

(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 7 月 7 日 (07.07.2016)



(10) 国际公布号
WO 2016/107352 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2015/095857

(22) 国际申请日: 2015 年 11 月 27 日 (27.11.2015)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201410849380.6 2014 年 12 月 29 日 (29.12.2014) CN
201410849382.5 2014 年 12 月 29 日 (29.12.2014) CN
201410849123.2 2014 年 12 月 29 日 (29.12.2014) CN

(71) 申请人: 北京奇虎科技有限公司 (BEIJING QIHOO TECHNOLOGY COMPANY LIMITED) [CN/CN]; 中国北京市西城区新街口外大街 28 号 D 座 112 室 (德胜园区), Beijing 100088 (CN)。奇智软件 (北京) 有限公司 (QIZHI SOFTWARE (BEIJING) COMPANY LIMITED) [CN/CN]; 中国北京市朝阳区

酒仙桥路 6 号院 2 号楼 B 座 2 层、3 层 301-306 室, Beijing 100015 (CN)。

(72) 发明人: 王智广 (WANG, Zhiguang); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。魏少俊 (WEI, Shaojun); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。

(74) 代理人: 北京市隆安律师事务所 (BEIJING LONGAN LAW FIRM); 中国北京市朝阳区建国门外大街 21 号北京国际俱乐部 188 室, Beijing 100020 (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,

[见续页]

(54) Title: SYSTEM AND METHOD FOR DETERMINING POI NAME AND FOR DETERMINING VALIDITY OF POI INFORMATION

(54) 发明名称: 确定 POI 名称、确定 POI 信息有效性的系统和方法

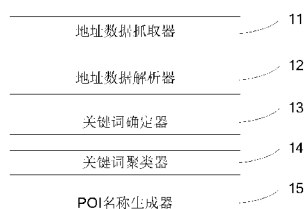


图 1 / Fig. 1

- 11 Address data capturing device
- 12 Address data parsing device
- 13 Key word determining device
- 14 Key word clustering device
- 15 POI name generator

(57) Abstract: A system and method for determining a POI name and for determining validity of POI information. The method comprises: capturing address data from network data (S11); respectively extracting a name field and address information from one or a plurality of items of captured address data (S12); on the basis of the name field, determining one or more key words (S13); clustering the key words corresponding to the same address information to generate at least one cluster (S14); and according to the clustered key word, determining a POI name corresponding to the address information (S15). The method enables a user to rapidly and accurately search for a POI name corresponding to the POI addresses at the same latitude and longitude, thereby improving the user experience.

(57) 摘要: 一种确定 POI 名称、确定 POI 信息有效性的系统和方法, 该方法包括: 从网络数据中抓取地址数据 (S11); 从抓取到的一个或多个地址数据中分别提取名称字段和地址信息 (S12); 基于所述名称字段确定一个或多个关键词 (S13); 将对应相同地址信息的所述关键词进行聚类, 生成至少一个类 (S14); 根据聚类后的关键词确定此地址信息对应的 POI 名称 (S15)。该方法使得用户能够快速、准确地搜索到同一经、纬度的 POI 地址对应的 POI 名称, 改善用户体验。

WO 2016/107352 A1



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。

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)。

- (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,

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

确定 POI 名称、确定 POI 信息有效性的系统和方法

技术领域

本发明涉及电子地图技术领域，具体而言，涉及一种基于聚类确定 POI 名称的系统和方法、一种基于聚类的 POI 名称确定系统和方法以及一种基于网络中的地址数据确定 POI 信息有效性的系统和方法。

5 背景技术

兴趣点 (Point of Interest, POI) 一般是电子地图中标注的地理信息点，通常包含 POI 标识、POI 名称、POI 类型、经度、纬度等信息。POI 可以在地图上标注出来，带有经纬度信息，可以用来查找并计算导航的地标点或者建筑物，例如商场、停车场、学校、医院、酒店、饭店、超市、公园、旅游景点等。

越来越多的用户在电子地图中查询 POI，数据库中存储的 POI 数据为 POI 查询提供数据支撑。目前，对
10 数据库中的 POI 数据进行更新主要通过进行数据实采，根据实采得到的数据对数据库中存储的 POI 数据进行更新，或是从互联网上的各个生活类信息网站上获取 POI 数据，只要获取的数据包括 POI 的名称和地址，即可将该条数据确定为一条 POI 数据。由于 POI 数据的获取及更新方式的特点，不可避免的导致互联网上存在着各种各样的 POI 数据。因此，从不同来源网站获取的 POI 数据中，有可能存在重复性数据，即多条 POI 数据实际描述的是同一 POI，其实际的 POI 经度、纬度相同，但是 POI 名称和 POI 地址的描述方式却不同。
15 重复性的 POI 数据导致用户无法快速、准确的搜索到同一 POI 地理位置（经纬度）的 POI 地址对应的 POI 名称，影响用户体验。

发明内容

鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决或者减缓上述问题的基于聚类确定 POI 名称的系统 and 相应的基于聚类确定 POI 名称的方法、一种基于聚类的 POI 名称确定系统和方法
20 以及一种基于网络中的地址数据确定 POI 信息有效性的系统和方法。

根据本发明的一个方面，提供了一种基于聚类确定 POI 名称的系统，该系统包括：

地址数据抓取器，用于从网络数据中抓取地址数据；

地址数据解析器，用于从抓取到的一个或多个地址数据中分别提取名称字段和地址信息；

关键词确定器，用于基于所述名称字段确定一个或多个关键词；

25 关键词聚类器，用于将对应相同地址信息的所述关键词进行聚类，生成至少一个类；

POI 名称生成器，用于根据聚类后的关键词确定此地址信息对应的 POI 名称。

根据本发明的另一个方面，提供了一种基于聚类确定 POI 名称的方法，包括：

从网络数据中抓取地址数据；

从抓取到的一个或多个地址数据中分别提取名称字段和地址信息；

30 基于所述名称字段确定一个或多个关键词；

将对应相同地址信息的所述关键词进行聚类，生成至少一个类；

根据聚类后的关键词确定此地址信息对应的 POI 名称。

根据本发明的又一个方面，提供了一种基于聚类的 POI 名称确定系统，该系统包括：

35 地址数据抓取器，用于基于搜索引擎从网络数据中抓取地址数据，所述地址数据包括名称字段和地址信息；

名称字段聚类器，用于将对应相同地址信息的名称字段按照关键词进行聚类；

第二频次统计器，用于统计聚类后各类别中名称字段出现的频次，作为第二频次；

POI 名称确定单元，用于根据所述第二频次确定该类别对应该地址信息的 POI 名称。

根据本发明的再一个方面，提供了一种基于聚类的 POI 名称确定方法，包括：

40 从网络数据中抓取地址数据，所述地址数据包括名称字段和地址信息；

将对应相同地址信息的名称字段按照关键词进行聚类；

统计聚类后各类别中名称字段出现的频次，作为第二频次；

根据所述第二频次确定该类别对应该地址信息的 POI 名称。

45 根据本发明的再一个方面，提供了一种基于网络中的地址数据确定 POI 信息有效性的系统，该系统包括：
POI 信息获取单元，用于基于搜索引擎利用网络中的地址数据获取对应相同 POI 名称的多个相关 POI 信息；

统计单元，用于统计所述 POI 信息在所述网络中的地址数据中的出现次数；

POI 信息确定单元，用于根据所述 POI 信息在所述网络中的地址数据中的出现次数确定对应所述相同

POI 名称的有效 POI 信息。

根据本发明的再一个方面，提供了一种基于网络中的地址数据确定 POI 信息有效性的方法，包括：

利用网络中的地址数据获取对应相同 POI 名称的多个相关 POI 信息；

统计所述 POI 信息在所述网络中的地址数据中的出现次数；

5 根据所述 POI 信息在所述网络中的地址数据中的出现次数确定对应所述相同 POI 名称的有效 POI 信息。

根据本发明的再一个方面，提出了一种计算机程序，包括计算机可读代码，当所述计算机可读代码在计算设备上运行时，导致所述计算设备执行上文所述的基于聚类确定 POI 名称的方法，或者导致所述计算设备执行上文所述的基于聚类的 POI 名称确定方法，或者导致所述计算设备执行上文所述的基于网络中的地址数据确定 POI 信息有效性的方法。

10 根据本发明的再一个方面，提出了一种计算机可读介质，其中存储了上述的计算机程序。

本发明的有益效果为：

本发明对从网络数据中抓取地址数据进行名称字段和地址信息的提取，基于名称字段确定一个或多个关键词，并将对应相同地址信息的关键词进行聚类，基于聚类后的关键词确定地址信息对应的 POI 名称，从而使得用户能够快速、准确地搜索到同一经、纬度的 POI 地址对应的 POI 名称，改善用户体验。

15 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

20 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

图 1 示意性示出了本发明一个实施例的基于聚类确定 POI 名称的系统的框图；

图 2 示意性示出了本发明另一个实施例的基于聚类确定 POI 名称的系统中的关键词确定器的框图；

图 3 示意性示出了本发明另一个实施例的基于聚类确定 POI 名称的系统中的 POI 名称生成器的框图；

25 图 4 示意性示出了本发明另一个实施例的基于聚类确定 POI 名称的系统中的 POI 名称生成器的框图；

图 5 示意性示出了本发明一个实施例的基于聚类确定 POI 名称的方法的流程图；

图 6 示意性示出了本发明另一个实施例的基于聚类确定 POI 名称的方法的步骤 S13 的细分流程图；

图 7 示意性示出了本发明另一个实施例的基于聚类确定 POI 名称的方法的步骤 S15 的细分流程图；以

及

30 图 8 示意性示出了本发明另一个实施例的基于聚类确定 POI 名称的方法的步骤 S15 的细分流程图；

图 9 示意性示出了本发明一个实施例的基于聚类的 POI 名称确定系统的框图；

图 10 示意性示出了本发明另一个实施例的基于聚类的 POI 名称确定系统中的名称字段聚类器的框图；

图 11 示意性示出了本发明另一个实施例的基于聚类的 POI 名称确定系统中的第二频次统计器的框图；

图 12 示意性示出了本发明一个实施例的基于聚类的 POI 名称确定方法的流程图；

35 图 13 示意性示出了本发明另一个实施例的基于聚类的 POI 名称确定方法的步骤 S122 的细分流程图；以及

图 14 示意性示出了本发明另一个实施例的基于聚类的 POI 名称确定方法的步骤 S123 的细分流程图。

图 15 示意性示出了本发明一个实施例的基于网络中的地址数据确定 POI 信息有效性的系统的框图；

图 16 示意性示出了本发明另一个实施例的基于网络中的地址数据确定 POI 信息有效性的系统中的统计单元的框图；

40 图 17 示意性示出了本发明另一个实施例的基于网络中的地址数据确定 POI 信息有效性的系统中的 POI 信息确定单元的框图；

图 18 示意性示出了本发明一个实施例的基于网络中的地址数据确定 POI 信息有效性的方法的流程图；

图 19 示意性示出了本发明另一个实施例的基于网络中的地址数据确定 POI 信息有效性的方法的步骤 S812 的细分流程图；以及

45 图 20 示意性示出了本发明另一个实施例的基于网络中的地址数据确定 POI 信息有效性的方法的步骤 S813 的细分流程图。

图 21 示意性地示出了用于执行根据本发明的方法的计算设备的框图；以及

图 22 示意性地示出了用于保持或者携带实现根据本发明的方法的程序代码的存储单元。

具体实施例

下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能解释为对本发明的限制。

5 本技术领域技术人员可以理解，除非特意声明，这里使用的单数形式“一”、“一个”、“所述”和“该”也可包括复数形式。应该进一步理解的是，本发明的说明书中使用的措辞“包括”是指存在所述特征、整数、步骤、操作、元件和/或组件，但是并不排除存在或添加一个或多个其他特征、整数、步骤、操作、元件、组件和/或它们的组。

10 本技术领域技术人员可以理解，除非另外定义，这里使用的所有术语（包括技术术语和科学术语），具有与本发明所属领域中的普通技术人员的一般理解相同的意义。还应该理解的是，诸如通用字典中定义的那些术语，应该被理解为具有与现有技术的上下文中的意义一致的意义，并且除非被特定定义，否则不会用理想化或过于正式的含义来解释。

图 1 示出了本发明一个实施例的基于聚类确定 POI 名称的系统的框图。

参照图 1，本发明实施例的基于聚类确定 POI 名称的系统，包括：

15 地址数据抓取器 11，用于从网络数据中抓取地址数据；

地址数据解析器 12，用于从抓取到的一个或多个地址数据中分别提取名称字段和地址信息；

关键词确定器 13，用于基于所述名称字段确定一个或多个关键词；

关键词聚类器 14，用于将对应相同地址信息的所述关键词进行聚类，生成至少一个类；

POI 名称生成器 15，用于根据聚类后的关键词确定此地址信息对应的 POI 名称。

20 本发明实施例，基于搜索引擎利用网络中的地址数据，所述地址数据包括名称字段、地址信息以及多个相关 POI 信息；本发明实施例中，所述多个相关 POI 信息为对应 POI 至少一个预设属性的信息。进一步地，所述预设属性为经纬度、地址、建筑物名称或所囊括单位名称。

25 本发明实施例，基于搜索引擎从网络数据中抓取地址数据，所述地址数据包括名称字段和地址信息，基于搜索引擎从互联网上挖掘的地图地址数据，比如 name: 某某地产集团**分公司公司；address: **市**区 8*财富中心 A 座写字楼 14 楼，其中“某某地产集团**分公司公司”为 POI 的名称，“**市**区 8*财富中心 A 座写字楼 14 楼”为此 POI 的地址，通过对地址的经纬度解析可以获得此地址所在的经纬度信息，比如地址“**市**区 8*财富中心 A 座写字楼 14 楼”经纬度解析得到的经纬度为：东经：102.733445 北纬：25.08108。另外，需要统计 POI 信息在互联网上出现的次数以及记录来源。

所以，最终从互联网上挖掘的地址数据对应的不同信息源的 POI 信息的格式如表 1 所示，具体如下：

表 1 不同信息源的 POI 信息的格式表

经度	纬度	名字	地址	次数
东经:99.1794081	北纬:25.1225831	保山博鑫源汽车贸易有限公司	云南保山隆阳区和平路 79 号啤酒厂东侧	1
东经:99.1794081	北纬:25.1225831	云南省澜沧江啤酒集团保山有限公司	中国云南保山市隆阳区云南省保山市隆阳区和平路 79 号	8
东经:99.1794081	北纬:25.1225831	云南省澜沧江啤酒集团保山有限公司	云南省保山市云南省保山市隆阳区和平路 79 号	7
东经:99.1794081	北纬:25.1225831	云南省澜沧江啤酒集团保山有限公司	云南省保山市隆阳区和平路 79 号	3

东经:99.1794081	北纬:25.1225831	云南省澜沧江啤酒集团保山有限公司(地图标注)	云南-保山市隆阳区云南省保山市隆阳区和平路 79 号	1
东经:99.1794081	北纬:25.1225831	保山明志汽车销售有限公司	保山隆阳区和平路 79 号澜沧江啤酒厂旁	1
东经:99.1794081	北纬:25.1225831	保山明志汽车销售服务有限公司	云南保山隆阳区和平路 79 号	3
东经:99.1794081	北纬:25.1225831	保山明志汽车销售服务有限公司	隆阳区和平路 79 号	1
东经:99.1794081	北纬:25.1225831	保山长城汽车 4S 店	保山隆阳区和平路 79 号	1
东经:99.1794081	北纬:25.1225831	保山融易通汽车销售有限公司(雪佛兰 4S 店)	保山隆阳区和平路 79 号	1

由表 1 可见，在同一地理位置（经纬度相同）从不同来源网站获取的 POI 数据中，有可能存在重复性数据，即同一个地址（经纬度）可能存在多个 POI 名字，如表 1 中同一个经纬度存在多个公司，其实际的 POI 经度、纬度相同，但是 POI 名称和 POI 地址的描述方式却不同；还可以看出，同一个 poi 名字可能多种不同的说法，比如“保山明志汽车销售有限公司”和“保山明志汽车销售服务有限公司”，重复性的 POI 数据导致用户无法快速、准确的搜索到同一 POI 地理位置（经纬度）的 POI 地址对应的 POI 名称。

对此，本发明实施例，从网络数据中抓取地址数据，从抓取到的一个或多个地址数据中分别提取名称字段和地址信息，基于所述名称字段确定一个或多个关键词；将对应相同地址信息的所述关键词进行聚类，生成至少一个类，根据聚类后的关键词确定此地址信息对应的 POI 名称，进而得到最佳的 poi 名字。

为了进一步体现发明的优越性，如下进一步揭示本发明基于聚类确定 POI 名称的系统中的关键词确定器 13 的在另一实施例中的内部结构，来体现依据关键词确定器 13 实现的另一实施例的细节。参照图 2，关键词确定器 13 进一步包括切词单元 131 以及关键词获取单元 132：

- 所述的切词单元 131，用于对所述名称字段中的名称进行切词处理生成分词；
- 所述的关键词获取单元 132，用于根据所述分词获取所述地址数据的关键词。
- 其中，关键词获取单元进一步包括：
- 15 第一频次统计模块，用于统计对应相同地址信息的每个分词出现的频次，作为第一频次；
- 关键词生成模块，用于根据所述第一频次生成所述地址数据的关键词。
- 其中，关键词生成模块选择频次最小并且是非地名的分词作为所述地址数据的关键词。
- 本发明实施例中，对所挖掘的地址数据中 POI 信息的名称切词，并且统计切词后每个词出现的次数，同一个 POI 名称中出现频次最少即包含的信息量最大，并且是非地名的那个词记为该 POI 名称的关键词，比如
- 20 表 1 中出现的地址数据对应的相关 POI 信息中 POI 名称切词后数据如表 2 所示（词频是根据约 9000 万的 poi 名字统计的），表 2 中第二列为获取到的关键词，具体如下：

表2 POI名称的切词后的数据表

名字	关键词/频次	切词数据/词频
保山博鑫源汽车贸易有限公司	博鑫源/74	保山/5358 博鑫源/74 汽车-贸易/31123 有限公司/17269098
云南省澜沧江啤酒集团保山有限公司	啤酒/20946	云南-省/65291 澜沧-江/457 啤酒/20946 集团/645368 保山/5358 有限公司/17269098
云南省澜沧江啤酒集团保	啤酒/20946	云南-省/65291 澜沧-江/457 啤酒

山有限公司		/20946 集团/645368 保山/5358 有限-公司/17269098
云南省澜沧江啤酒集团保山有限公司	啤酒/20946	云南-省/65291 澜沧-江/457 啤酒 /20946 集团/645368 保山/5358 有限-公司/17269098
云南省澜沧江啤酒集团保山有限公司(地图标注)	啤酒/20946	云南-省/65291 澜沧-江/457 啤酒 /20946 集团/645368 保山/5358 有限-公司/17269098 (/7519121 地图/1665298 标注 /1663423)/5355654
保山明志汽车销售有限公司	明志/1139	保山/5358 明志/1139 汽车 /702730 销售/174240 有限-公司 /17269098
保山明志汽车销售服务有限公司	明志/1139	保山/5358 明志/1139 汽车 /702730 销售/174240 服务 /885349 有限-公司/17269098
保山明志汽车销售服务有限公司	明志/1139	保山/5358 明志/1139 汽车 /702730 销售/174240 服务 /885349 有限-公司/17269098
保山长城汽车 4S 店	汽车-4S-店 /220	保山/5358 长城/50304 汽车 -4S-店/220
保山融易通汽车销售有限公司(雪佛兰 4S 店)	融易通/45	保山/5358 融易通/45 汽车 /702730 销售/174240 有限-公司 /17269098 (/7519121 雪佛兰 /1726 4S-店/1702)/5355654

根据关键词聚类：同一个关键词对应的 POI 名称记为同一类，上述几个 POI 名称可以归为 5 个类，也就是说在此 POI 地址上存在 5 个不同的 poi 名字。

为了进一步体现发明的优越性，如下进一步揭示本发明基于聚类确定 POI 名称的系统中的 POI 名称生成器 15 的在另一实施例中的内部结构，来体现依据 POI 名称生成器 15 实现的另一实施例的细节。参照图 3，

5 POI 名称生成器 15 进一步包括频率统计单元 151、类标识名称确定单元 152 以及 POI 名称确定单元 153：

所述的频率统计单元 151，用于计算各个类中名称字段的出现频率；

所述的类标识名称确定单元 152，用于将所述各个类中出现频率最高的名称字段作为类标识名称；

所述的 POI 名称确定单元 153，用于将每个类标识名称均作为 POI 名称。

本实施例中，将每个类标识名称均作为 POI 名称，进一步为：根据关键词聚类：同一个关键词对应的

10 POI 名称记为同一类，上述几个 POI 名称可以归为 5 个类，也就是说在此 POI 地址上存在 5 个不同的 poi 名字，分别为：

A：保山博鑫源汽车贸易有限公司；

B：云南省澜沧江啤酒集团保山有限公司云南省澜沧江啤酒集团保山有限公司(地图标注)；

C：保山明志汽车销售有限公司保山明志汽车销售服务有限公司

15 D：保山长城汽车 4S 店；

E：保山融易通汽车销售有限公司(雪佛兰 4S 店)。

为了进一步体现发明的优越性，如下进一步揭示本发明基于聚类确定 POI 名称的系统中的 POI 名称生成器 15 的在另一实施例中的内部结构，来体现依据 POI 名称生成器 15 实现的另一实施例的细节。参照图 4，

20 POI 名称生成器 15 进一步包括频率统计单元 151'、类标识名称确定单元 152' 以及 POI 名称确定单元 153'：

频率统计单元 151'，用于计算各个类中名称字段的出现频率；

类标识名称确定单元 152'，用于将所述各个类中出现频率最高的名称字段作为类标识名称；

POI 名称确定单元 153'，用于选择出现频率最高的类标识名称作为 POI 名称。

本实施例中，在同一类的 POI 名称中，选取最佳的 POI 名称是根据互联上的“投票”来解决，所谓“投票”主要是根据此 POI 名称在互联网上出现的频次以及来源的可信度，互联网上出现的频次最高、来源最可信的那个名字为要选取的最佳名字。比如：

A 类中只有一个名字，最佳的也是这一个。

5 B 类中有两个名字，其中“云南省澜沧江啤酒集团保山有限公司”出现的频率最高，作为最佳名字。

C 类中有两个名字，其中“保山明志汽车销售服务有限公司”出现的频率最高，作为最佳名字。

D 类和 E 类中同样是只有一个名字，类似 A。

本发明实施例中，所述可靠来源为具有预定可信度的来源。其中，所述来源为网站或者网页。

10 其中，可靠来源的网站或者网页包括但不限于，如新浪、凤凰网等大型网站、通过官方认证的网站、访问频次比较高、数据流量大的网站以及不携带恶意链接、病毒链接且客户满意度交高的网站等。

本发明实施例中，可靠来源的网站或者网页的可信度是可量化的，可根据用户的访问次数以及客户评价等对各个网站或网页的可信度进行量化。而且各个网站或网页的可信度是动态变化的，若当前网站出现病毒、欺诈广告或被其他恶意欺诈网站所利用，则其可信度会随之降低，本发明通过网站可信度的量化和动态调整，进一步保证获取的 POI 信息的可靠、有效。

15 本发明实施例提供的基于聚类确定 POI 名称的系统，根据切词后词频次的多少来挖掘 poi 名字的关键词，并且以此关键词来聚类，把不同说法的同一个 poi 名字聚为一类，解决同一个经纬度对应多个 poi 名字的问题，利用互联网“投票”机制来选取最佳的 poi 名字。

图 5 示出了本发明一个实施例的基于聚类确定 POI 名称的方法的流程图。

参照图 5，本发明实施例的基于聚类确定 POI 名称的方法包括以下步骤：

20 S11、从网络数据中抓取地址数据；

S12、从抓取到的一个或多个地址数据中分别提取名称字段和地址信息；

S13、基于所述名称字段确定一个或多个关键词；

S14、将对应相同地址信息的所述关键词进行聚类，生成至少一个类；

S15、根据聚类后的关键词确定此地址信息对应的 POI 名称。

25 本发明实施例，基于搜索引擎从网络数据中抓取地址数据，所述地址数据包括名称字段和地址信息，基于搜索引擎从互联网上挖掘的地图地址数据，比如 name: 恒大地产集团昆明公司；address: 昆明市盘龙区北辰财富中心 A 座写字楼 14 楼，其中“恒大地产集团昆明公司”为 POI 的名称，“昆明市盘龙区北辰财富中心 A 座写字楼 14 楼”为此 POI 的地址，通过对地址的经纬度解析可以获得此地址所在的经纬度信息，比如地址“昆明市盘龙区北辰财富中心 A 座写字楼 14 楼”经纬度解析得到的经纬度为：东经：102.733445 北纬：25.08108。另外，需要统计 POI 信息在互联网上出现的次数以及记录来源。在同一地理位置（经纬度相同）从不同来源网站获取的 POI 数据中，有可能存在重复性数据，即同一个地址（经纬度）可能存在多个 POI 名字，如同一个经纬度存在多个公司，其实际的 POI 经度、纬度相同，但是 POI 名称和 POI 地址的描述方式却不同；还可以看出，同一个 poi 名字可能多种不同的说法，比如“保山明志汽车销售有限公司”和“保山明志汽车销售服务有限公司”，重复性的 POI 数据导致用户无法快速、准确的搜索到同一 POI 地理位置（经纬度）的 POI 30 地址对应的 POI 名称。

对此，本发明实施例，从网络数据中抓取地址数据，从抓取到的一个或多个地址数据中分别提取名称字段和地址信息，基于所述名称字段确定一个或多个关键词；将对应相同地址信息的所述关键词进行聚类，生成至少一个类，根据聚类后的关键词确定此地址信息对应的 POI 名称，进而得到最佳的 poi 名字。

40 为了进一步体现发明的优越性，如下进一步揭示本发明基于聚类确定 POI 名称的方法中步骤 S13 的细分步骤，来体现依据本步骤实现的另一实施例。参照图 6，本步骤的细分步骤包括：

S131、对所述名称字段中的名称进行切词处理生成分词；

S132、根据所述分词获取所述地址数据的关键词。

其中，步骤 S132：根据所述分词获取所述地址数据的关键词，进一步包括：

统计对应相同地址信息的每个分词出现的频次作为第一频次；

45 根据所述第一频次生成所述地址数据的关键词。

其中步骤：根据所述第一频次生成所述地址数据的关键词具体为：

选择频次最小并且是非地名的分词作为所述地址数据的关键词。

为了进一步体现发明的优越性，如下进一步揭示本发明基于聚类确定 POI 名称的方法中步骤 S15 的细

分步骤, 来体现依据本步骤实现的另一实施例。参照图 7, 本步骤的细分步骤包括:

S151、计算各个类中名称字段的出现频率;

S152、将所述各个类中出现频率最高的名称字段作为类标识名称;

S153、将每个类标识名称均作为 POI 名称。

5 本实施例中, 将每类标识名称均作为对应该地址信息的 POI 名称, 根据关键词聚类: 同一个关键词对应的 POI 名称记为同一类, 上述几个 POI 名称可以归为 5 个类, 也就是说在此 POI 地址上存在 5 个不同的 poi 名字, 分别为:

A: 保山博鑫源汽车贸易有限公司;

B: 云南省澜沧江啤酒集团保山有限公司云南省澜沧江啤酒集团保山有限公司(地图标注);

10 C: 保山明志汽车销售有限公司保山明志汽车销售服务有限公司

D: 保山长城汽车 4S 店;

E: 保山融易通汽车销售有限公司(雪佛兰 4S 店)。

为了进一步体现发明的优越性, 如下进一步揭示本发明基于聚类确定 POI 名称的方法中步骤 S15 的细分步骤, 来体现依据本步骤实现的另一实施例。参照图 8, 本步骤的细分步骤包括:

15 S151'、计算各个类中名称字段的出现频率;

S152'、将所述各个类中出现频率最高的名称字段作为类标识名称;

S153'、选择出现频率最高的类标识名称作为 POI 名称。

本实施例中, 在同一类的 POI 名称中, 选取最佳的 POI 名称是根据互联上的“投票”来解决, 所谓“投票”主要是根据此 POI 名称在互联网上出现的频次以及来源的可信度, 互联网上出现的频次最高、来源最可信的那个名字为要选取的最佳名字。比如:

20 A 类中只有一个名字, 最佳的也是这一个。

B 类中有两个名字, 其中“云南省澜沧江啤酒集团保山有限公司”出现的频率最高, 作为最佳名字。

C 类中有两个名字, 其中“保山明志汽车销售服务有限公司”出现的频率最高, 作为最佳名字。

D 类和 E 类中同样是只有一个名字, 类似 A。

25 本发明实施例中, 所述可靠来源为具有预定可信度的来源。其中, 所述来源为网站或者网页。

其中, 可靠来源的网站或者网页包括但不限于, 如新浪、凤凰网等大型网站、通过官方认证的网站、访问频次比较高、数据流量大的网站以及不携带恶意链接、病毒链接且客户满意度交高的网站等。

本发明实施例中, 可靠来源的网站或者网页的可信度是可量化的, 可根据用户的访问次数以及客户评价等对各个网站或网页的可信度进行量化。而且各个网站或网页的可信度是动态变化的, 若当前网站出现病毒、30 欺诈广告或被其他恶意欺诈网站所利用, 则其可信度会随之降低, 本发明通过网站可信度的量化和动态调整, 进一步保证获取的 POI 信息的可靠、有效。

本发明实施例提供的基于聚类确定 POI 名称的方法, 根据切词后词频次的多少来挖掘 poi 名字的关键词, 并且以此关键词来聚类, 把不同说法的同一个 poi 名字聚为一类, 解决同一个经纬度对应多个 poi 名字的问题, 利用互联网“投票”机制来选取最佳的 poi 名字。

35 综上所述, 本发明的上述实施例中通过对从网络数据中抓取地址数据进行名称字段和地址信息的提取, 基于名称字段确定关键词, 并将对应相同地址信息的关键词进行聚类, 基于聚类后的关键词确定地址信息对应的 POI 名称, 从而使得用户能够快速、准确地搜索到同一经、纬度的 POI 地址对应的 POI 名称, 改善用户体验。

40 图 9 示出了本发明一个实施例的基于聚类的 POI 名称确定系统的框图。

参照图 9, 本发明实施例的基于聚类的 POI 名称确定系统, 包括:

地址数据抓取器 91, 用于基于搜索引擎从网络数据中抓取地址数据, 所述地址数据包括名称字段和地址信息;

名称字段聚类器 92, 用于将对应相同地址信息的名称字段按照关键词进行聚类;

45 第二频次统计器 93, 用于统计聚类后各类别中名称字段出现的频次, 作为第二频次;

POI 名称确定单元 94, 用于根据所述第二频次确定该类别对应该地址信息的 POI 名称。

本发明实施例, 基于搜索引擎利用网络中的地址数据, 所述地址数据包括名称字段、地址信息以及多个相关 POI 信息; 本发明实施例中, 所述多个相关 POI 信息为对应 POI 至少一个预设属性的信息。进一步地, 所

述预设属性为经纬度、地址、建筑物名称或所囊括单位名称。

由前述的表 1 可见,在同一地理位置(经纬度相同)从不同来源网站获取的 POI 数据中,有可能存在重复性数据;还可以看出,同一个 poi 名字可能多种不同的说法。

对此,本发明实施例,基于搜索引擎从网络数据中抓取地址数据,所述地址数据包括名称字段和地址信息,5 将对应相同地址信息的名称字段按照关键词进行聚类,统计聚类后各类别中名称字段出现的频次,作为第二频次,根据所述第二频次确定该类别对应该地址信息的 POI 名称,进而得到最佳的 poi 名字。

为了进一步体现发明的优越性,如下进一步揭示本发明基于聚类的 POI 名称确定系统中的名称字段聚类器 92 的在另一实施例中的内部结构,来体现依据名称字段聚类器 92 实现的另一实施例的细节。参照图 10,名称字段聚类器 92 进一步包括关键词确定单元 921、关键词聚类单元 922 以及名称字段聚类确定单元 923:

10 所述关键词确定单元 921,用于基于所述名称字段确定一个或多个关键词;
所述关键词聚类单元 922,用于将对应相同地址信息的所述关键词进行聚类;
所述名称字段聚类确定单元 923,用于根据聚类后的关键词确定聚类后的名称字段。

更进一步地,所述关键词确定单元 921 进一步包括切词模块和关键词获取模块:所述切词模块,用于对所15 述名称字段中的名称进行切词处理生成分词;所述关键词获取模块,用于根据所述分词获取所述名称字段的关键词。

更进一步地,所述关键词获取模块进一步包括第一频次统计子模块和关键词生成子模块:所述第一频次统计子模块,用于统计对应相同地址信息的每个分词出现的频次,作为第一频次;所述关键词生成子模块,用于根据所述第一频次生成所述名称字段的关键词。

其中,所述关键词生成子模块选择所述第一频次最小并且是非地名的分词作为所述名称字段的关键词。

20 本发明实施例中,对所挖掘的地址数据中 POI 信息的名称切词,并且统计切词后每个词出现的次数,同一个 POI 名称中出现频次最少即包含的信息量最大,并且是非地名的那个词记为该 POI 名称的关键词,比如前述表 1 中出现的地址数据对应的相关 POI 信息中 POI 名称切词后数据如前述表 2 所示(词频是根据约 9000 万的 poi 名字统计的)。为了进一步体现发明的优越性,如下进一步揭示本发明基于聚类的 POI 名称确定系统中的第二频次统计器 93 的在另一实施例中的内部结构,来体现依据第二频次统计器 93 实现的另一实施例的细节。参照图 11,第二频次统计器 93 进一步包括名称字段来源获取单元 931、来源可靠性判断单元 932 以及第25 二频次统计单元 933:

所述名称字段来源获取单元 931,用于获取所述名称字段的来源;
所述来源可靠性判断单元 932,用于判断所述来源是否属于可靠来源;
所述第二频次统计单元 933,用于在判断为是的情况下,统计所述名称字段出现的频次,作为第二频次,30 否则不统计。

本发明实施例中,所述可靠来源为具有预定可信度的来源。其中,所述来源为网站或者网页。

其中,可靠来源的网站或者网页包括但不限于,如新浪、凤凰网等大型网站、通过官方认证的网站、访问频次比较高、数据流量大的网站以及不携带恶意链接、病毒链接且客户满意度交高的网站等。

35 本发明实施例中,可靠来源的网站或者网页的可信度是可量化的,可根据用户的访问次数以及客户评价等对各个网站或网页的可信度进行量化。而且各个网站或网页的可信度是动态变化的,若当前网站出现病毒、欺诈广告或被其他恶意欺诈网站所利用,则其可信度会随之降低,本发明通过网站可信度的量化和动态调整,进一步保证获取的 POI 信息的可靠、有效。

为了进一步体现发明的优越性,如下进一步揭示本发明基于聚类的 POI 名称确定系统中的 POI 名称确定单元 94 的在另一实施例中的内部结构,来体现依据 POI 名称确定单元 94 实现的另一实施例的细节。POI 名称40 确定单元 94 进一步包括第一类标识名称确定模块以及第一 POI 名称确定模块:

所述第一类标识名称确定模块,用于将所述各个类中所述第二频次最高的名称字段作为类标识名称;
所述第一 POI 名称确定模块,用于将每类标识名称均作为对应该地址信息的 POI 名称。

本实施例中,将每类标识名称均作为对应该地址信息的 POI 名称,根据关键词聚类:同一个关键词对应的 POI 名称记为同一类,参见表 1 和表 2,上述几个 POI 名称可以归为 5 个类,也就是说在此 POI 地址上存45 在 5 个不同的 poi 名字,分别为:

A: 保山博鑫源汽车贸易有限公司;
B: 云南省澜沧江啤酒集团保山有限公司 云南省澜沧江啤酒集团保山有限公司(地图标注);
C: 保山明志汽车销售有限公司 保山明志汽车销售服务有限公司

D: 保山长城汽车 4S 店;

E: 保山融易通汽车销售有限公司(雪佛兰 4S 店)。

为了进一步体现发明的优越性, 如下进一步揭示本发明基于聚类的 POI 名称确定系统中的 POI 名称确定单元 94 的在另一实施例中的内部结构, 来体现依据 POI 名称确定单元 94 实现的另一实施例的细节。POI 名称确定单元 94 进一步包括第二类标识名称确定模块以及第二 POI 名称确定模块:

所述第二类标识名称确定模块, 用于将所述各个类中第二频次最高的名称字段作为类标识名称;

所述第二 POI 名称确定模块, 用于将网络上出现次数最多的类标识名称作为对应该地址信息的 POI 名称。

本实施例中, 在同一类的 POI 名称中, 选取最佳的 POI 名称是根据互联上的“投票”来解决, 所谓“投票”主要是根据此 POI 名称在互联网上出现的频次以及来源的可信度, 互联网上出现的频次最高、来源最可信的那个名字为要选取的最佳名字。比如:

A 类中只有一个名字, 最佳的也是这一个。

B 类中有两个名字, 其中“云南省澜沧江啤酒集团保山有限公司”出现的频率最高, 作为最佳名字。

C 类中有两个名字, 其中“保山明志汽车销售服务有限公司”出现的频率最高, 作为最佳名字。

D 类和 E 类中同样是只有一个名字, 类似 A。

本发明实施例提供的基于聚类的 POI 名称确定系统, 根据切词后词频次的多少来挖掘 POI 名字的关键词, 并且以此关键词来聚类, 把不同说法的同一个 POI 名字聚为一类, 解决同一个经纬度对应多个 POI 名字的问题, 利用互联网“投票”机制来选取最佳的 POI 名字。

图 12 示出了本发明一个实施例的基于聚类的 POI 名称确定方法的流程图。

参照图 12, 本发明实施例的基于聚类的 POI 名称确定方法包括以下步骤:

S121、从网络数据中抓取地址数据, 所述地址数据包括名称字段和地址信息;

S122、将对应相同地址信息的名称字段按照关键词进行聚类;

S123、统计聚类后各类别中名称字段出现的频次, 作为第二频次;

S124、根据所述第二频次确定该类别对应该地址信息的 POI 名称。

基于搜索引擎利用网络中的地址数据, 所述地址数据包括名称字段、地址信息以及多个相关 POI 信息; 本发明实施例中, 所述多个相关 POI 信息为对应 POI 至少一个预设属性的信息。进一步地, 所述预设属性为经纬度、地址、建筑物名称或所囊括单位名称。

本发明实施例, 基于搜索引擎从网络数据中抓取地址数据, 所述地址数据包括名称字段和地址信息, 基于搜索引擎从互联网上挖掘的地图地址数据, 比如 name: 恒大地产集团昆明公司; address: 昆明市盘龙区北辰财富中心 A 座写字楼 14 楼, 其中“恒大地产集团昆明公司”为 POI 的名称, “昆明市盘龙区北辰财富中心 A 座写字楼 14 楼”为此 POI 的地址, 通过对地址的经纬度解析可以获得此地址所在的经纬度信息, 比如地址“昆明市盘龙区北辰财富中心 A 座写字楼 14 楼”经纬度解析得到的经纬度为: 东经: 102.733445 北纬: 25.08108。另外, 需要统计 POI 信息在互联网上出现的次数以及记录来源。

但是, 在同一地理位置(经纬度相同)从不同来源网站获取的 POI 数据中, 有可能存在重复性数据, 即同一个地址(经纬度)可能存在多个 POI 名字, 如同一个经纬度存在多个公司, 其实际的 POI 经度、纬度相同, 但是 POI 名称和 POI 地址的描述方式却不同; 还可以看出, 同一个 poi 名字可能多种不同的说法, 比如“保山明志汽车销售有限公司”和“保山明志汽车销售服务有限公司”, 重复性的 POI 数据导致用户无法快速、准确的搜索到同一 POI 地理位置(经纬度)的 POI 地址对应的 POI 名称。

为了进一步体现发明的优越性, 如下进一步揭示本发明基于聚类的 POI 名称确定方法中步骤 S122 的细分步骤, 来体现依据本步骤实现的另一实施例。参照图 13, 本步骤的细分步骤包括:

S1221、基于所述名称字段确定一个或多个关键词;

S1222、将对应相同地址信息的所述关键词进行聚类;

S1223、根据聚类后的关键词确定聚类后的名称字段。

更进一步地, 所述步骤 S1221: 基于所述名称字段确定一个或多个关键词, 进一步包括: 对所述名称字段进行切词处理生成分词; 根据分词获取所述名称字段的关键词。

更进一步地, 所述步骤: 根据分词获取所述名称字段的关键词, 进一步包括: 统计对应相同地址信息的每个分词出现的频次, 作为第一频次; 根据所述第一频次确定所述名称字段的关键词。

更进一步地, 所述步骤根据所述第一频次确定所述名称字段的关键词具体为: 选择第一频次最小并且是非地名的分词作为所述名称的关键词。

本发明实施例中，对所挖掘的地址数据中POI信息的名称切词，并且统计切词后每个词出现的次数，同一个POI名称中出现频次最少即包含的信息量最大，并且是非地名的那个词记为该POI名称的关键词，根据关键词聚类：同一个关键词对应的POI名称记为同一类。

5 为了进一步体现发明的优越性，如下进一步揭示本发明基于聚类的POI名称确定方法中步骤S123的细分步骤，来体现依据本步骤实现的另一实施例。参照图14，本步骤的细分步骤包括：

S1231、获取所述名称字段的来源；

S1232、判断所述来源是否属于可靠来源，如果是，则执行S1233；

S1233、统计所述名称字段出现的频次，作为第二频次。

本发明实施例中，所述可靠来源为具有预定可信度的来源。其中，所述来源为网站或者网页。

10 其中，可靠来源的网站或者网页包括但不限于，如新浪、凤凰网等大型网站、通过官方认证的网站、访问频次比较高、数据流量大的网站以及不携带恶意链接、病毒链接且客户满意度交高的网站等。

15 本发明实施例中，可靠来源的网站或者网页的可信度是可量化的，可根据用户的访问次数以及客户评价等对各个网站或网页的可信度进行量化。而且各个网站或网页的可信度是动态变化的，若当前网站出现病毒、欺诈广告或被其他恶意欺诈网站所利用，则其可信度会随之降低，本发明通过网站可信度的量化和动态调整，进一步保证获取的POI信息的可靠、有效。

为了进一步体现发明的优越性，如下进一步揭示本发明基于聚类的POI名称确定方法中步骤S124的细分步骤，来体现依据本步骤实现的另一实施例。本步骤的细分步骤包括：

将所述各个类中所述第二频次最高的名称字段作为类标识名称；将每类标识名称均作为对应该地址信息的POI名称。

20 为了进一步体现发明的优越性，如下进一步揭示本发明基于聚类的POI名称确定方法中步骤S124的细分步骤，来体现依据本步骤实现的另一实施例。本步骤的细分步骤包括：

将所述各个类中所述第二频次最高的名称字段作为类标识名称；将网络上出现次数最多的类标识名称作为对应该地址信息的POI名称。

25 本发明实施例提供的基于聚类的POI名称确定方法，根据切词后词频次的多少来挖掘POI名字的关键词，并且以此关键词来聚类，把不同说法的同一个POI名字聚为一类，解决同一个经纬度对应多个POI名字的问题，利用互联网“投票”机制来选取最佳的POI名字。

综上所述，本发明的上述实施例通过对从网络数据中抓取地址数据进行名称字段和地址信息的提取，基于名称字段确定关键词，并将对应相同地址信息的关键词进行聚类，基于聚类后的关键词确定地址信息对应的POI名称，从而使得用户能够快速、准确地搜索到同一经、纬度的POI地址对应的POI名称，改善用户体验。

30 图15示意性示出了本发明一个实施例的基于网络中的地址数据确定POI信息有效性的系统的框图。

参照图15，本发明实施例的基于网络中的地址数据确定POI信息有效性的系统，包括：

POI信息获取单元511，用于基于搜索引擎利用网络中的地址数据获取对应相同POI名称的多个相关POI信息；

35 本发明实施例中，所述多个相关POI信息为对应POI至少一个预设属性的信息。进一步地，所述预设属性为经纬度、地址、建筑物名称或所囊括单位名称。

统计单元512，用于统计所述POI信息在所述网络中的地址数据中的出现次数；

POI信息确定单元513，用于根据所述POI信息在所述网络中的地址数据中的出现次数确定对应所述相同POI名称的有效POI信息。

40 由前述表1可见，在同一地理位置（经纬度相同）从不同来源网站获取的POI数据中，有可能存在重复性数据，即同一个地址（经纬度）可能存在多个POI名字；还可以看出，同一个poi名字可能多种不同的说法。

45 本发明实施例中，基于搜索引擎利用网络中的地址数据获取对应相同POI名称的多个相关POI信息，其中，多个相关POI信息为对应POI至少一个预设属性的信息，所述预设属性为经纬度、地址、建筑物名称或所囊括单位名称，根据所述POI信息在所述网络中的地址数据中的出现次数确定对应所述相同POI名称的有效POI信息。

进一步地，根据所述POI信息在所述网络中的地址数据中的出现次数确定对应所述相同POI名称的有效POI信息，包括：根据相关POI信息的预设属性的信息将对应相同地址信息的名称字段按照关键词进行聚类，统计聚类后各类别中名称字段出现的频次，作为第二频次，根据所述第二频次确定该类别对应该地址信息的POI名

称, 根据所述第二频次确定该类别对应该地址信息的POI名称, 利用互联网“投票”机制来选取相同POI名称的可信的、有效POI信息。

更进一步地, 基于所述名称字段确定一个或多个关键词, 将对应相同地址信息的所述关键词进行聚类, 根据聚类后的关键词确定聚类后的名称字段。

5 更进一步地, 对所述名称字段中的名称进行切词处理生成分词, 根据所述分词获取所述名称字段的关键词。

更进一步地, 统计对应相同地址信息的每个分词出现的频次, 作为第一频次, 根据所述第一频次生成所述名称字段的关键词, 具体为, 选择所述第一频次最小并且是非地名的分词作为所述名称字段的关键词。

10 更进一步地, 本发明可以将所述各个类中所述第二频次最高的名称字段作为类标识名称, 将每类标识名称均作为对应该地址信息的POI名称; 或, 将所述各个类中第二频次最高的名称字段作为类标识名称, 将网络上出现次数最多的类标识名称作为对应该地址信息的POI名称。

其中, 对所挖掘的地址数据中 POI 信息的名称切词, 并且统计切词后每个词出现的次数, 同一个 POI 名称中出现频次最少即包含的信息量最大, 并且是非地名的那个词记为该 POI 名称的关键词, 比如表 1 中出现的地址数据对应的相关 POI 信息中 POI 名称切词后数据如前述表 2 所示(词频是根据约 9000 万的 poi 名字统计的)。为了进一步体现发明的优越性, 如下进一步揭示本发明基于网络中的地址数据确定 POI 信息有效性的系统中的统计单元 512 的在另一实施例中的内部结构, 来体现依据统计单元 512 实现的另一实施例的细节。参照图 16, 统计单元 512 进一步包括 POI 信息来源获取模块 5121、POI 信息来源可靠性判断模块 5122 以及统计模块 5123:

所述的 POI 信息来源获取模块 5121, 用于获取所述 POI 信息的来源;

所述的 POI 信息来源可靠性判断模块 5122, 用于判断所述来源是否属于可靠来源;

20 所述的统计模块 5123, 用于在来源属于可靠来源的情况下统计所述 POI 信息在所述网络中的地址数据中的出现次数; 否则不统计。

本实施例中, 在同一类的 POI 名称中, 选取最佳的 POI 名称是根据互联上的“投票”来解决, 所谓“投票”主要是根据此 POI 名称在互联网上出现的频次以及来源的可信度, 互联网上出现的频次最高、来源最可信的那个名字为要选取的最佳名字。比如:

25 A 类中只有一个名字, 最佳的也是这一个。

B 类中有两个名字, 其中“云南省澜沧江啤酒集团保山有限公司”出现的频率最高, 作为最佳名字。

C 类中有两个名字, 其中“保山明志汽车销售服务有限公司”出现的频率最高, 作为最佳名字。

D 类和 E 类中同样是只有一个名字, 类似 A。

30 为了进一步体现发明的优越性, 如下进一步揭示本发明基于网络中的地址数据确定 POI 信息有效性的系统中的 POI 信息确定单元 513 的在另一实施例中的内部结构, 来体现依据 POI 信息确定单元 513 实现的另一实施例的细节。参照图 17, POI 信息确定单元 513 进一步包括判断子单元 5131 以及信息点信息确定子单元 5132:

所述的判断子单元 5131, 用于判断所述 POI 信息在所述网络中的地址数据中的出现次数是否高于预定阈值;

35 所述的信息点信息确定子单元 5132, 用于在所述判断子单元判断为是的情况下, 确定所获取的 POI 信息有效。

本发明实施例中, POI 信息在互联上出现的频率越高、来源的可信度越可信, 则 POI 信息越可信。对最终选取的最佳 POI 名字根据其在互联上出现的频次以及来源来过滤, 高于一定阈值的则为最终挖掘的可信的 POI 信息。

本发明实施例中, 所述可靠来源为具有预定可信度的来源。其中, 所述来源为网站或者网页。

40 本发明实施例中, 预定可信度的来源的网站或者网页包括但不限于, 如新浪、凤凰网等大型网站、通过官方认证的网站、访问频次比较高、数据流量大的网站以及不携带恶意链接、病毒链接且客户满意度交高的网站等。

45 本发明实施例中, 可信度是可量化的, 可根据用户的访问次数以及客户评价等对各个网站或网页的可信度进行量化。而且各个网站或网页的可信度是动态变化的, 若当前网站出现病毒、欺诈广告或被其他恶意欺诈网站所利用, 则其可信度会随之降低, 本发明通过网站可信度的量化和动态调整, 进一步保证获取的 POI 信息的可靠、有效。

本实施例对利用网络中的地址数据获取对应相同 POI 名称的多个相关 POI 信息, 根据 POI 信息在网络中的地址数据中的出现次数确定对应所述相同 POI 名称的有效 POI 信息, 从而使得用户能够快速、准确地搜索

到同一经、纬度的 POI 地址对应的一个或多个 POI 名称, 然后利用网络投票机制从一个或多个 POI 名称按照信息来源以及其在互联网上出现的频次进行过滤, 选出可信度高的 POI 名称作为当前 POI 地址对应的 POI 名称, 提高 POI 信息的有效性。

图18示意性示出了本发明一个实施例的基于网络中的地址数据确定POI信息有效性的方法的流程图。

5 参照图18, 本发明实施例的基于网络中的地址数据确定POI信息有效性的方法包括以下步骤:

S811、利用网络中的地址数据获取对应相同 POI 名称的多个相关 POI 信息;

S812、统计所述 POI 信息在所述网络中的地址数据中的出现次数;

S813、根据所述 POI 信息在所述网络中的地址数据中的出现次数确定对应所述相同 POI 名称的有效 POI 信息。

10 本发明实施例中, 所述多个相关 POI 信息为对应 POI 至少一个预设属性的信息。其中, 所述预设属性为经纬度、地址、建筑物名称或所囊括单位名称。

在同一地理位置(经纬度相同)从不同来源网站获取的 POI 数据中, 有可能存在重复性数据, 同一个 poi 名字可能多种不同的说法。

15 对此, 本发明实施例, 对所挖掘的地址数据中 POI 信息的名称切词, 并且统计切词后每个词出现的次数, 同一个 POI 名称中出现频次最少即包含的信息量最大, 并且是非地名的那个词记为该 POI 名称的关键词。

为了进一步体现发明的优越性, 如下进一步揭示本发明基于网络中的地址数据确定 POI 信息有效性的方法中步骤 S812 的细分步骤, 来体现依据本步骤实现的另一实施例。参照图 19, 本步骤的细分步骤包括:

S8121、获取所述 POI 信息的来源;

S8122、判断所述来源是否属于可靠来源, 如果是, 则执行步骤 S123;

20 S8123、当所述来源属于可靠来源时, 统计所述 POI 信息在所述网络中的地址数据里的出现次数, 否则不统计。

本实施例中, 在同一类的 POI 名称中, 选取最佳的 POI 名称是根据互联上的“投票”来解决, 所谓“投票”主要是根据此 POI 名称在互联网上出现的频次以及来源的可信度, 互联网上出现的频次最高、来源最可信的那个名字为要选取的最佳名字。

25 为了进一步体现发明的优越性, 如下进一步揭示本发明基于网络中的地址数据确定 POI 信息有效性的方法中步骤 S813 的细分步骤, 来体现依据本步骤实现的另一实施例。参照图 20, 本步骤的细分步骤包括:

S8131、判断所述 POI 信息在所述网络中的地址数据中的出现次数是否高于预定阈值; 如果是, 则执行步骤 S8132,

S8132、确定所述 POI 信息有效。

30 本发明实施例中, POI 信息在互联上出现的频率越高、来源的可信度越可信, 则 POI 信息越可信。对最终选取的最佳 POI 名字根据其在互联上出现的频次以及来源来过滤, 高于一定阈值的则为最终挖掘的可信的 POI 信息。

本发明实施例中, 所述可靠来源为具有预定可信度的来源。其中, 所述来源为网站或者网页。

35 本发明实施例中, 预定可信度的来源的网站或者网页包括但不限于, 如新浪、凤凰网等大型网站、通过官方认证的网站、访问频次比较高、数据流量大的网站以及不携带恶意链接、病毒链接且客户满意度交高的网站等。

40 本发明实施例中, 可信度是可量化的, 可根据用户的访问次数以及客户评价等对各个网站或网页的可信度进行量化。而且各个网站或网页的可信度是动态变化的, 若当前网站出现病毒、欺诈广告或被其他恶意欺诈网站所利用, 则其可信度会随之降低, 本发明通过网站可信度的量化和动态调整, 进一步保证获取的 POI 信息的可靠、有效。

通过采用本发明实施例提供的基于网络中的地址数据确定POI信息有效性的方法, 根据切词后词频次的多少来挖掘poi名字的关键词, 并且以此关键词来聚类, 把不同说法的同一个poi名字聚为一类, 解决同一个经纬度对应多个poi名字的问题, 利用互联网“投票”机制来选取最佳的poi名字, 利用互联上“投票”机制来选取可信的poi信息。

45 综上所述, 本发明上述实施例对利用网络中的地址数据获取对应相同POI名称的多个相关POI信息, 根据POI信息在网络中的地址数据中的出现次数确定对应所述相同POI名称的有效POI信息, 从而使得用户能够快速、准确地搜索到同一经、纬度的POI地址对应的一个或多个POI名称, 然后利用网络投票机制从一个或多个POI名称按照信息来源以及其在互联网上出现的频次进行过滤, 选出可信度高的POI名称作为当前POI地址对应

的POI名称,提高POI信息的有效性。

应当注意,在此提供的算法和公式不与任何特定计算机、虚拟系统或者其它设备固有相关。各种通用系统也可以与基于在此的示例一起使用。根据上面的描述,构造这类系统所要求的结构是显而易见的。此外,本
5 发明也不针对任何特定编程语言。应当明白,可以利用各种编程语言实现在此描述的本发明的内容,并且上面
对特定语言所做的描述是为了披露本发明的最佳实施方式。

在此处所提供的说明书中,说明了大量具体细节。然而,能够理解,本发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中,并未详细示出公知的方法、结构和技术,以便不模糊对本说明书的理解。

类似地,应当理解,为了精简本发明并帮助理解本发明各个方面中的一个或多个,在上面对本发明的示
10 例性实施例的描述中,本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而,并不
应将该公开的方法和装置解释成反映如下意图:即所要求保护的本发明要求比在每个权利要求中所明确记载的
特征更多的特征。更确切地说,如权利要求书所反映,发明方面在于少于前面公开的单个实施例的所有特征。
因此,遵循具体实施方式的权利要求书由此明确地并入该具体实施方式,其中每个权利要求本身都作为本发明的
单独实施例。

15 本领域那些技术人员可以理解,可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置
与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件,以
及此外可以把它们分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者单元中的至少一些是相
互排斥之外,可以采用任何组合对本说明书(包括伴随的权利要求、摘要和附图)中公开的所有特征以及如此
公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述,本说明书(包括伴随的权利要求、
20 摘要和附图)中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

此外,本领域的技术人员能够理解,尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而
不是其它特征,但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。

本发明的各个部件实施例可以以硬件实现,或者以在一个或者多个处理器上运行的软件模块实现,或者
以它们的组合实现。本领域的技术人员应当理解,可以在实践中使用微处理器或者数字信号处理器(DSP)来
25 实现根据本发明实施例的基于聚类确定 POI 名称的系统中的一些或者全部部件的一些或者全部功能。本发明
还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序(例如,计算机程序和计算机
程序产品)。这样的实现本发明的程序可以存储在计算机可读介质上,或者可以具有一个或者多个信号的形式。
这样的信号可以从因特网网站上下载得到,或者在载体信号上提供,或者以任何其他形式提供。

例如,图 21 示意性地示出了用于执行根据本发明的方法的计算设备的框图。该计算设备传统上包括处
30 理器 2110 和以存储器 2120 形式的计算机程序产品或者计算机可读介质。存储器 2120 可以是诸如闪存、
EEPROM(电可擦除可编程只读存储器)、EPROM、硬盘或者 ROM 之类的电子存储器。存储器 2120 具有用
于执行上述方法中的任何方法步骤的程序代码 2131 的存储空间 2130。例如,用于程序代码的存储空间 2130
可以包括分别用于实现上面的方法中的各种步骤的各个程序代码 2131。这些程序代码可以从一个或者多个计
算机程序产品中读出或者写入到这一个或者多个计算机程序产品中。这些计算机程序产品包括诸如硬盘、紧致
35 盘(CD)、存储卡或者软盘之类的程序代码载体。这样的计算机程序产品通常为如参考图 22 所述的便携式或
者固定存储单元。该存储单元可以具有与图 21 的计算设备中的存储器 2120 类似布置的存储段、存储空间等。
程序代码可以例如以适当形式进行压缩。通常,存储单元包括用于执行根据本发明的方法步骤的计算机可读代
码 2131',即可以由例如诸如 2110 之类的处理器读取的代码,这些代码当由计算设备运行时,导致该计算设
备执行上面所描述的方法中的各个步骤。

40 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制,并且本领域技术人员在不脱离所
附权利要求的范围的情况下可设计出替换实施例。在权利要求中,不应将位于括号之间的任何参考符号构造成
对权利要求的限制。单词“包括”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或
“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计
算机来实现。在列举了若干装置的单元权利要求中,这些装置中的若干个可以是通过同一个硬件项来具体体现。
45 单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

此外,还应当注意,本说明书中使用的语言主要是为了可读性和教导的目的而选择的,而不是为了解释
或者限定本发明的主题而选择的。因此,在不偏离所附权利要求书的范围和精神的情况下,对于本技术领域的
普通技术人员来说许多修改和变更都是显而易见的。对于本发明的范围,对本发明所做的公开是说明性的,而

非限制性的，本发明的范围由所附权利要求书限定。

本发明可以应用于计算机系统/服务器，其可与众多其它通用或专用计算系统环境或配置一起操作。适于与计算机系统/服务器一起使用的众所周知的计算系统、环境和/或配置的例子包括但不限于：个人计算机系统、服务器计算机系统、瘦客户机、厚客户机、手持或膝上设备、基于微处理器的系统、机顶盒、可编程消费电子产品、网络个人电脑、小型计算机系统、大型计算机系统和包括上述任何系统的分布式云计算技术环境，等等。

计算机系统/服务器可以在由计算机系统执行的计算机系统可执行指令（诸如程序模块）的一般语境下描述。通常，程序模块可以包括例程、程序、目标程序、组件、逻辑、数据结构等等，它们执行特定的任务或者实现特定的抽象数据类型。计算机系统/服务器可以在分布式云计算环境中实施，分布式云计算环境中，任务是由通过通信网络链接的远程处理设备执行的。在分布式云计算环境中，程序模块可以位于包括存储设备的本地或远程计算系统存储介质上。

本文中所称的“一个实施例”、“实施例”或者“一个或者多个实施例”意味着，结合实施例描述的特定特征、结构或者特性包括在本发明的至少一个实施例中。此外，请注意，这里“在一个实施例中”的词语例子不一定全指同一个实施例。

以上所述仅是本发明的部分实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

权 利 要 求

- 1、一种基于聚类确定 POI 名称的系统，该系统包括：
地址数据抓取器，用于从网络数据中抓取地址数据；
5 地址数据解析器，用于从抓取到的一个或多个地址数据中分别提取名称字段和地址信息；
关键词确定器，用于基于所述名称字段确定一个或多个关键词；
关键词聚类器，用于将对应相同地址信息的所述关键词进行聚类，生成至少一个类；
POI 名称生成器，用于根据聚类后的关键词确定此地址信息对应的 POI 名称。
- 2、如权利要求 1 所述的系统，所述关键词确定器进一步包括：
10 切词单元，用于对所述名称字段中的名称进行切词处理生成分词；
关键词获取单元，用于根据所述分词获取所述地址数据的关键词。
- 3、如权利要求 1-2 任一项所述的系统，所述关键词获取单元进一步包括：
第一频次统计模块，用于统计对应相同地址信息的每个分词出现的频次，作为第一频次；
关键词生成模块，用于根据所述第一频次生成所述地址数据的关键词。
- 15 4、如权利要求 1-3 任一项所述的系统，所述关键词生成模块选择频次最小并且是非地名的分词作为所述地址数据的关键词。
- 5、如权利要求 1-4 任一项所述的系统，所述 POI 名称生成器进一步包括：
频率统计单元，用于计算各个类中名称字段的出现频率；
类标识名称确定单元，用于将所述各个类中出现频率最高的名称字段作为类标识名称；
20 POI 名称确定单元，用于将每个类标识名称均作为 POI 名称。
- 6、如权利要求 1-4 任一项所述的系统，所述 POI 名称生成器进一步包括：
频率统计单元，用于计算各个类中名称字段的出现频率；
类标识名称确定单元，用于将所述各个类中出现频率最高的名称字段作为类标识名称；
POI 名称确定单元，用于选择出现频率最高的类标识名称作为 POI 名称。
- 25 7、一种基于聚类确定 POI 名称的方法，包括：
从网络数据中抓取地址数据；
从抓取到的一个或多个地址数据中分别提取名称字段和地址信息；
基于所述名称字段确定一个或多个关键词；
将对应相同地址信息的所述关键词进行聚类，生成至少一个类；
30 根据聚类后的关键词确定此地址信息对应的 POI 名称。
- 8、如权利要求 7 所述的方法，所述步骤：基于所述名称字段确定一个或多个关键词，进一步包括：
对所述名称字段中的名称进行切词处理生成分词；
根据所述分词获取所述地址数据的关键词。
- 9、如权利要求 7-8 任一项所述的方法，所述步骤：根据所述分词获取所述地址数据的关键词，进一步
35 包括：
统计对应相同地址信息的每个分词出现的频次作为第一频次；
根据所述第一频次生成所述地址数据的关键词。
- 10、如权利要求 7-9 任一项所述的方法，所述步骤根据所述第一频次生成所述地址数据的关键词具体为：
选择频次最小并且是非地名的分词作为所述地址数据的关键词。
- 40 11、如权利要求 7-10 任一项所述的方法，所述步骤：根据聚类后的关键词确定此地址信息对应的 POI 名称，进一步包括：
计算各个类中名称字段的出现频率；
将所述各个类中出现频率最高的名称字段作为类标识名称；
将每个类标识名称均作为 POI 名称。
- 45 12、如权利要求 7-11 任一项所述的方法，所述步骤：根据聚类后的关键词确定此地址信息对应的 POI 名称，进一步包括：
计算各个类中名称字段的出现频率；
将所述各个类中出现频率最高的名称字段作为类标识名称；

选择出现频率最高的类标识名称作为 POI 名称。

13、一种基于聚类的 POI 名称确定系统，包括：

地址数据抓取器，用于基于搜索引擎从网络数据中抓取地址数据，所述地址数据包括名称字段和地址信息；

- 5 名称字段聚类器，用于将对应相同地址信息的名称字段按照关键词进行聚类；
第二频次统计器，用于统计聚类后各类别中名称字段出现的频次，作为第二频次；
POI 名称确定单元，用于根据所述第二频次确定该类别对应该地址信息的 POI 名称。

14、如权利要求 13 所述系统，所述名称字段聚类器进一步包括：

- 10 关键词确定单元，用于基于所述名称字段确定一个或多个关键词；
关键词聚类单元，用于将对应相同地址信息的所述关键词进行聚类；
名称字段聚类确定单元，用于根据聚类后的关键词确定聚类后的名称字段。

15、如权利要求 13-14 任一项所述的系统，所述关键词确定单元进一步包括：

切词模块，用于对所述名称字段中的名称进行切词处理生成分词；

关键词获取模块，用于根据所述分词获取所述名称字段的关键词。

- 15 16、如权利要求 13-15 任一项所述的系统，所述关键词获取模块进一步包括：

第一频次统计子模块，用于统计对应相同地址信息的每个分词出现的频次，作为第一频次；

关键词生成子模块，用于根据所述第一频次生成所述名称字段的关键词。

17、如权利要求 13-16 任一项所述的系统，所述关键词生成子模块选择所述第一频次最小并且是非地名的分词作为所述名称字段的关键词。

- 20 18、如权利要求 13-17 任一项所述系统，所述第二频次统计器进一步包括：

名称字段来源获取单元，用于获取所述名称字段的来源；

来源可靠性判断单元，用于判断所述来源是否属于可靠来源；

第二频次统计单元，用于在判断为是的情况下，统计所述名称字段出现的频次，作为第二频次，否则不统计。

- 25 19、如权利要求 13-18 任一项所述系统，所述 POI 名称确定单元进一步包括：

类标识名称确定模块，用于将所述各个类中所述第二频次最高的名称字段作为类标识名称；

第一 POI 名称确定模块，用于将每类标识名称均作为对应该地址信息的 POI 名称。

20、如权利要求 13-19 任一项所述系统，所述 POI 名称确定单元进一步包括：

类标识名称确定模块，用于将所述各个类中第二频次最高的名称字段作为类标识名称；

- 30 第二 POI 名称确定模块，用于将网络上出现次数最多的类标识名称作为对应该地址信息的 POI 名称。

21、一种基于聚类的 POI 名称确定方法，包括：

从网络数据中抓取地址数据，所述地址数据包括名称字段和地址信息；

将对应相同地址信息的名称字段按照关键词进行聚类；

- 35 统计聚类后各类别中名称字段出现的频次，作为第二频次；

根据所述第二频次确定该类别对应该地址信息的 POI 名称。

22、如权利要求 21 所述方法，所述将对应相同地址信息的名称字段按照关键词进行聚类，进一步包括：

基于所述名称字段确定一个或多个关键词；

- 40 将对应相同地址信息的所述关键词进行聚类；

根据聚类后的关键词确定聚类后的名称字段。

23、如权利要求 21-22 任一项所述的方法，所述基于所述名称字段确定一个或多个关键词，进一步包括：

对所述名称字段进行切词处理生成分词；

- 45 根据分词获取所述名称字段的关键词。

24、如权利要求 21-23 任一项所述的方法，所述根据分词获取所述名称字段的关键词，进一步包括：

统计对应相同地址信息的每个分词出现的频次，作为第一频次；

根据所述第一频次确定所述名称字段的关键词。

25、如权利要求 21-24 任一项所述的方法，所述根据所述第一频次确定所述名称字段的关键词具体为：

选择第一频次最小并且是非地名的分词作为所述名称的关键词。

5 26、如权利要求 21-25 任一项所述的方法，所述统计聚类后各类别中名称字段出现的频次，作为第二频次，进一步包括：

获取所述名称字段的来源；

判断所述来源是否属于可靠来源，如果是，则统计所述名称字段出现的频次，作为第二频次。

10 27、如权利要求 21-26 任一项所述方法，所述根据所述第二频次确定该类别对应该地址信息的 POI 名称，进一步包括：

将所述各个类中所述第二频次最高的名称字段作为类标识名称；

将每类标识名称均作为对应该地址信息的 POI 名称。

15 28、如权利要求 21-27 任一项所述方法，所述根据所述第二频次确定该类别对应该地址信息的 POI 名称，进一步包括：

将所述各个类中所述第二频次最高的名称字段作为类标识名称；

将网络上出现次数最多的类标识名称作为对应该地址信息的 POI 名称。

29、一种基于网络中的地址数据确定 POI 信息有效性的系统，该系统包括：

20 POI 信息获取单元，用于基于搜索引擎利用网络中的地址数据获取对应相同 POI 名称的多个相关 POI 信息；

统计单元，用于统计所述 POI 信息在所述网络中的地址数据中的出现次数；

POI 信息确定单元，用于根据所述 POI 信息在所述网络中的地址数据中的出现次数确定对应所述相同 POI 名称的有效 POI 信息。

30、如权利要求 29 所述的系统，所述多个相关 POI 信息为对应 POI 至少一个预设属性的信息。

25 31、如权利要求 29-30 任一项所述的系统，所述预设属性为经纬度、地址、建筑物名称或所囊括单位名称。

32、如权利要求 29-31 任一项所述的系统，所述统计单元进一步包括：

POI 信息来源获取模块，用于获取所述 POI 信息的来源；

POI 信息来源可靠性判断模块，用于判断所述来源是否属于可靠来源；

30 统计模块，用于在来源属于可靠来源的情况下统计所述 POI 信息在所述网络中的地址数据中的出现次数；否则不统计。

33、如权利要求 29-32 任一项所述的系统，所述 POI 信息确定单元进一步包括：

判断子单元，用于判断所述 POI 信息在所述网络中的地址数据中的出现次数是否高于预定阈值；

信息点信息确定子单元，用于在所述判断子单元判断为是的情况下，确定所获取的 POI 信息有效。

34、如权利要求 29-33 任一项所述的系统，所述可靠来源为具有预定可信度的来源。

35 35、如权利要求 29-34 任一项所述的系统，所述来源为网站或者网页。

36、一种基于网络中的地址数据确定 POI 信息有效性的方法，该方法包括：

利用网络中的地址数据获取对应相同 POI 名称的多个相关 POI 信息；

统计所述 POI 信息在所述网络中的地址数据中的出现次数；

40 根据所述 POI 信息在所述网络中的地址数据中的出现次数确定对应所述相同 POI 名称的有效 POI 信息。

37、如权利要求 36 所述的方法，所述多个相关 POI 信息为对应 POI 至少一个预设属性的信息。

38、如权利要求 36-37 任一项所述的方法，所述预设属性为经纬度、地址、建筑物名称或所囊括单位名称。

45 39、如权利要求 36-38 任一项所述的方法，所述步骤：统计所述 POI 信息在所述网络中的地址数据中的出现次数，进一步包括：

获取所述 POI 信息的来源；

判断所述来源是否属于可靠来源，如果是，则统计所述 POI 信息在所述网络中的地址数据里的出现次数，否则不统计。

40、如权利要求 36-39 任一项所述的方法，所述步骤：根据所述 POI 信息在所述网络中的地址数据中的出现次数确定对应所述相同 POI 名称的有效 POI 信息，进一步包括：

判断所述 POI 信息在所述网络中的地址数据中的出现次数是否高于预定阈值；

如果是，则确定所述 POI 信息有效。

5 41、如权利要求 36-40 任一项所述的方法，所述可靠来源为具有预定可信度的来源。

42、如权利要求 36-41 任一项所述的方法，所述来源为网站或者网页。

43、一种计算机程序，包括计算机可读代码，当所述计算机可读代码在计算设备上运行时，导致所述计算设备执行根据权利要求 7-12 中的任一项所述的基于聚类确定 POI 名称的方法，或者，导致所述计算设备执行根据权利要求 21-28 中的任一项所述的基于聚类的 POI 名称确定方法，或者导致所述计算设备执行根据权利要求 36-42 中的任一项所述的基于网络中的地址数据确定 POI 信息有效性的方法。

10 44、一种计算机可读介质，其中存储了如权利要求 43 所述的计算机程序。

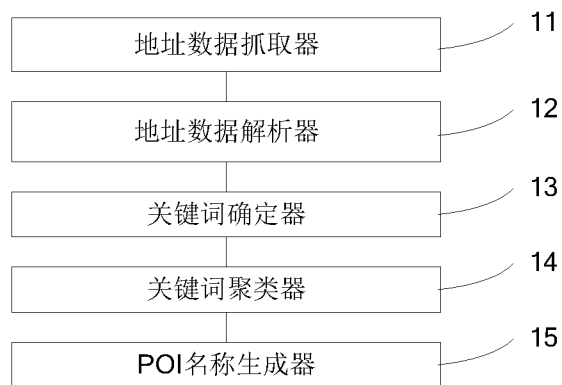


图 1

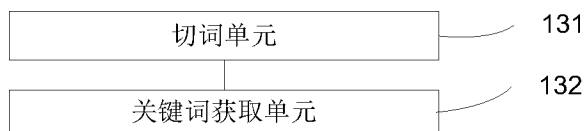


图 2

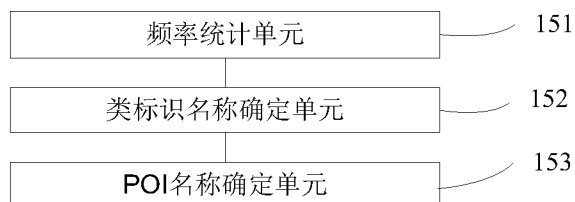


图 3

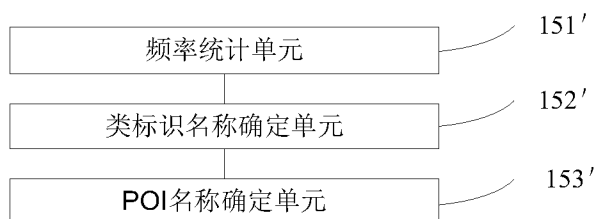


图 4

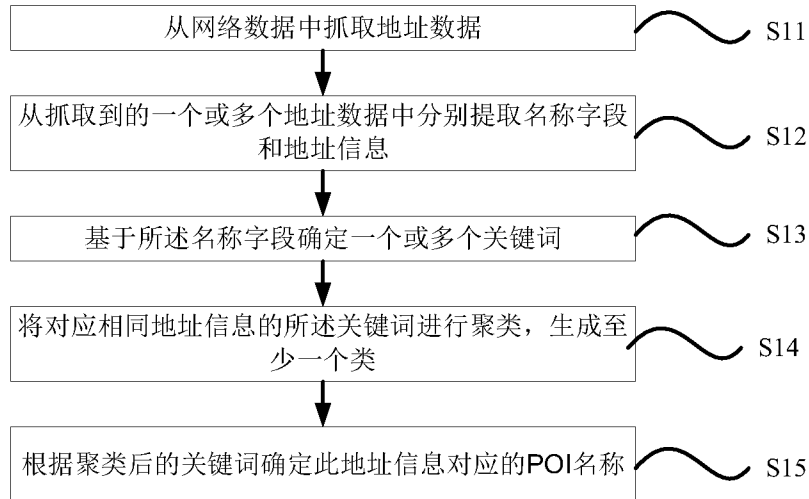


图 5

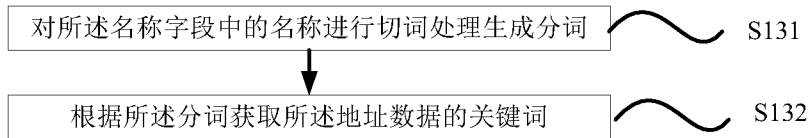


图 6

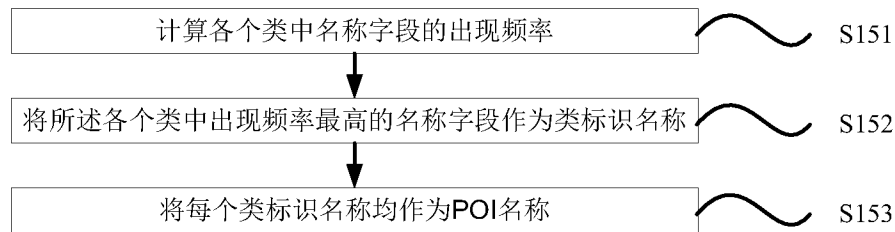


图 7

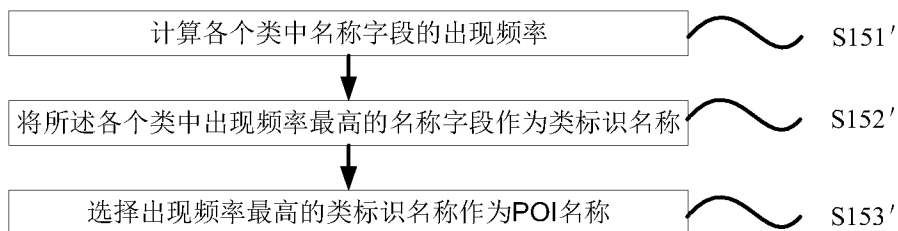


图 8

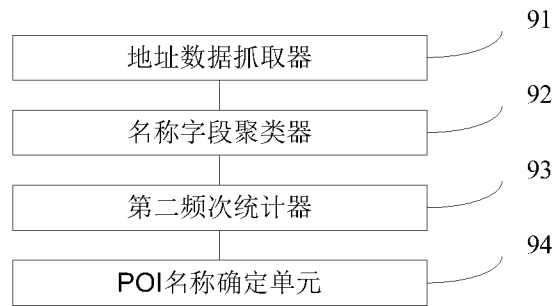


图 9

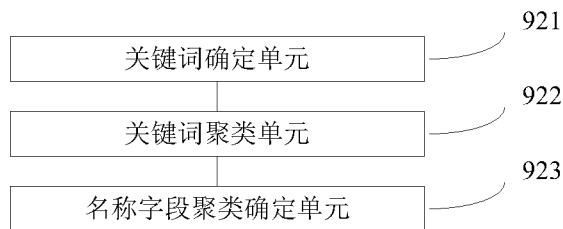


图 10

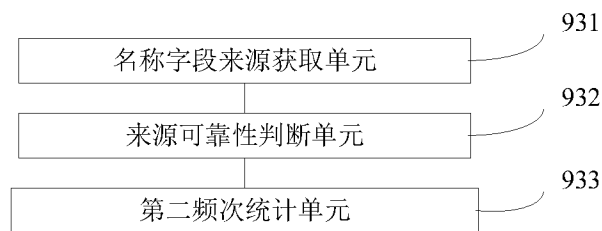


图 11

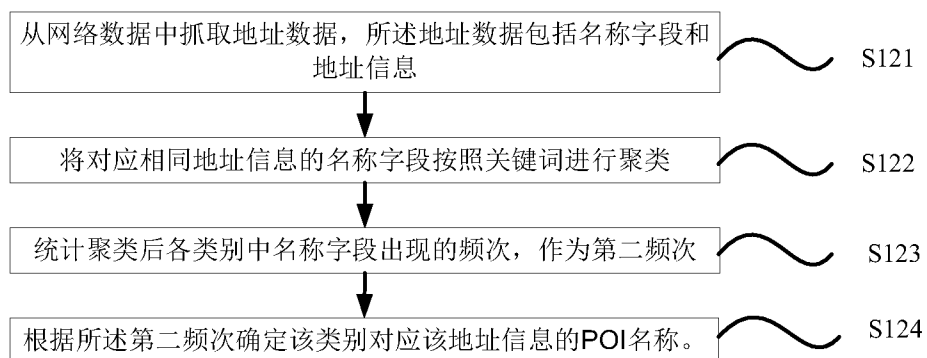


图 12

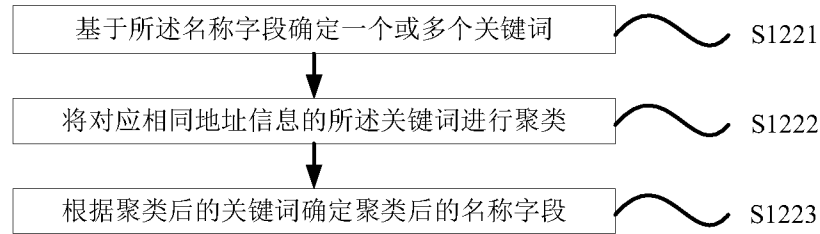


图 13

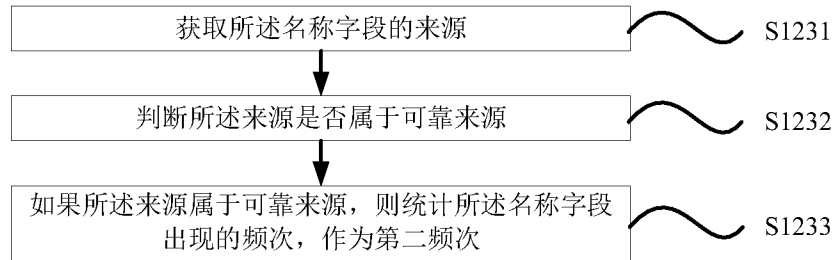


图 14

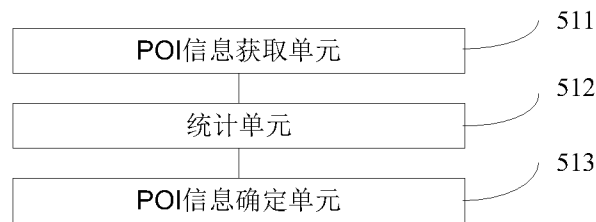


图 15

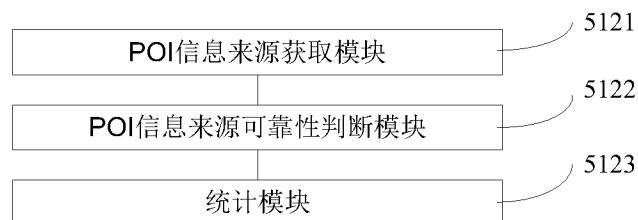


图 16

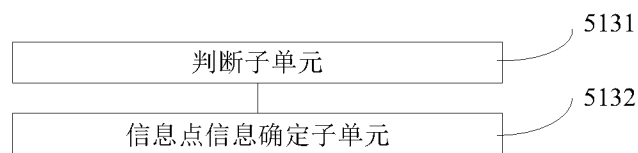


图 17

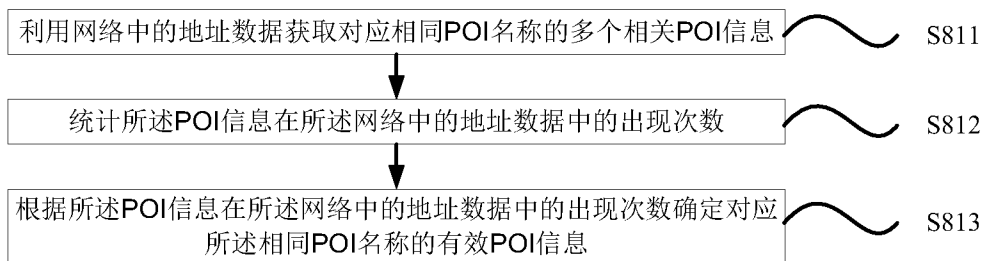


图 18

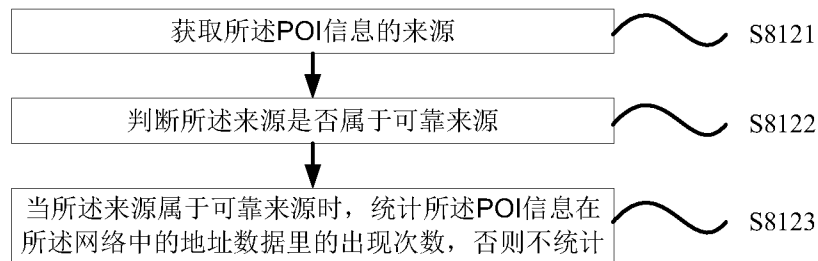


图 19

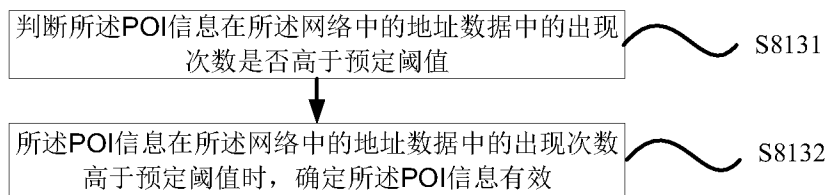


图 20

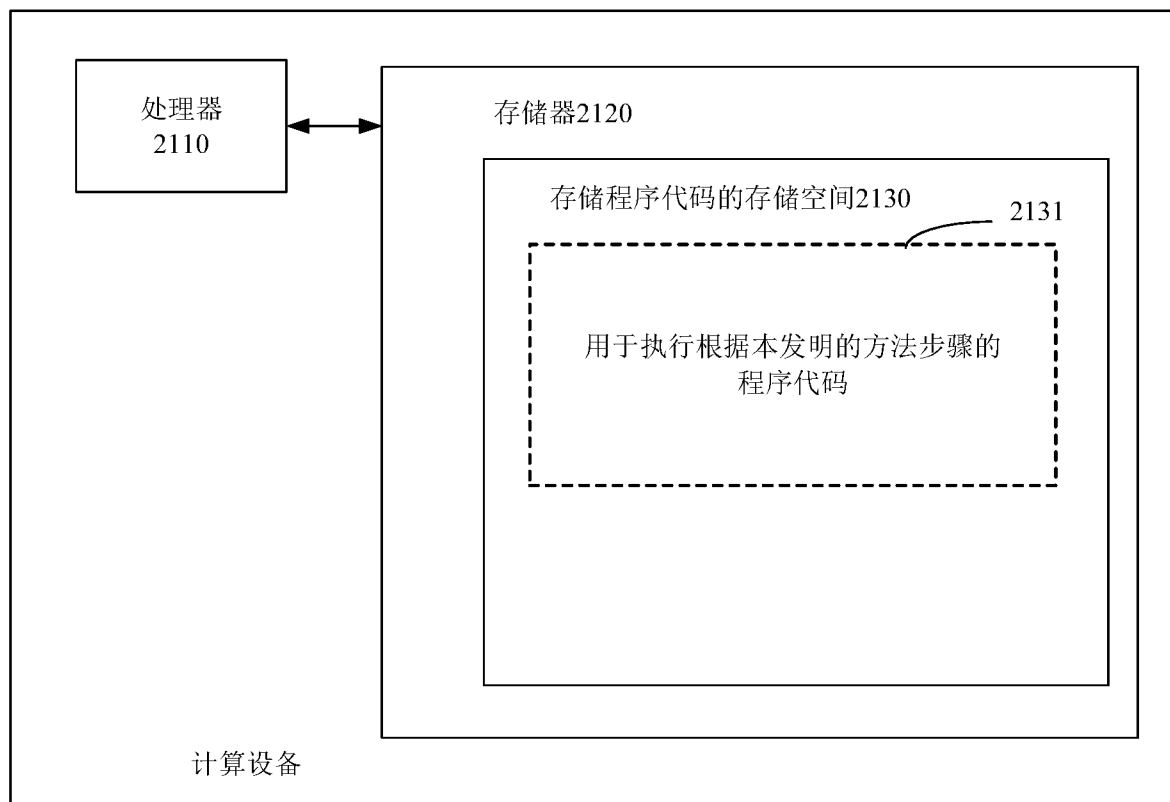


图21

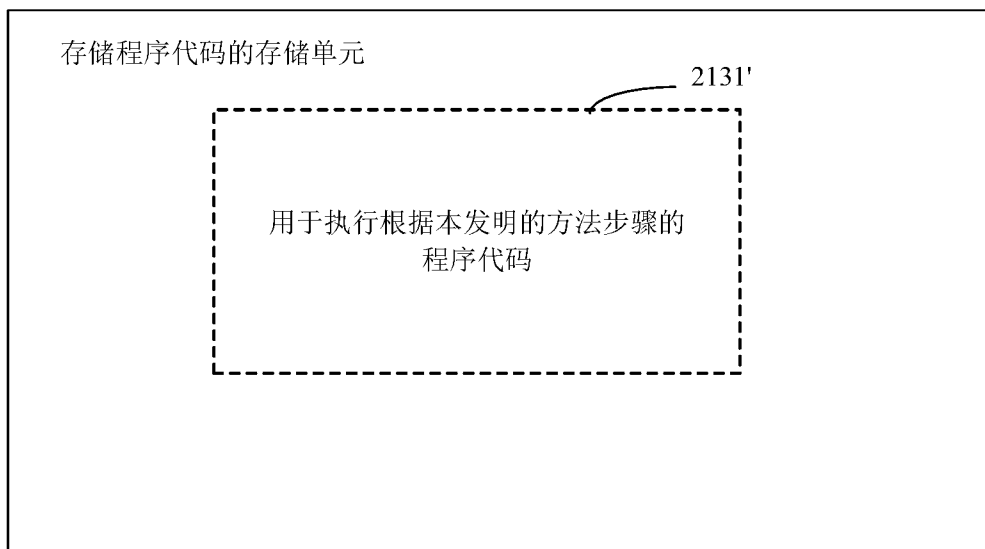


图22

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2015/095857

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)

G06F

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)

CNPAT, CNKI, WPI, EPODOC, GOOGLE: POI, frequency of occurrence, name, address, keyword, clustering, frequency

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 104572957 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.) 29 April 2015 (29.04.2015) claims 1-10	1-12, 43, 44
PX	CN 104572956 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.) 29 April 2015 (29.04.2015) claims 1-10	29-44
PX	CN 104572955 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.) 29 April 2015 (29.04.2015) claims 1-10	13-28, 43, 44
Y	LI, Ruishan, "Multi-source Poi Information Fusion Based on Natural Language Processing" Master degree theses of master of China excellent full-text database, Information Technology Series, vol., no. 3, 15 March 2014 (15.03.2014), ISSN:1674-0246, pages 1-4, 17, 18-20 and 31-35	1-44

☒ 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 2 February 2016	Date of mailing of the international search report 24 February 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 ZHANG, Xia Telephone No. (86-10) 61648105

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2015/095857

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 102063460 A (CAI, Lianghua) 18 May 2011 (18.05.2011) description, paragraphs [0044]-[0058], and figure 2	1-44
A	CN 104077295 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 01 October 2014 (01.10.2014) the whole document	1-44
A	CN 102479229 A (BEIJING NAVINFO CO., LTD.) 30 May 2012 (30.05.2012) the whole document	1-44
A	US 2009110302 A1 (MICROSOFT CORPORATION) 30 April 2009 (30.04.2009) the whole document	1-44
A	JP 2014086045 A (KDDI CORP.) 12 May 2014 (12.05.2014) the whole document	1-44

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2015/095857

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104572957 A	29 April 2015	None	
CN 104572956 A	29 April 2015	None	
CN 104572955 A	29 April 2015	None	
CN 102063460 A	18 May 2011	None	
CN 104077295 A	01 October 2014	None	
CN 102479229 A	30 May 2012	None	
US 2009110302 A1	30 April 2009	None	
JP 2014086045 A	12 May 2014	None	

国际检索报告

国际申请号

PCT/CN2015/095857

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)

G06F

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))

CNPAT, CNKI, WPI, EPODOC, GOOGLE: POI, 名称, 地址, 关键词, 聚类, 频次, 出现次数, name, address, keyword, clustering, frequency

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
PX	CN 104572957 A (北京奇虎科技有限公司 等) 2015年 4月 29日 (2015 - 04 - 29) 权利要求1-10	权利要求1-12, 43, 44
PX	CN 104572956 A (北京奇虎科技有限公司 等) 2015年 4月 29日 (2015 - 04 - 29) 权利要求1-10	权利要求29-44
PX	CN 104572955 A (北京奇虎科技有限公司 等) 2015年 4月 29日 (2015 - 04 - 29) 权利要求1-10	权利要求13-28, 43, 44
Y	李瑞娜. "基于自然语言处理的多源POI数据融合的研究" 中国优秀硕士学位论文全文数据库, 信息科技辑, 第/卷, 第3期, 2014年 3月 15日 (2014 - 03 - 15), ISSN: 1674-0246, 第1-4, 17, 18-20, 31-35页	1-44
Y	CN 102063460 A (蔡亮华) 2011年 5月 18日 (2011 - 05 - 18) 说明书第[0044]-[0058]段, 图2	1-44

☒ 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 2月 2日

国际检索报告邮寄日期

2016年 2月 24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

授权官员

张霞

传真号 (86-10)62019451

电话号码 (86-10)61648105

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
A	CN 104077295 A (百度在线网络技术北京有限公司) 2014年 10月 1日 (2014 - 10 - 01) 全文	1-44
A	CN 102479229 A (北京四维图新科技股份有限公司) 2012年 5月 30日 (2012 - 05 - 30) 全文	1-44
A	US 2009110302 A1 (MICROSOFT CORPORATION) 2009年 4月 30日 (2009 - 04 - 30) 全文	1-44
A	JP 2014086045 A (KDDI 株式会社) 2014年 5月 12日 (2014 - 05 - 12) 全文	1-44

国际申请号	PCT/CN2015/095857
-------	-------------------

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	104572957	A	2015年 4月 29日	无	
CN	104572956	A	2015年 4月 29日	无	
CN	104572955	A	2015年 4月 29日	无	
CN	102063460	A	2011年 5月 18日	无	
CN	104077295	A	2014年 10月 1日	无	
CN	102479229	A	2012年 5月 30日	无	
US	2009110302	A1	2009年 4月 30日	无	
JP	2014086045	A	2014年 5月 12日	无	

表 PCT/ISA/210 (同族专利附件) (2009年7月)