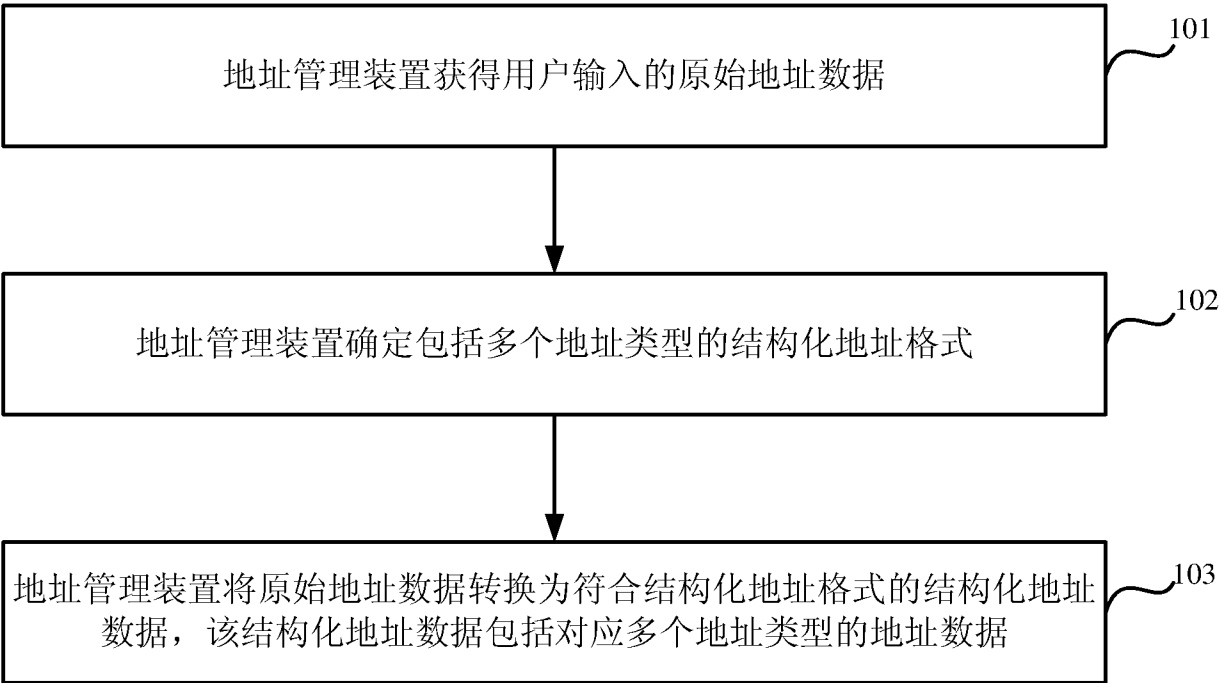


图 1

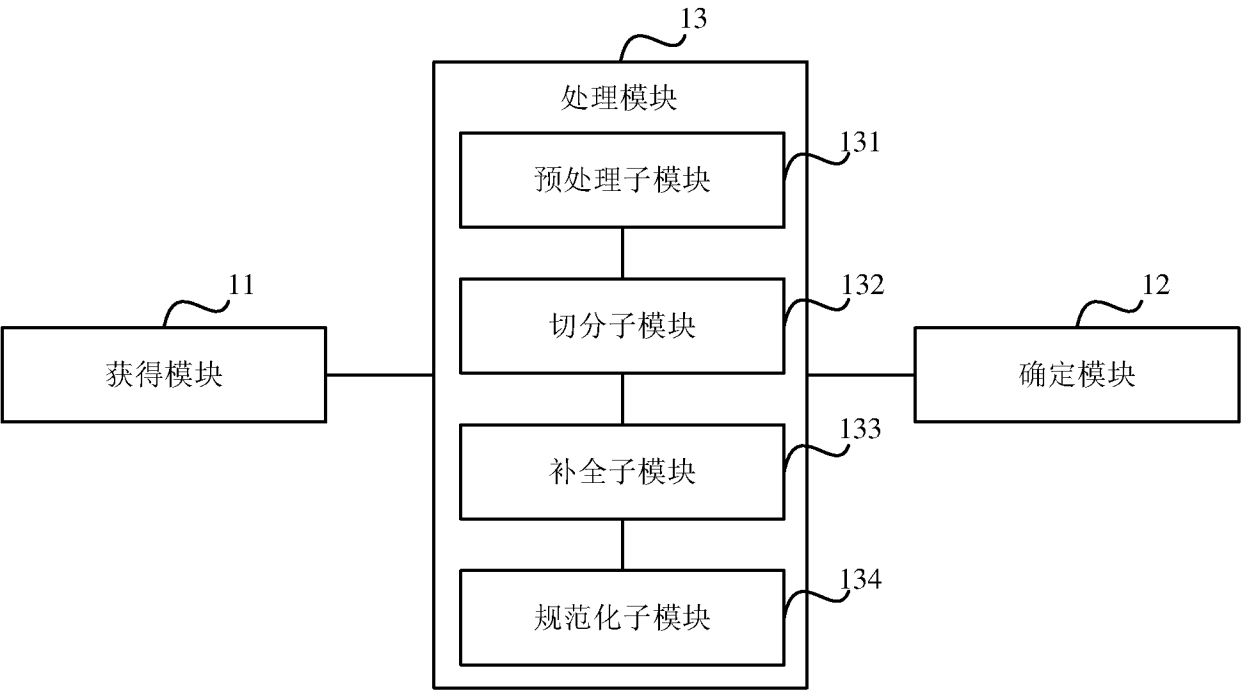


图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/077297

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06F 17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; VEN; CNKI: address, standardiz????, structur???, segment????, divid???

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103473289 A (SHENZHEN AUDAQUE DATA TECHNOLOGY LTD.) 25 December 2013 (25.12.2013) description, paragraphs [0022]-[0028], and figures 1 and 2	1-12
A	CN 101719128 A (UNIV ZHEJIANG TECHNOLOGY) 02 June 2010 (02.06.2010) the whole document	1-12
A	CN 102955832 A (SHENZHEN HUAAO DATA TECHNOLOGY CO., LTD.) 06 March 2013 (06.03.2013) the whole document	1-12
A	CN 102955832 A (SHENZHEN HUAAO DATA TECHNOLOGY CO., LTD.) 06 March 2013 (06.03.2013) the whole document	1-12
A	US 2012047179 A1 (INT BUSINESS MACHINES CORP.) 23 February 2012 (23.02.2012) the whole document	1-12

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&” document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 08 June 2016	Date of mailing of the international search report 28 June 2016
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer HAO, Xiaoli Telephone No. (86-10) 62089281

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/077297

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103473289 A	25 December 2013	WO 2015027837 A1	05 March 2015
CN 101719128 A	02 June 2010	CN 101719128 B	23 May 2012
CN 102955833 A	06 March 2013	CN 102955833 B	25 November 2015
CN 102955832 A	06 March 2013	CN 102955832 B	25 November 2015
US 2012047179 A1	23 February 2012	None	

A. 主题的分类 G06F 17/30 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类		
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F17/30 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNABS; CNTXT; VEN; CNKI: 地址, 标准, 规范, 结构化, 分词, 切分; address, standardiz????, structur????, segment????, divid???,		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 103473289 A (深圳市华傲数据技术有限公司) 2013年 12月 25日 (2013 - 12 - 25) 说明书第[0022]-[0028]段、附图1-2	1-12
A	CN 101719128 A (浙江工业大学) 2010年 6月 2日 (2010 - 06 - 02) 全文	1-12
A	CN 102955833 A (深圳市华傲数据技术有限公司) 2013年 3月 6日 (2013 - 03 - 06) 全文	1-12
A	CN 102955832 A (深圳市华傲数据技术有限公司) 2013年 3月 6日 (2013 - 03 - 06) 全文	1-12
A	US 2012047179 A1 (INT BUSINESS MACHINES CORP) 2012年 2月 23日 (2012 - 02 - 23) 全文	1-12
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。		
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件		
国际检索实际完成的日期 2016年 6月 8日		国际检索报告邮寄日期 2016年 6月 28日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		授权官员 郝晓丽 电话号码 (86-10)62089281

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/077297

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	103473289	A	2013年 12月 25日	WO	2015027837	A1	2015年 3月 5日
CN	101719128	A	2010年 6月 2日	CN	101719128	B	2012年 5月 23日
CN	102955833	A	2013年 3月 6日	CN	102955833	B	2015年 11月 25日
CN	102955832	A	2013年 3月 6日	CN	102955832	B	2015年 11月 25日
US	2012047179	A1	2012年 2月 23日	无			