(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102654881 A
(43)申请公布日 2012.09.05

(21)申请号 201110056065.4

(22)申请日 2011.03.03

(71)申请人 富士通株式会社
地址 日本神奈川县

(72)发明人 王新文 夏迎炬 孟遥 张姝
贾文杰 于浩

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 杜诚 李春晖

(51)Int.Cl.
G06F 17/30(2006.01)

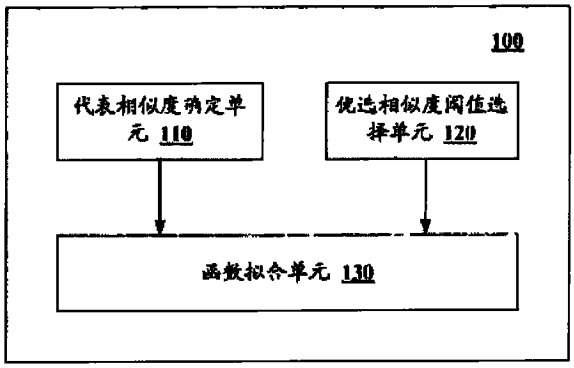
权利要求书 2 页 说明书 10 页 附图 4 页

(54)发明名称

用于名称消岐聚类的装置和方法

(57)摘要

提供了用于名称消岐聚类的装置和方法。对名称训练集进行数据处理的装置包括：代表相似度确定单元，用于确定名称训练集的代表相似度，该代表相似度为该名称训练集中的文本间相似度的代表值；优选相似度阈值选择单元，用于采用不同的相似度阈值对该名称训练集进行聚类以选择使聚类效果较佳的相似度阈值作为优选相似度阈值；以及函数拟合单元，用于根据至少两个名称训练集中的每个名称训练集的代表相似度和优选相似度阈值拟合表示代表相似度与优选相似度阈值之间对应关系的函数。



CN 102654881 A

1. 一种对名称训练集进行数据处理的装置，包括：

代表相似度确定单元，用于确定名称训练集的代表相似度，所述代表相似度为所述名称训练集中的文本间相似度的代表值；

优选相似度阈值选择单元，用于采用不同的相似度阈值对所述名称训练集进行聚类以选择使聚类效果较佳的相似度阈值作为优选相似度阈值；以及

函数拟合单元，用于根据至少两个名称训练集中的每个名称训练集的所述代表相似度和所述优选相似度阈值拟合表示所述代表相似度与所述优选相似度阈值之间对应关系的函数。

2. 根据权利要求 1 所述的装置，其中，所述代表相似度确定单元通过对所述名称训练集的文本间相似度进行加权平均来确定所述名称训练集的代表相似度。

3. 根据权利要求 1 或 2 所述的装置，其中，所述代表相似度确定单元包括：

相似度序列生成单元，用于计算所述名称训练集中所有文本相互之间的相似度，并对所计算的相似度进行排序以生成相似度序列；

相似度序列划分单元，用于将所述相似度序列划分成两个或更多个块；以及

代表相似度计算单元，用于对所述相似度序列中所划分的每个块的平均相似度进行加权平均以确定所述代表相似度。

4. 根据权利要求 3 所述的装置，其中，所述相似度序列划分单元识别所述相似度序列中相似度发生跳跃性变化的位置，并在所述跳跃性变化的位置划分所述相似度序列。

5. 根据权利要求 3 所述的装置，其中，所述相似度序列划分单元将所述相似度序列划分成三块，首尾两块的长度小于中间一块的长度。

6. 根据权利要求 3 所述的装置，其中，所述代表相似度计算单元从所划分的各个块中选择关键相似度块，在所述加权平均中对所述关键相似度块赋予较高权重。

7. 根据权利要求 6 所述的装置，其中，所述代表相似度计算单元根据下式选择所述关键相似度块：

$$k=\begin{cases}(n+1)/2 & n\text{为奇数}\\ n/2 & n\text{为偶数}\end{cases}$$

其中，n 为所述相似度序列中所划分的块的总数，k 为关键相似度块在 n 个块中的序号。

8. 一种对名称训练集进行数据处理的方法，包括步骤：

确定至少两个名称训练集中每个名称训练集的代表相似度，所述代表相似度为相应名称训练集中的文本间相似度的代表值；

针对所述至少两个名称训练集中的每个名称训练集，采用不同的相似度阈值进行聚类以选择使聚类效果较佳的相似度阈值作为优选相似度阈值；以及

根据所述至少两个名称训练集中每个名称训练集的所述代表相似度和所述优选相似度阈值，拟合表示所述代表相似度与所述优选相似度阈值之间对应关系的函数。

9. 一种用于名称消岐的聚类装置，包括：

代表相似度确定单元，用于确定所述待消岐名称集的代表相似度；

优选相似度阈值估算单元，用于根据表示代表相似度与使聚类效果较佳的优选相似度阈值之间对应关系的预定函数，确定与所确定的代表相似度对应的所述优选相似度阈值；

以及

聚类单元，用于利用所确定的优选相似度阈值对所述待消岐名称集进行聚类。

10. 一种用于名称消岐的聚类方法，包括步骤：

确定待消岐名称集的代表相似度；

根据表示代表相似度与使聚类效果较佳的优选相似度阈值之间对应关系的预定函数，确定与所确定的代表相似度对应的所述优选相似度阈值；以及

利用所确定的优选相似度阈值对所述待消岐名称集进行聚类。

用于名称消岐聚类的装置和方法

技术领域

[0001] 本发明涉及名称消岐聚类，尤其涉及利用动态阈值进行名称消岐聚类的装置和方法。

背景技术

[0002] 名称消岐是最近兴起的一个研究方向。名称消岐是针对由于同一名称（人名、地名、组织机构名等）被现实中的多个实体使用而产生的名称歧义现象而提出的，目前大多数的名称消岐方案采用文本聚类的方法。例如，当利用搜索引擎搜索某个名称时，作为搜索结果返回大量包含该名称的网页 $D = \{d_1, d_2, \cdots, d_n\}$，这些网页中的名称可能指向现实中的不同实体，聚类的目的是按照不同实体将这些网页构成的文本集合聚合为若干类 $C = \{c_1, c_2, \cdots, c_m\}$，其中每个类 c_i 对应于现实中的一个实体，从而达到消岐的目的。

[0003] 典型的聚类算法不同程度地涉及对阈值的选择，而阈值的选择直接影响聚类效果。然而，由于名称歧义的特殊性，传统的文本聚类方法在名称消岐上的聚类效果差强人意，导致这个问题的主要原因在于对于不同的名称的聚类最优截断阈值不同而导致聚类结果不理想。例如，针对某个常用人名的文本集合与针对某个生僻人名的文本集合具有不同的相似度特性，相应地，这两个文本集合在聚类过程中具有最佳聚类效果的相似度阈值也存在差异。因此，如果采用固定的阈值进行聚类，难以针对具有不同相似度特性的文本集合达到理想的聚类效果。

发明内容

[0004] 本发明的目的在于提供一种利用动态阈值进行名称消岐聚类的装置和方法，以至少部分地克服现有技术的上述缺陷。

[0005] 根据本发明的一个实施例，提供一种对名称训练集进行数据处理的装置，包括：代表相似度确定单元，用于确定名称训练集的代表相似度，该代表相似度为该名称训练集中的文本间相似度的代表值；优选相似度阈值选择单元，用于采用不同的相似度阈值对该名称训练集进行聚类以选择使聚类效果较佳的相似度阈值作为优选相似度阈值；以及函数拟合单元，用于根据至少两个名称训练集中的每个名称训练集的代表相似度和优选相似度阈值拟合表示代表相似度与优选相似度阈值之间对应关系的函数。

[0006] 根据本发明的另一个实施例，提供一种对名称训练集进行数据处理的方法，包括步骤：确定至少两个名称训练集中每个名称训练集的代表相似度，该代表相似度为相应名称训练集中的文本间相似度的代表值；针对至少两个名称训练集中的每个名称训练集，采用不同的相似度阈值进行聚类以选择使聚类效果较佳的相似度阈值作为优选相似度阈值；以及根据至少两个名称训练集中每个名称训练集的代表相似度和优选相似度阈值，拟合表示代表相似度与优选相似度阈值之间对应关系的函数。

[0007] 根据本发明的再一个实施例，提供一种用于名称消岐的聚类装置，包括：代表相似度确定单元，用于确定待消岐名称集的代表相似度；优选相似度阈值估算单元，用于根据表

示代表相似度与使聚类效果较佳的优选相似度阈值之间对应关系的预定函数，确定与所确定的代表相似度对应的优选相似度阈值；以及聚类单元，用于利用所确定的优选相似度阈值对待消岐名称集进行聚类。

[0008] 根据本发明的又一个实施例，提供一种用于名称消岐的聚类方法，包括步骤：确定待消岐名称集的代表相似度；根据表示代表相似度与使聚类效果较佳的优选相似度阈值之间对应关系的预定函数，确定与所确定的代表相似度对应的优选相似度阈值；以及利用所确定的优选相似度阈值对待消岐名称集进行聚类。

[0009] 根据本发明的实施例，通过对每个名称集的聚类阈值进行动态调整，能够减少由于使用固定阈值而造成的聚类效果偏差，可以针对不同的名称集给出优选的聚类阈值，从而提高聚类的自适应性，并且提高最终聚类效果。

附图说明

[0010] 参照下面结合附图对本发明实施例进行的说明，会更加容易地理解本发明的以上和其它目的、特点和优点。为了避免因不必要的细节而模糊了本发明，在附图中仅仅示出了与根据本发明的方案密切相关的装置结构和 / 或处理步骤，而省略了与本发明关系不大的其它细节。

[0011] 图 1 是示出根据本发明实施例的用于对名称训练集进行数据处理的装置 100 的配置的框图；

[0012] 图 2 是示出图 1 所示的代表相似度确定单元 110 的配置的框图；

[0013] 图 3 是示出根据本发明实施例的对名称训练集进行数据处理的方法的流程图；

[0014] 图 4 是示出图 3 所示的确定名称训练集的代表相似度的步骤 S310 中的过程的流程图；

[0015] 图 5 是示出根据本发明实施例的用于名称消岐的聚类装置的配置的框图；

[0016] 图 6 是示出根据本发明实施例的用于名称消岐的聚类方法的流程图；

[0017] 图 7 是示出其中实现本发明的装置和方法的计算机的示例性结构的框图。

具体实施方式

[0018] 下面参照附图说明本发明的实施例。应当注意，为了清楚的目的，附图和说明中省略了与本发明无关的、本领域普通技术人员已知的部件和处理的表示和描述。

[0019] 图 1 是示出根据本发明实施例的对用于名称训练集进行数据处理的装置的配置的框图。

[0020] 如图 1 所示，对名称训练集进行数据处理的装置 100 包括代表相似度确定单元 110、优选相似度阈值选择单元 120 以及函数拟合单元 130。

[0021] 作为装置 100 的处理对象的名称训练集中的每个名称训练集包括针对同一名称的多个文本，并且该多个文本的聚类关系已知。

[0022] 例如，当文本是网页时，文本可以经过网页预处理、特征向量提取、约束条件建立等处理。在网页预处理中，可以对网页进行内容提取、有效 url 提取、标题提取等操作，并可以将网页用 xml 文件格式保存。在特征向量提取过程中，根据网页的标题和内容建立特征向量组，并对每个特征向量赋予权重，从而得到文本的向量表示。例如，文本中每个特征向

量的权重可以通过 TF（词汇频率）方法确定。另外，标题中的特征的权重值可以被设置为高于网页内容中的特征的权重值。

[0023] 可选地，可以根据特殊的特征对文本建立约束条件。特殊的特征例如包括链接互指、所属单位命名实体、住址命名实体、电子邮件命名实体等。根据这些特征约束某些文本不可以被聚合为一类，某些文本应该被聚合为一类。例如，当涉及同一人名的两个网页上包含相同的电子邮件地址时，基本上可以确定该两个网页针对现实中的同一人，因此可以将约束条件建立为将该两个网页聚合为同一类。

[0024] 作为处理对象的名称训练集中的文本被提供给代表相似度确定单元 110 和优选相似度阈值选择单元 120。代表相似度确定单元 110 用于确定每个名称训练集的代表相似度，名称训练集的代表相似度是指该名称训练集中的文本间的相似度的代表值。可以采用不同的预定规则确定名称训练集的代表相似度。例如，代表相似度确定单元 110 可以通过对名称训练集中的文本相互之间的相似度进行加权平均来确定该名称训练集的代表相似度。

[0025] 图 2 示出了根据本发明的一个实施例的代表相似度确定单元的配置。代表相似度确定单元 110 包括相似度序列生成单元 210、相似度序列划分单元 220 和代表相似度计算单元 230。

[0026] 相似度序列生成单元 210 例如采用欧式距离公式、余弦距离公式等常用相似度计算方法计算一个名称训练集中所有文本相互之间的相似度，并对所计算的相似度进行排序以生成相似度序列。例如，当某个名称训练集共包含 N 个文本时，相似度序列生成单元 210 总共计算C_N^2个相似度值，并将这C_N^2个相似度值进行排序以生成相似度序列。

[0027] 相似度序列生成单元 210 将所生成的相似度序列提供给相似度序列划分单元 220，相似度序列划分单元 220 用于将相似度序列划分成 n 块。

[0028] 在本发明的一个实施例中，相似度序列划分单元 220 识别相似度序列中相似度发生跳跃性变化的位置，并在发生跳跃性变化的位置划分相似度序列。相似度序列中的跳跃性变化位置例如可能对应于同类文本间相似度与非同类文本间相似度的分界、距离较近的不同类别中的文本间相似度与距离较远的不同类别中的文本间相似度的分界等等，因此根据相似度跳跃性变化的位置划分出的各个块分别包含对于不同类型的文本间相似度具有代表性的相似度值。

[0029] 或者，相似度序列划分单元 220 可以根据预定比例将相似度序列划分成 n 块。根据本发明的一个实施例，相似度序列划分单元 220 将相似度序列划分成 3 块，中间一块的长度较大，首尾两块的长度较小。例如，中间一块占相似度序列的 50％，首位两块各占相似度序列的 25％。

[0030] 相似度序列划分单元 220 将经划分的相似度序列提供给代表相似度计算单元 230，代表相似度计算单元 230 可以根据相似度序列的每个块中的相似度值计算该名称训练集的代表相似度。例如，代表相似度计算单元 230 可以对所划分的每个块中的相似度值求平均以得到 n 个平均相似度值，并对这 n 个平均相似度进一步进行加权平均来计算该名称训练集的代表相似度。

[0031] 根据本发明的一个具体实施例，代表相似度确定单元 230 根据下式计算名称训练集的代表相似度。

[0032] $k = [n*coef]$ （等式 1）

[0033] $M = \sum_{i=1}^{n}(n - |k - i|)$（等式 2）

[0034] $Sim = \frac{1}{M}\sum_{i=1}^{n}(n - |k - i|) * Si$（等式 3）

[0035] 其中，n 表示相似度序列中所划分的块的总数；k 表示关键相似度块的序号；coef 表示关键位置百分比；方括号表示四舍五入取整；Si 表示第 i 个块中的相似度的均值；Sim 表示名称训练集的代表相似度。

[0036] 根据等式 1，通过对相似度块的总数 n 和关键位值百分比 coef 的乘积取整来确定关键块序号 k，可以将 coef 设置为不同值以选择相似度序列的各个块中不同位置的块作为关键相似度块。然后，根据等式 2 和等式 3，通过对各相似度块的平均相似度进 Si 行加权平均来确定该名称集的代表相似度。在等式 3 所表示的加权平均中，第 i 个块的权重 $w_i = (n-|k-i|)$，关键相似度块（第 k 块）被赋予最高的权重，即 $w_k = n$，而距离关键相似度块越远的块被赋予的权重越低。另外，可以按照类似规则采用其它具体公式确定每个块的权重。

[0037] 根据本发明的一个具体实施例，关键位置百分比的取值为 coef = 50%。根据等式 1，关键相似度块序号 k = [n/2]，即，当 n 为奇数时，k = (n+1)/2，当 n 为偶数时，k = n/2。也就是，选取相似度序列的各块中位于中间的块作为关键相似度块。

[0038] 以上作为示例描述了代表相似度确定单元 110 的配置。然而，代表相似度确定单元 110 也可以具有其它配置。例如，除了相似度序列生成单元 210 和相似度序列划分单元 220 之外，或代替代相似度序列生成单元 210 和相似度序列划分单元 220，代表相似度确定单元 110 可以包括相似度抽样单元（未示出）。相似度抽样单元可以从所处理的名称训练集的全部文本中抽取一定比例（例如 30%）的文本并计算所抽取的文本间的相似度。代表相似度计算单元 230 例如可以通过对相似度抽样单元提供的相似度进行加权平均来确定代表相似度。

[0039] 以这种方式，代表相似度确定单元 110 可以确定的一组名称训练集中的每个名称训练集的代表相似度 Sim，并将其提供给函数拟合单元 130。

[0040] 优选相似度阈值选择单元 120 采用不同的相似度阈值对名称训练集进行聚类（例如采用层次聚类 (HAC)，在聚类过程中，只有当两个簇 (cluster) 的相似度超过相似度阈值时才允许将该两个簇合并），并且通过将使用不同阈值得到的聚类效果与该名称训练集的已知聚类关系进行比较来确定聚类效果较佳的优选相似度阈值。例如，优选相似度阈值选择单元 120 可以在 0 至 0.3（余弦相似度）的范围内选取不同的相似度阈值对名称训练集进行聚类，并选择其中使聚类效果较好的相似度阈值作为优选相似度阈值。

[0041] 例如，优选相似度阈值选择单元 120 可以将使用某一相似度阈值得到的聚类结果与该训练集的已知聚类关系进行比较，根据准确率、召回率等常用指标评估聚类效果。

[0042] 以这种方式，优选相似度阈值选择单元 120 可以确定一组名称训练集中每个名称训练集的优选相似度阈值 Sim_op，并将其提供给函数拟合单元 130。

[0043] 函数拟合单元 130 根据代表相似度确定单元 110 提供的代表相似度 Sim 和优选相似度阈值选择单元 120 提供的优选相似度阈值 Sim_op 可以得到针对单个名称训练集的代表相似度和优选相似度阈值的对 <Sim，Sim_op>，对于一组名称训练集，可以得到代表相似度和优选相似度阈值的对的集合 $S_{<Sim,Sim\_op>}$。函数拟合单元 130 根据集合 $S_{<Sim,Sim\_op>}$ 拟合表

示代表相似度 Sim 与优选相似度阈值 Sim_op 之间的对应关系的函数 Sim_op = f(Sim)，该函数可以是线性函数、二次函数等。可以根据函数拟合的需要确定所要处理的名称训练集的数量。在采用最小二乘法拟合代表相似度与优选相似度阈值间的二次函数的情况下，例如可以对 100 个以上的名称训练集进行数据处理以进行函数拟合。

[0044] 图 3 是示出根据本发明实施例的对名称训练集进行数据处理的方法的流程图。

[0045] 在步骤 S310，确定一组名称训练集中一个名称训练集的代表相似度；

[0046] 在步骤 S320，采用不同相似度阈值对该名称训练集进行聚类，并选择使聚类效果较佳的相似度阈值作为优选相似度阈值。需要指出的是，虽然图 3 中示例性地示出步骤 S310 在步骤 S320 之前，但步骤 S310 和步骤 S320 之间不存在先后顺序的限制，可以并行地执行或者以任意顺序相继执行步骤 S310 和步骤 S320；

[0047] 当对该组名称训练集中的每个名称训练集完成步骤 S310 和 S320 时，在步骤 S330，根据通过步骤 S310 得到的每个名称训练集的代表相似度和通过步骤 S320 得到的每个名称训练集的优选相似度阈值，拟合表示代表相似度与优选相似度阈值之间的对应关系的函数。

[0048] 根据本发明的一个实施例，步骤 S310 包括图 4 所示的过程，其中：

[0049] 在步骤 S410，计算该名称训练集中所有文本相互之间的相似度，并对所计算的相似度进行排序以生成相似度序列；

[0050] 在步骤 S420，将相似度序列划分成 n 块；

[0051] 在步骤 S430，对相似度序列的所划分的每个块的平均相似度进行加权平均以确定该名称训练集的代表相似度。

[0052] 根据本发明的一个具体实施例，在步骤 S430 中采用等式 1-3 计算代表相似度。

[0053] 图 5 是示出根据本发明实施例的用于名称消岐的聚类装置的配置的框图。用于名称消岐的聚类装置 500 包括代表相似度确定单元 510、优选相似度阈值估算单元 520 以及聚类单元 530。

[0054] 代表相似度确定单元 510 确定待消岐名称集的代表相似度。根据本发明的一个实施例，代表相似度确定单元 510 具有与图 2 所示的代表相似度确定单元 110 类似的配置，在此省略对该具体配置的描述。代表相似度确定单元 510 将待消岐名称集的代表相似度提供给优选相似度阈值估算单元 520。

[0055] 优选相似度阈值估算单元 520 根据表示名称集的代表相似度与使聚类效果较佳的优选相似度阈值之间的对应关系的预定函数，确定与代表相似度确定单元提供的待消岐名称集的代表相似度相对应的优选相似度阈值估算优选相似度阈值，并将所估算的优选相似度阈值提供给聚类单元 530。例如，该预定函数可以是通过利用图 1 所示的装置 100 或利用图 3 所示的方法对名称训练集进行数据处理而得到的函数。

[0056] 聚类单元 530 利用由优选相似度阈值估算单元 520 估算的优选相似度阈值对待消岐名称集进行聚类。

[0057] 根据本发明的一个实施例，聚类单元 530 采用层次聚类方法对名称集进行聚类，在聚类过程中，只有当两个簇的相似度超过由优选相似度阈值估算单元 520 估算的优选相似度阈值时才允许将该两个簇合并。

[0058] 图 6 是示出根据本发明实施例的用于名称消岐的聚类方法的流程图。

[0059] 在步骤 S610，确定待消岐名称集的代表相似度；

[0060] 在步骤 S620，根据表示代表相似度与使聚类效果较佳的优选相似度阈值之间的对应关系的预定函数来估算优选相似度阈值；

[0061] 在步骤 S630，利用步骤 S620 中估算的优选相似度阈值对待消岐名称集进行聚类。

[0062] 根据本发明的一个实施例，确定代表相似度的步骤 S610 具有与图 4 所示的过程类似的过程，在此省略对该过程的具体描述。

[0063] 步骤 S620 中所使用的预定函数例如可以是通过利用图 1 所示的装置 100 或利用图 3 所示的方法对名称训练集进行数据处理而得到的函数。

[0064] 根据本发明的一个实施例，步骤 S630 采用层次聚类方法对名称集进行聚类，在聚类过程中，只有当两个簇的相似度超过在步骤 S620 中估算的优选相似度阈值时才允许将该两个簇合并。

[0065] 所属技术领域的技术人员知道，本发明可以体现为装置、方法或计算机程序产品。因此，本发明可以具体实现为以下形式，即，可以是完全的硬件、完全的软件（包括固件、驻留软件、微代码等）、或者软件部分与硬件部分的组合。此外，本发明还可以采取体现在任何有形的表达介质中的计算机程序产品的形式，该介质中包含计算机可用的程序码。

[0066] 可以使用一个或多个计算机可读介质的任何组合。计算机可读介质可以是计算机可读信号介质或计算机可读存储介质，计算机可读存储介质例如可以是—但不限于—电的、磁的、光的、电磁的、红外线的、或半导体的系统、装置、器件或传播介质、或前述各项的任何适当的组合。计算机可读存储介质的更具体的例子（非穷举的列表）包括：有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器 (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、或前述各项的任何适当的组合。在本文语境中，计算机可读存储介质可以是任何含有或存储供指令执行系统、装置或器件使用的或与指令执行系统、装置或器件相联系的程序的有形介质。

[0067] 用于执行本发明的操作的计算机程序码，可以以一种或多种程序设计语言的任何组合来编写，所述程序设计语言包括面向对象的程序设计语言—诸如 Java、Smalltalk、C++ 之类，还包括常规的过程式程序设计语言—诸如”C”程序设计语言或类似的程序设计语言。程序码可以完全地在用户的计算机上执行、部分地在用户的计算机上执行、作为一个独立的软件包执行、部分在用户的计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在后一种情形中，远程计算机可以通过任何种类的网络—包括局域网 (LAN) 或广域网 (WAN)—连接到用户的计算机，或者，可以（例如利用因特网服务提供商来通过因特网）连接到外部计算机。

[0068] 图 7 是示出其中实现本发明的设备和方法的计算机的示例性结构的框图。

[0069] 在图 7 中，中央处理单元 (CPU)701 根据只读存储器 (ROM)702 中存储的程序或从存储部分 708 加载到随机存取存储器 (RAM)703 的程序执行各种处理。在 RAM 703 中，也根据需要存储当 CPU 701 执行各种处理等等时所需的数据。

[0070] CPU 701、ROM 702 和 RAM 703 经由总线 704 彼此连接。输入 / 输出接口 705 也连接到总线 704。

[0071] 下述部件连接到输入 / 输出接口 705：输入部分 706，包括键盘、鼠标等等；输出部

分 707，包括显示器，比如阴极射线管 (CRT)、液晶显示器 (LCD) 等等，和扬声器等等；存储部分 708，包括硬盘等等；和通信部分 709，包括网络接口卡比如 LAN 卡、调制解调器等等。通信部分 709 经由网络比如因特网执行通信处理。

[0072] 根据需要，驱动器 710 也连接到输入 / 输出接口 705。可拆卸介质 711 比如磁盘、光盘、磁光盘、半导体存储器等等根据需要被安装在驱动器 710 上，使得从中读出的计算机程序根据需要被安装到存储部分 708 中。

[0073] 在通过软件实现上述步骤和处理的情况下，从网络比如因特网或存储介质比如可拆卸介质 711 安装构成软件的程序。

[0074] 本领域的技术人员应当理解，这种存储介质不局限于图 7 所示的其中存储有程序、与方法相分离地分发以向用户提供程序的可拆卸介质 711。可拆卸介质 711 的例子包含磁盘、光盘 (包含光盘只读存储器 (CD-ROM) 和数字通用盘 (DVD))、磁光盘 (包含迷你盘 (MD)) 和半导体存储器。或者，存储介质可以是 ROM 702、存储部分 708 中包含的硬盘等等，其中存有程序，并且与包含它们的方法一起被分发给用户。

[0075] 权利要求中的对应结构、操作以及所有功能性限定的装置或步骤的等同替换，旨在包括任何用于与在权利要求中具体指出的其它单元相组合地执行该功能的结构或操作。所给出的对本发明的描述其目的在于示意和描述，并非是穷尽性的，也并非是要把本发明限定到所表述的形式。对于所属技术领域的普通技术人员来说，在不偏离本发明范围和精神的情况下，显然可以作出许多修改和变型。对实施例的选择和说明，是为了最好地解释本发明的原理和实际应用，使所属技术领域的普通技术人员能够明了，本发明可以有适合所要的特定用途的具有各种改变的各种实施方式。

[0076] 附记

[0077] 附记 1. 一种对名称训练集进行数据处理的装置，包括：

[0078] 代表相似度确定单元，用于确定名称训练集的代表相似度，所述代表相似度为所述名称训练集中的文本间相似度的代表值；

[0079] 优选相似度阈值选择单元，用于采用不同的相似度阈值对所述名称训练集进行聚类以选择使聚类效果较佳的相似度阈值作为优选相似度阈值；以及

[0080] 函数拟合单元，用于根据至少两个名称训练集中的每个名称训练集的所述代表相似度和所述优选相似度阈值拟合表示所述代表相似度与所述优选相似度阈值之间对应关系的函数。

[0081] 附记 2. 根据附记 1 所述的装置，其中，所述代表相似度确定单元通过对所述名称训练集的文本间相似度进行加权平均来确定所述名称训练集的代表相似度。

[0082] 附记 3. 根据附记 1 或 2 所述的装置，其中，所述代表相似度确定单元包括：

[0083] 相似度序列生成单元，用于计算所述名称训练集中所有文本相互之间的相似度，并对所计算的相似度进行排序以生成相似度序列；

[0084] 相似度序列划分单元，用于将所述相似度序列划分成两个或更多个块；以及

[0085] 代表相似度计算单元，用于对所述相似度序列中所划分的每个块的平均相似度进行加权平均以确定所述代表相似度。

[0086] 附记 4. 根据附记 3 所述的装置，其中，所述相似度序列划分单元识别所述相似度序列中相似度发生跳跃性变化的位置，并在所述跳跃性变化的位置划分所述相似度序列。

[0087] 附记 5. 根据附记 3 所述的装置，其中，所述相似度序列划分单元将所述相似度序列划分成三块，首尾两块的长度小于中间一块的长度。

[0088] 附记 6. 根据附记 3 所述的装置，其中，所述代表相似度计算单元从所划分的各个块中选择关键相似度块，在所述加权平均中对所述关键相似度块赋予较高权重。

[0089] 附记 7. 根据附记 6 所述的装置，其中，所述代表相似度计算单元根据下式选择所述关键相似度块：

[0090]

$$k=\begin{cases}(n+1)/2 & n\text{为奇数}\\ n/2 & n\text{为偶数}\end{cases}$$

[0091] 其中，n 为所述相似度序列中所划分的块的总数，k 为关键相似度块在 n 个块中的序号。

[0092] 附记 8. 一种对名称训练集进行数据处理的方法，包括步骤：

[0093] 确定至少两个名称训练集中每个名称训练集的代表相似度，所述代表相似度为相应名称训练集中的文本间相似度的代表值；

[0094] 针对所述至少两个名称训练集中的每个名称训练集，采用不同的相似度阈值进行聚类以选择使聚类效果较佳的相似度阈值作为优选相似度阈值；以及

[0095] 根据所述至少两个名称训练集中每个名称训练集的所述代表相似度和所述优选相似度阈值，拟合表示所述代表相似度与所述优选相似度阈值之间对应关系的函数。

[0096] 附记 9. 根据附记 8 所述的方法，其中，通过对所述名称训练集的文本间相似度进行加权平均来确定所述名称训练集的代表相似度。

[0097] 附记 10. 根据附记 8 或 9 所述的方法，其中，所述确定至少两个名称训练集中每个名称训练集的代表相似度的步骤包括：

[0098] 计算所述名称训练集中所有文本相互之间的相似度，并对所计算的相似度进行排序以生成相似度序列；

[0099] 将所述相似度序列划分成两个或更多个块；以及

[0100] 对所述相似度序列中所划分的每个块的平均相似度进行加权平均以确定所述代表相似度。

[0101] 附记 11. 根据附记 10 所述的方法，其中，将所述相似度序列划分成两个或更多个块的步骤包括：识别所述相似度序列中相似度发生跳跃性变化的位置，并在所述跳跃性变化的位置划分所述相似度序列。

[0102] 附记 12. 根据附记 10 所述的方法，其中，将所述相似度序列划分成两个或更多个块的步骤包括：将所述相似度序列划分成三块，首尾两块的长度小于中间一块的长度。

[0103] 附记 13. 根据附记 10 所述的方法，其中，在对所述相似度序列中所划分的每个块的平均相似度进行加权平均以确定所述代表相似度的步骤中，从所划分的各个块中选择关键相似度块，在所述加权平均中对所述关键相似度块赋予较高权重。

[0104] 附记 14. 根据附记 13 所述的方法，其中，根据下式从所划分的各个块中选择关键相似度块：

[0105]

$$k=\begin{cases}(n+1)/2 & n\text{为奇数}\\ n/2 & n\text{为偶数}\end{cases}$$

[0106] 其中，n 为所述相似度序列中所划分的块的总数，k 为关键相似度块在 n 个块中的序号。

[0107] 附记 15. 一种用于名称消岐的聚类装置，包括：

[0108] 代表相似度确定单元，用于确定所述待消岐名称集的代表相似度；

[0109] 优选相似度阈值估算单元，用于根据表示代表相似度与使聚类效果较佳的优选相似度阈值之间对应关系的预定函数，确定与所确定的代表相似度对应的所述优选相似度阈值；以及

[0110] 聚类单元，用于利用所确定的优选相似度阈值对所述待消岐名称集进行聚类。

[0111] 附记 16. 根据附记 15 所述的装置，其中，所述代表相似度确定单元通过对所述名称训练集的文本间相似度进行加权平均来确定所述名称训练集的代表相似度。

[0112] 附记 17. 根据附记 15 或 16 所述的装置，其中，所述代表相似度确定单元包括：

[0113] 相似度序列生成单元，用于计算所述名称训练集中所有文本相互之间的相似度，并对所计算的相似度进行排序以生成相似度序列；

[0114] 相似度序列划分单元，用于将所述相似度序列划分成两个或更多个块；以及

[0115] 代表相似度计算单元，用于对所述相似度序列中所划分的每个块的平均相似度进行加权平均以确定所述代表相似度。

[0116] 附记 18. 根据附记 17 所述的装置，其中，所述相似度序列划分单元识别所述相似度序列中相似度发生跳跃性变化的位置，并在所述跳跃性变化的位置划分所述相似度序列。

[0117] 附记 19. 根据附记 17 所述的装置，其中，所述相似度序列划分单元将所述相似度序列划分成三块，首尾两块的长度小于中间一块的长度。

[0118] 附记 20. 根据附记 17 所述的装置，其中，所述代表相似度计算单元从所划分的各个块中选择关键相似度块，在所述加权平均中对所述关键相似度块赋予较高权重。

[0119] 附记 21. 根据附记 20 所述的装置，其中，所述代表相似度计算单元根据下式选择所述关键相似度块：

[0120]

$$k=\begin{cases}(n+1)/2 & n\text{为奇数}\\ n/2 & n\text{为偶数}\end{cases}$$

[0121] 其中，n 为所述相似度序列中所划分的块的总数，k 为关键相似度块在 n 个块中的序号。

[0122] 附记 22. 一种用于名称消岐的聚类方法，包括步骤：

[0123] 确定待消岐名称集的代表相似度；

[0124] 根据表示代表相似度与使聚类效果较佳的优选相似度阈值之间对应关系的预定函数，确定与所确定的代表相似度对应的所述优选相似度阈值；以及

[0125] 利用所确定的优选相似度阈值对所述待消岐名称集进行聚类。

[0126] 附记 23. 根据附记 22 所述的方法，其中，通过对所述名称训练集的文本间相似度

进行加权平均来确定所述名称训练集的代表相似度。

[0127] 附记 24. 根据附记 22 或 23 所述的方法，其中，所述确定至少两个名称训练集中每个名称训练集的代表相似度的步骤包括：

[0128] 计算所述名称训练集中所有文本相互之间的相似度，并对所计算的相似度进行排序以生成相似度序列；

[0129] 将所述相似度序列划分成两个或更多个块；以及

[0130] 对所述相似度序列中所划分的每个块的平均相似度进行加权平均以确定所述代表相似度。

[0131] 附记 25. 根据附记 24 所述的方法，其中，将所述相似度序列划分成两个或更多个块的步骤包括：识别所述相似度序列中相似度发生跳跃性变化的位置，并在所述跳跃性变化的位置划分所述相似度序列。

[0132] 附记 26. 根据附记 24 所述的方法，其中，将所述相似度序列划分成两个或更多个块的步骤包括：将所述相似度序列划分成三块，首尾两块的长度小于中间一块的长度。

[0133] 附记 27. 根据附记 24 所述的方法，其中，在对所述相似度序列中所划分的每个块的平均相似度进行加权平均以确定所述代表相似度的步骤中，从所划分的各个块中选择关键相似度块，在所述加权平均中对所述关键相似度块赋予较高权重。

[0134] 附记 28. 根据附记 27 所述的方法，其中，根据下式从所划分的各个块中选择关键相似度块：

[0135]

$$k=\begin{cases}(n+1)/2 & n\text{为奇数}\\ n/2 & n\text{为偶数}\end{cases}$$

[0136] 其中，n 为所述相似度序列中所划分的块的总数，k 为关键相似度块在 n 个块中的序号。

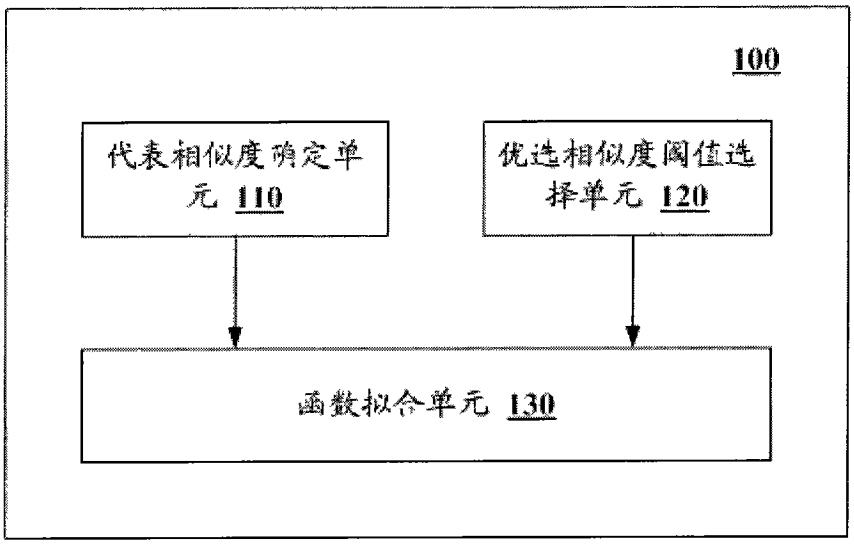

100
代表相似度确定单元 110
优选相似度阈值选择单元 120
函数拟合单元 130

图 1

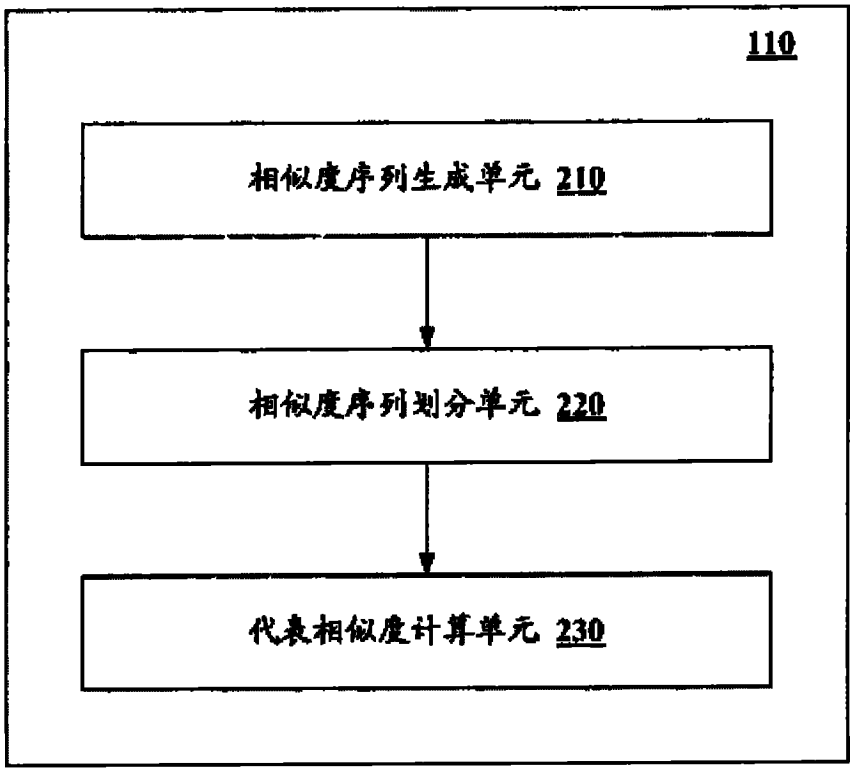

110
相似度序列生成单元 210
相似度序列划分单元 220
代表相似度计算单元 230

图 2

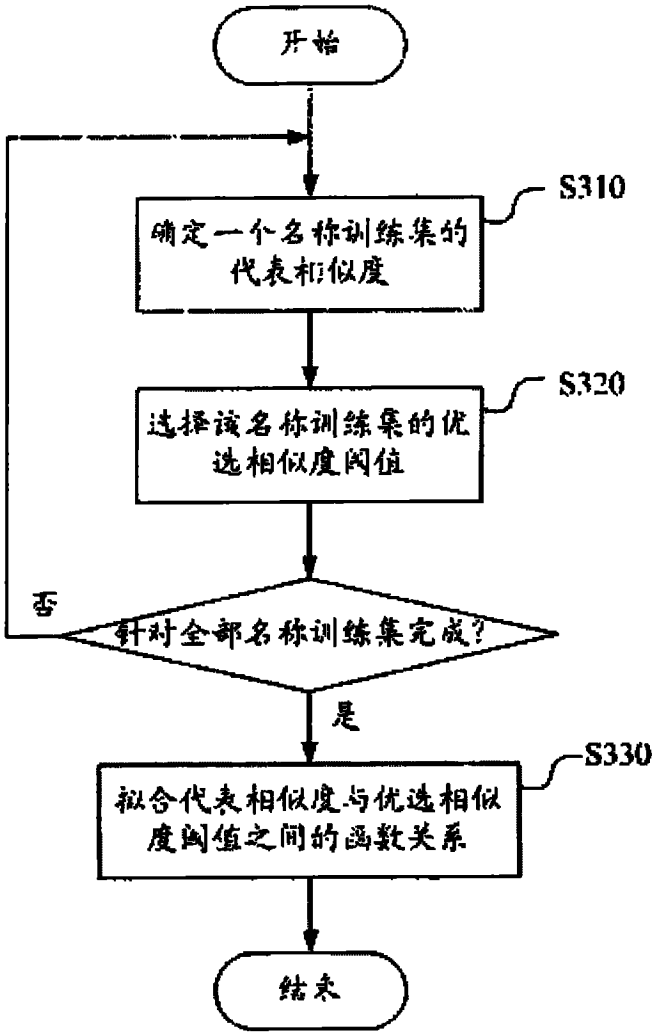

开始
确定一个名称训练集的代表相似度
S310
选择该名称训练集的优选相似度阈值
S320
否
针对全部名称训练集完成?
是
拟合代表相似度与优选相似度阈值之间的函数关系
S330
结束

图 3

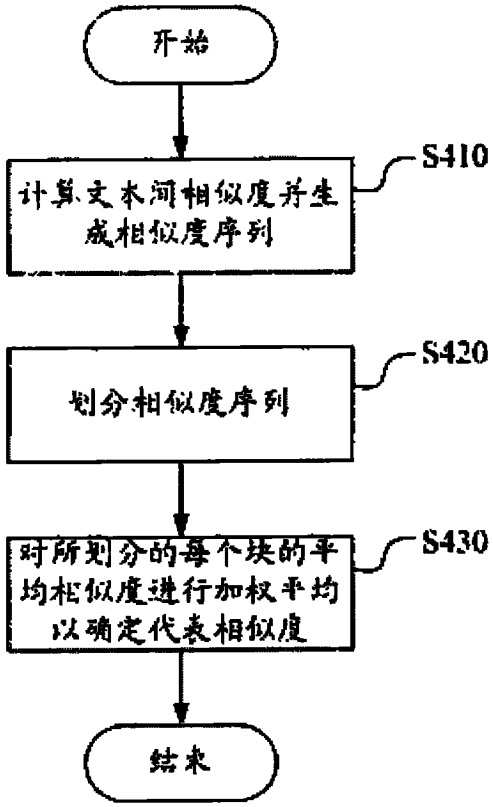

开始
计算文本间相似度并生成相似度序列
S410
划分相似度序列
S420
对所划分的每个块的平均相似度进行加权平均以确定代表相似度
S430
结束

图 4

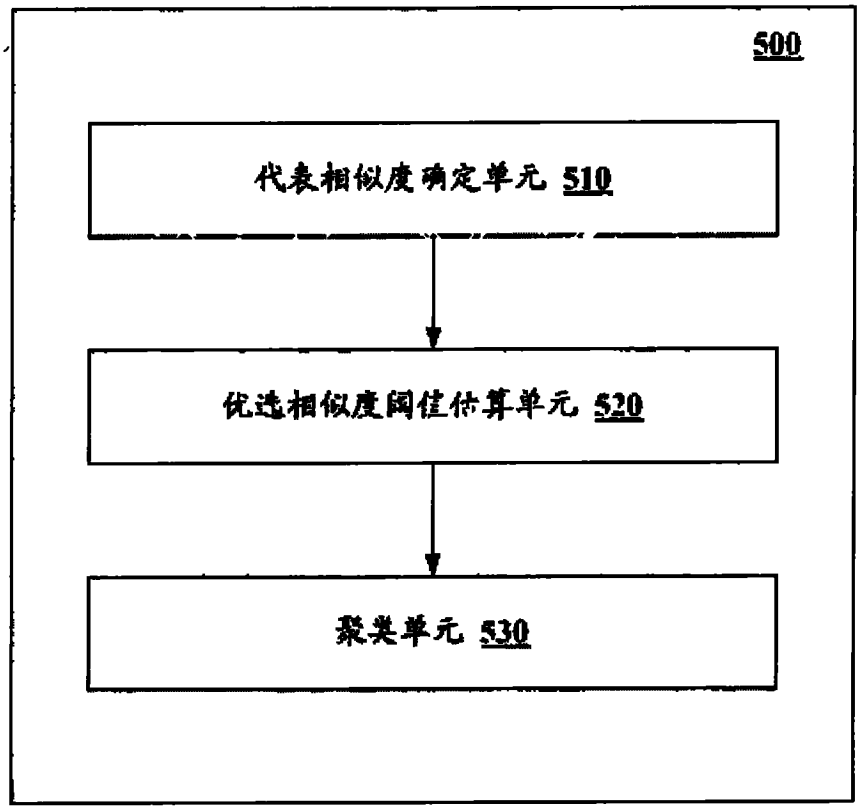


图 5

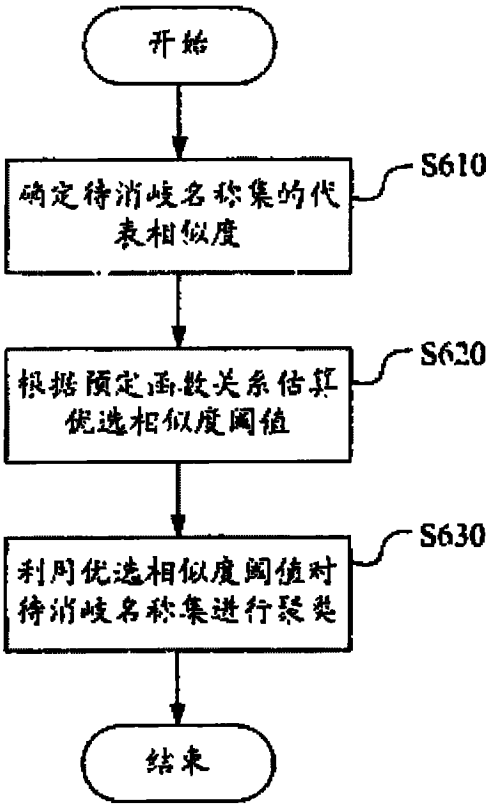

开始
确定待消岐名称集的代表相似度
S610
根据预定函数关系估算优选相似度阈值
S620
利用优选相似度阈值对待消岐名称集进行聚类
S630
结束

图 6

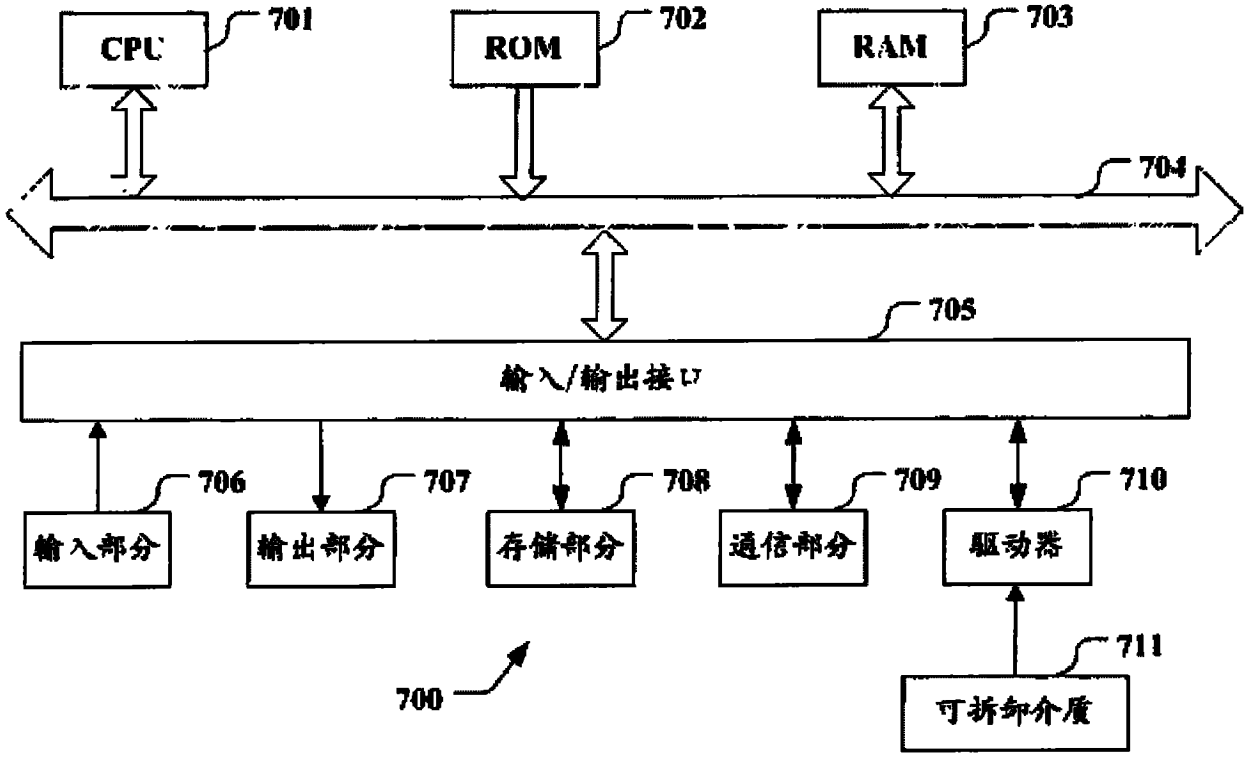

CPU
701
ROM
702
RAM
703
704
705
输入/输出接口
706
707
708
709
710
输入部分
输出部分
存储部分
通信部分
驱动器
711
可拆卸介质
700

图 7