(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106569994 B
(45)授权公告日 2019.02.26

(21)申请号 201510652677.8

(22)申请日 2015.10.10

(65)同一申请的已公布的文献号
申请公布号 CN 106569994 A

(43)申请公布日 2017.04.19

(73)专利权人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱

(72)发明人 陆青

(74)专利代理机构 北京亿腾知识产权代理事务所 11309
代理人 陈霁

(51)Int.Cl.
G06F 17/27(2006.01)

(56)对比文件
CN 103399907 A,2013.11.20,
CN 104679728 A,2015.06.03,
CN 104699668 A,2015.06.10,
US 5423032 A,1995.06.06,
CN 103425640 A,2013.12.04,

审查员 董立波

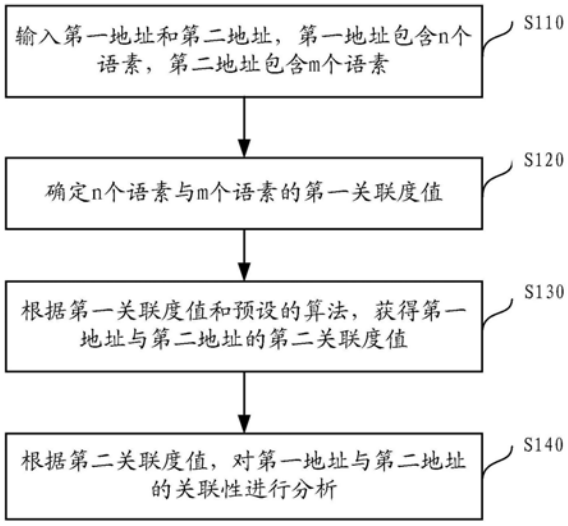
权利要求书2页 说明书6页 附图1页

(54)发明名称

地址的分析方法及装置

(57)摘要

本申请实施例涉及一种地址的分析方法及装置，包括：确定第一地址中n个语素与第二地址中m个语素的第一关联度值；根据所述第一关联度值和预设的算法，获得所述第一地址与所述第二地址的第二关联度值；根据所述第二关联度值，对所述第一地址与所述第二地址的关联性进行分析。由此，可以提高地址的分析的准确性和适用性。



CN 106569994 B

1.一种地址的分析方法,其特征在于,所述方法包括:

输入第一地址和第二地址,所述第一地址包含n个语素,所述第二地址包含m个语素,其中,所述语素是指所述地址中最小的语义单位,n和m均为自然数;

确定所述n个语素与所述m个语素的第一关联度值;

根据所述第一关联度值和预设的算法,从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素,并记录所述第一语素与所述第二语素的目标关联度值;针对所述n个语素中的各个语素执行此操作,直至记录n个目标关联度值;

根据所述n个目标关联度值,获得所述第一地址与所述第二地址的第二关联度值;

根据所述第二关联度值,对所述第一地址与所述第二地址的关联性进行分析。

2.根据权利要求1所述的方法,其特征在于,所述确定所述n个语素与所述m个语素的第一关联度值,包括:

对所述n个语素中的每个语素,确定所述每个语素与所述m个语素中各个语素的第一关联度值,以获得n×m个第一关联度值。

3.根据权利要求2所述的方法,其特征在于,所述根据所述第一关联度值和预设的算法,从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素,包括:

根据所述n×m个第一关联度值,构建n×m的矩阵;

根据预设的算法,对所述n×m的矩阵进行预处理;

根据预处理后的n×m的矩阵,从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素。

4.根据权利要求1-3任一项所述的方法,其特征在于,所述第一关联度值包括:

编辑距离值、汉明距离值、杰卡德距离值、N邻近字N-Gram距离值、JW距离值或者余弦距离值。

5.一种地址的分析装置,其特征在于,所述装置包括:输入单元、确定单元、获取单元和分析单元;

所述输入单元,用于输入第一地址和第二地址,所述第一地址包含n个语素,所述第二地址包含m个语素,其中,所述语素是指所述地址中最小的语义单位,n和m均为自然数;

所述确定单元,用于确定所述n个语素与所述m个语素的第一关联度值;

所述获取单元,用于从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素,并记录所述第一语素与所述第二语素的目标关联度值;针对所述n个语素中的各个语素执行此操作,直至记录n个目标关联度值;以及根据所述n个目标关联度值,获得所述第一地址与所述第二地址的第二关联度值;

所述分析单元,用于根据所述获取单元获得的所述第二关联度值,对所述第一地址与所述第二地址的关联性进行分析。

6.根据权利要求5所述的装置,其特征在于,所述确定单元具体用于:

对所述n个语素中的每个语素,确定所述每个语素与所述m个语素中各个语素的第一关联度值,以获得n×m个第一关联度值。

7.根据权利要求6所述的装置,其特征在于,所述根据所述第一关联度值和预设的算法,从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素,包括:

根据所述n×m个第一关联度值,构建n×m的矩阵;

根据预设的算法,对所述n×m的矩阵进行预处理;

根据预处理后的n×m的矩阵,从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素。

8.根据权利要求5-7任一项所述的装置,其特征在于,所述第一关联度值包括:

编辑距离值、汉明距离值、杰卡德距离值、N邻近字N-Gram距离值、JW距离值或者余弦距离值。

地址的分析方法及装置

技术领域

[0001] 本申请涉及计算机技术领域，尤其涉及一种地址的分析方法及装置。

背景技术

[0002] 传统技术中，在对地址进行分析时，如，在对两个英文地址的关联性进行分析时，直接将两个英文地址作为两个长字符串，通过计算两个长字符串间的编辑距离来对两个英文地址的关联性进行分析；或者，先对两个英文地址进行预处理，以对一个英文地址进行预处理为例来说，可以将英文地址中的单词按照首字母进行排序，然后删除重复的单词；之后再将预处理后的两个英文地址作为两个长字符串，计算两个长字符串间的编辑距离，最后根据编辑距离来对两个英文地址的关联性进行分析。

[0003] 然而，上述第一种方法对英文地址中单词的位置依赖性较大，一旦英文地址中有前后顺序颠倒或重复的情况发生，就会导致编辑距离的扩大，这影响了地址分析的准确性；而第二种方法，虽然减小了单词的顺序及重复的影响，但对单词的首字母有了依赖性，而对于非英语母语的英文地址而言，部分单词的首字母拼写确实有着若干种不同的用法(比如俄语中，Hl inka与Gl inka均是Глинка的常见英文写法)，这种情况下，会导致编辑距离更大化的扩大，也即第二种方法的适用性较差。

发明内容

[0004] 本申请实施例提供了一种地址的分析方法及装置，可以提高地址的分析的准确性和适用性。

[0005] 第一方面，提供了一种地址的分析方法，该方法包括：

[0006] 输入第一地址和第二地址，所述第一地址包含n个语素，所述第二地址包含m个语素，其中，所述语素是指所述地址中最小的语义单位，n和m均为自然数；

[0007] 确定所述n个语素与所述m个语素的第一关联度值；

[0008] 根据所述第一关联度值和预设的算法，获得所述第一地址与所述第二地址的第二关联度值；

[0009] 根据所述第二关联度值，对所述第一地址与所述第二地址的关联性进行分析。

[0010] 第二方面，提供了一种地址的分析装置，该装置包括：输入单元、确定单元、获取单元和分析单元；

[0011] 所述输入单元，用于输入第一地址和第二地址，所述第一地址包含n个语素，所述第二地址包含m个语素，其中，所述语素是指所述地址中最小的语义单位，n和m均为自然数；

[0012] 所述确定单元，用于确定所述n个语素与所述m个语素的第一关联度值；

[0013] 所述获取单元，用于根据所述确定单元确定的所述第一关联度值和预设的算法，获得所述第一地址与所述第二地址的第二关联度值；

[0014] 所述分析单元，用于根据所述获取单元获得的所述第二关联度值，对所述第一地址与所述第二地址的关联性进行分析。

[0015] 本申请提供的地址的分析方法及装置，确定第一地址中n个语素与第二地址中m个语素的第一关联度值；根据所述第一关联度值和预设的算法，获得所述第一地址与所述第二地址的第二关联度值；根据所述第二关联度值，对所述第一地址与所述第二地址的关联性进行分析。即本申请中，首先确定两个地址中语素间的第一关联度值，然后根据语素间的第一关联度值和预设的算法获得地址见的第二关联度值，根据第二关联度值对地址进行分析，由此，可以避免现有技术中将两个地址作为两个字符串，然后通过计算两个字符串间的编辑距离来对地址的关联性进行分析时，导致的对单个词的依赖性较大的问题，从而可以提高地址的分析的准确性和适用性。

附图说明

[0016] 图1为本申请一种实施例提供的地址的分析方法流程图；

[0017] 图2为本申请另一种实施例提供的地址的分析装置示意图。

具体实施方式

[0018] 为使本申请实施例的目的、技术方案和优点更加清楚，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，所描述的实施例是本申请一部分实施例，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。

[0019] 为便于对本申请实施例的理解，下面将结合附图以具体实施例做进一步的解释说明，实施例并不构成对本申请实施例的限定。

[0020] 本申请实施例提供的地址的分析方法及装置，适用于对地址间的关联性进行分析的场景，此处的关联性可以包括差异性和相似性，如，可以用于对实物交易的电子商务交易中的地址间的关联性进行分析。

[0021] 需要说明的是，上述地址间的关联性的分析结果可以作为聚类分析、词频分析以及地址标准化等的基础。

[0022] 图1为本申请一种实施例提供的地址的分析方法流程图。所述方法的执行主体可以为具有处理能力的设备：服务器或者系统或者装置，如图1所示，所述方法具体可以包括：

[0023] 步骤110，输入第一地址和第二地址，所述第一地址包含n个语素，所述第二地址包含m个语素，其中，所述语素是指所述地址中最小的语义单位，n和m均为自然数。

[0024] 此处，第一地址和第二地址的定义相同，以第一地址为例来说，第一地址可以包括中文地址或者英文地址等。对于中文地址，则先要对中文地址进行标准化处理，如，将繁体字转换为简体字等，之后在对标准化处理后的中文地址进行分词处理，就可以得到n个词，将该n个词作为n个语素；对于英文地址，则直接可以将第一地址中包含的n个单词作为n个语素。

[0025] 可以理解的是，当第一地址或者第二地址为中文地址时，则一个语素可以为一个词；而当第一地址或者第二地址为英文地址时，则一个语素可以为一个单词。

[0026] 步骤120，确定所述n个语素与所述m个语素的第一关联度值。

[0027] 此处，第一关联度值包括差异度值和相似度值，其中，差异度值可以包括：编辑距离值；而相似度值可以包括：汉明(hamming)距离值、杰卡德(Jaccard)距离值、N邻近字(N-

Gram)距离值、Jaro-Winkler距离值或者余弦(cos ine)距离值。在此说明书中,以第一关联度值为编辑距离值为例来说,编辑距离值是针对两个语素而言的,即一个语素经过多少次编辑变换(包括:删除变换、插入变换和替换变换等)可以变为另一个语素,其可以是根据经典的编辑距离算法计算的,也可以是根据调整过的编辑距离算法(如,增加替换变换的差异度或者减少元音缺失的差异度等)计算的。举例来说,假设一个语素为cafe,另一个语素为coffee,从cafe变为coffee的过程为:cafe→caffe→coffe→coffee,也即需要经过三次编辑变换,因此cafe与coffee的编辑距离值为3。

[0028] 需要说明的是,在第一地址或者第二地址为中文地址时,则在执行步骤120之前,还可以对第一地址或者第二地址中包含的语素做如下处理:

[0029] 以第一地址为例来说,可以将第一地址中的每个语素(即,词)转换为拼音或者笔画,也即处理后的第一地址可以包含n组拼音或者n组笔画,然后确定第一地址包含的n组拼音与第二地址包含的m组拼音的第一关联度值,或者确定第一地址包含的n组笔画与第二地址包含的m组笔画的第一关联度值,且确定方法与确定第一地址包含的n个单词与第二地址包含的m个单词的第一关联度值类似,本申请对此不作赘述。

[0030] 其中,步骤120具体可以包括:

[0031] 对所述n个语素中的每个语素,确定所述每个语素与所述m个语素中各个语素的第一关联度值,以获得n×m个第一关联度值。

[0032] 步骤120也可以描述为如下步骤:

[0033] 1)按顺序取得n个语素中的第i个语素,其中,i=1,2,…,n;

[0034] 2)按顺序取得m个语素中的第j个语素,其中,j=1,2,…,m;

[0035] 3)计算第i个语素与第j个语素的第一关联度值;

[0036] 4)遍历所有的n个语素以及所有的m个语素,获得n×m个第一关联度值。

[0037] 以第一地址和第二地址为英文地址来说,即第一地址包含n个单词,第二地址包含m个单词,假设n为3,第一地址包含的3个单词分别为:X、Y和Z,且假设m为4,第二地址包含的4个单词分别为:A、B、C和D,则对于X,分别确定其与A、B、C和D四个语素的第一关联度值;对于Y,分别确定其与A、B、C和D四个语素的第一关联度值;对于Z,分别确定其与A、B、C和D四个语素的第一关联度值,最后获得3×4个第一关联度值。当X与A的第一关联度值可以表示为d(A,X)时,获得的3×4个第一关联度值可以如表1所示。

[0038] 表1

[0039]

	A	B	C	D
X	d(A,X)	d(B,X)	d(C,X)	d(D,X)
Y	d(A,Y)	d(B,Y)	d(C,Y)	d(D,Y)
Z	d(A,Z)	d(B,Z)	d(C,Z)	d(D,Z)

[0040] 步骤130,根据所述第一关联度值和预设的算法,获得所述第一地址与所述第二地址的第二关联度值。

[0041] 此处,预设的算法可以包括匈牙利算法或者穷举法等。在此说明书中,以预设的算法为匈牙利算法为例。

[0042] 步骤130具体可以包括:

[0043] 步骤A:根据所述n×m个第一关联度值和预设的算法,从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素,并记录所述第一语素与所述第二语素的目标关联度值;直至记录n个目标关联度值。

[0044] 此处,第一语素是第一地址中的任一语素,可以理解的是,当第一地址中包含n个语素时,则第一语素的个数为n,与第一语素匹配的第二语素的个数为n,因此,记录的目标关联度值的个数也为n个。

[0045] 其中,步骤A中根据所述n×m个第一关联度值和预设的算法,从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素进一步包括:

[0046] 根据所述n×m个第一关联度值,构建n×m的矩阵;

[0047] 根据预设的算法,对所述n×m的矩阵进行预处理;

[0048] 根据预处理后的n×m的矩阵,从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素。

[0049] 需要说明的是,当第一关联度值为编辑距离值时,则上述为第一语素选取最匹配的第二语素的过程即为求解编辑距离值的最优匹配问题,也即确保第一地址中的每个单词在第二地址中找到对应的单词,并且总体的差异度值最小。

[0050] 在求解编辑距离值的最优匹配问题时,预处理可以包括将矩阵中每行元素减去该行最小的元素和/或将每列元素减去该列最小的元素等。

[0051] 具体地,将第i个语素与第j个语素的第一关联度值作为矩阵第i行第j列的元素,则如前述例子中的3×4个第一关联度值构建的矩阵如下所示:

[0052] $$\begin{bmatrix} 3 & 5 & 1 & 2 \\ 4 & 3 & 5 & 3 \\ 4 & 2 & 2 & 4 \end{bmatrix}$$

[0053] 且根据预设的算法,对上述矩阵做进行预处理后,得到如下矩阵:

[0054] $$\begin{bmatrix} 1 & 4 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 2 \end{bmatrix}$$

[0055] 需要说明的是,上述对矩阵的预处理过程属于现有技术,在此不复赘述。此外,在最后得到的矩阵中,还对每行独立的零元素进行了标记,也即第1行第3列的0进行了标记,对第2行第1列的0进行了标记,并对第3行2列的0进行了标记,以便获得两个地址中语素间的最优匹配组合。根据上述预处理后的矩阵以及标记的独立零元素,可以将表1更新为表2。

[0056] 表2

[0057]

	A	B	C	D
X	1	4	0	1
Y	0	0	2	0

Z	1	0	0	2

[0058] 其中，用下划线标记的0即为预处理后的矩阵中标记出的独立零元素。从表2可以看出，当第一语素为X时，则与第一语素最匹配的第二语素为C，并记录X与C的目标关联度值d(C,X)；当第一语素为Y时，则与第一语素最匹配的第二语素为A，并记录Y与A的目标关联度值d(A,Y)；当第一语素为Z时，则与第一语素最匹配的第二语素为B，并记录Z与B的目标关联度值d(B,Z)，即可以记录3个目标关联度值。

[0059] 步骤B：根据所述n个目标关联度值，获得所述第一地址与所述第二地址的第二关联度值。

[0060] 在一个例子中，可以对n个目标关联度值求和，将n个目标关联度值之和作为述第一地址与第二地址的第二关联度值。如前述例子中，第一地址与第二地址的第二关联度值＝d(C,X)+d(A,Y)+d(B,Z)＝1+4+2＝7，此处的1，4和7是根据表1读取的。

[0061] 步骤140，根据所述第二关联度值，对所述第一地址与所述第二地址的关联性进行分析。

[0062] 可以理解的是，第二关联度值也可以包括差异度值和相似度值。

[0063] 需要说明的是，当步骤130中确定的第二关联度值为差异度值时，则第一地址与第二地址的第二关联度值越大，则表示两者越不相似；而当步骤120中确定的第二关联度值为相似度值时，则第一地址与第二地址的相似度值越接近1，则两者越相似，第一地址与第二地址的相似度值越接近0，则两者越不相似。此外，当步骤130中确定的第二关联度值为相似度值时，则可以对相似度值求log后取负号转化为差异度值。

[0064] 此外，本申请的关联性的分析结果可以作为聚类分析、词频分析以及地址标准化等的基础。

[0065] 本申请提供的地址的分析方法，确定第一地址中n个语素与第二地址中m个语素的第一关联度值；根据所述第一关联度值和预设的算法，获得所述第一地址与所述第二地址的第二关联度值；根据所述第二关联度值，对所述第一地址与所述第二地址的关联性进行分析。由此，可以提高地址的分析的准确性和适用性。

[0066] 综上，在实物交易的电子商务交易中，对于涉及境外交易订单而言，英文书写的收货地址特别是非英语母语地区的收货地址，拼写上的细微差别，书写顺序上的习惯等各种现实情况给地址关联性分析带来了进一步的挑战，因此本申请的地址的分析方法是必要的。

[0067] 与上述地址的分析方法对应地，本申请实施例还提供的一种地址的分析装置，如图2所示，该装置包括：输入单元201、确定单元202、获取单元203和分析单元204。

[0068] 输入单元201，用于输入第一地址和第二地址，所述第一地址包含n个语素，所述第二地址包含m个语素，其中，所述语素是指所述地址中最小的语义单位，n和m均为自然数。

[0069] 确定单元202，用于确定所述n个语素与所述m个语素的第一关联度值。

[0070] 确定单元202具体用于：

[0071] 对所述n个语素中的每个语素，确定所述每个语素与所述m个语素中各个语素的第一关联度值，以获得n×m个第一关联度值。

[0072] 其中，所述第一关联度值包括：编辑距离值、汉明距离值、杰卡德距离值、N-Gram距离值、JW距离值或者余弦距离值。

[0073] 获取单元203，用于根据确定单元202确定的所述第一关联度值和预设的算法，获得所述第一地址与所述第二地址的第二关联度值。

[0074] 获取单元203具体用于：

[0075] 根据所述n×m个第一关联度值和预设的算法，从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素，并记录所述第一语素与所述第二语素的目标关联度值；直至记录n个目标关联度值；

[0076] 根据所述n个目标关联度值，获得所述第一地址与所述第二地址的第二关联度值。

[0077] 其中，所述根据所述n×m个第一关联度值和预设的算法，从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素，包括：

[0078] 根据所述n×m个第一关联度值，构建n×m的矩阵；

[0079] 根据预设的算法，对所述n×m的矩阵进行预处理；

[0080] 根据预处理后的n×m的矩阵，从所述m个语素中选取与所述n个语素中第一语素最匹配的第二语素。

[0081] 分析单元204，用于根据获取单元203获得的所述第二关联度值，对所述第一地址与所述第二地址的关联性进行分析。

[0082] 本申请实施例提供的地址的分析装置，输入单元201输入第一地址和第二地址，所述第一地址包含n个语素，所述第二地址包含m个语素，其中，所述语素是指所述地址中最小的语义单位，n和m均为自然数；确定单元202确定所述n个语素与所述m个语素的第一关联度值；获取单元203根据确定的所述第一关联度值和预设的算法，获得所述第一地址与所述第二地址的第二关联度值；分析单元204根据所述第二关联度值，对所述第一地址与所述第二地址的关联性进行分析。由此，可以提高地址的分析的准确性和适用性。

[0083] 专业人员应该还可以进一步意识到，结合本文中所公开的实施例描述的各示例的对象及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[0084] 结合本文中所公开的实施例描述的方法或算法的步骤可以用硬件、处理器执行的软件模块，或者二者的结合来实施。软件模块可以置于随机存储器(RAM)、内存、只读存储器(ROM)、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

[0085] 以上所述的具体实施方式，对本申请的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本申请的具体实施方式而已，并不用于限定本申请的保护范围，凡在本申请的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

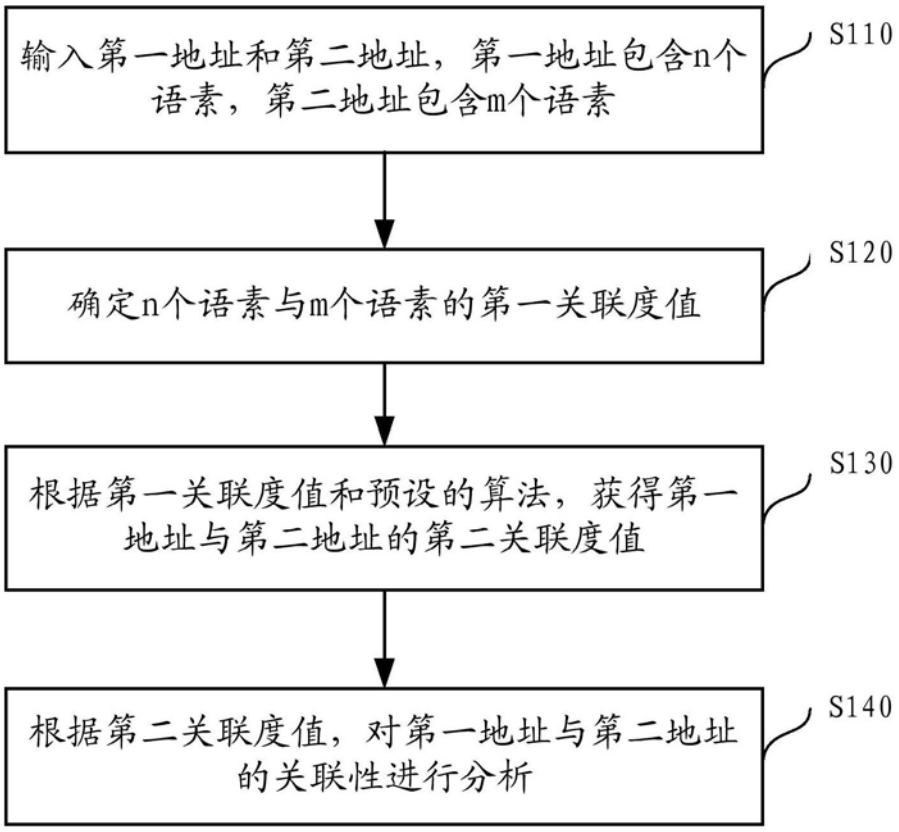

输入第一地址和第二地址，第一地址包含n个语素，第二地址包含m个语素
S110
确定n个语素与m个语素的第一关联度值
S120
根据第一关联度值和预设的算法，获得第一地址与第二地址的第二关联度值
S130
根据第二关联度值，对第一地址与第二地址的关联性进行分析
S140

图1

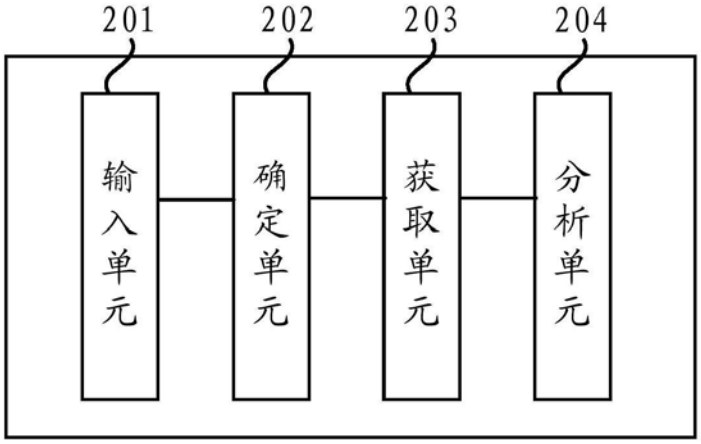

201
202
203
204
输入单元
确定单元
获取单元
分析单元

图2