(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113051919 B
(45)授权公告日 2023.04.04

(21)申请号 201911369966.1

(22)申请日 2019.12.26

(65)同一申请的已公布的文献号
申请公布号 CN 113051919 A

(43)申请公布日 2021.06.29

(73)专利权人 中国电信股份有限公司
地址 100033 北京市西城区金融大街31号

(72)发明人 高芷乔

(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 马景辉

(51)Int.Cl.
G06F 40/295(2020.01)

(56)对比文件
CN 109918680 A,2019.06.21
CN 109858040 A,2019.06.07
CN 110516247 A,2019.11.29
US 2017287474 A1,2017.10.05

审查员 王艳臣

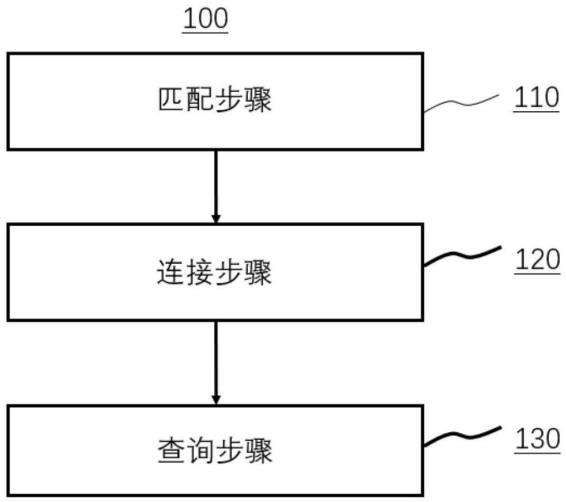
权利要求书2页 说明书9页 附图4页

(54)发明名称

一种识别命名实体的方法和装置

(57)摘要

本发明涉及一种识别命名实体的方法，包括：匹配步骤，基于命名实体中槽位的预定的优先级，按照优先级从高到低的顺序，将文本中的字段与各个槽位进行匹配；连接步骤，按照优先级和预定的逻辑关系将匹配到的槽位中的字段连接以得到连接结果；以及查询步骤，将所述连接结果在命名实体备选列表中进行查询，确定所述命名实体备选列表中与所述连接结果匹配的一个或多个命名实体，将所述一个或多个命名实体作为识别的命名实体。



CN 113051919 B

1.一种识别命名实体的方法,包括:

匹配步骤,基于命名实体中槽位的预定的优先级,按照优先级从高到低的顺序,将文本中的字段与各个槽位进行匹配;

连接步骤,按照优先级和预定的逻辑关系将匹配到的槽位中的字段连接以得到连接结果;以及

查询步骤,将所述连接结果在命名实体备选列表中进行查询,确定所述命名实体备选列表中与所述连接结果匹配的一个或多个命名实体,将所述一个或多个命名实体作为识别的命名实体;

相似度计算步骤,在所述命名实体备选列表中没有与所述连接结果匹配的命名实体的情况下,利用字符串相似度计算方法,计算所述连接结果与命名实体备选列表中的命名实体的相似度,并将相似度高于预定阈值的一个或多个命名实体作为相似的命名实体,并确定所述相似的命名实体中的最长公共子序列中的最短连续子串的长度大于第二阈值的命名实体作为识别的命名实体。

2.如权利要求1所述的方法,其中所述字符串相似度计算方法包括计算所述连接结果与命名实体备选列表中的命名实体的最长公共子序列长度,并且将所述最长公共子序列长度和连接结果的长度的比值作为所述相似度。

3.如权利要求2所述的方法,其中所述字符串相似度计算方法还包括利用动态规划矩阵,计算所述最长公共子序列长度。

4.如权利要求1所述的方法,其中:

根据动态规划矩阵,确定所述相似的命名实体中的最长公共子序列中的最短连续子串的长度大于第二阈值的命名实体作为识别的命名实体。

5.如权利要求1所述的方法,其中所述的预定的逻辑关系包括:具有相同优先级的槽位中的字段之间逻辑关系设为“或”,并且将不同优先级的槽位中的字段之间的逻辑关系设为“与”。

6.如权利要求1所述的方法,其中所述命名实体中槽位的预定的优先级是根据对命名实体备选列表中的所有命名实体的统计,将列表中的命名实体的字段归类为多个槽位,并对每个槽位赋予优先级而获得的。

7.如权利要求1所述的方法,其中所述查询步骤还包括,在所述命名实体备选列表中没有相似度高于预定阈值的命名实体的情况下,仅保留所述文本中与优先级最高的槽位匹配的第一字段,并将第一字段在命名实体备选列表中进行查询,确定到所述命名实体备选列表中包含第一字段的一个或多个命名实体,将所述一个或多个命名实体作为识别的命名实体。

8.一种识别命名实体的装置,包括:

匹配单元,所述匹配单元被配置为基于命名实体中槽位的预定的优先级,按照优先级从高到低的顺序,将文本中的字段与各个槽位进行匹配;

连接单元,所述连接单元被配置为按照优先级和预定的逻辑关系将匹配到的槽位中的字段连接以得到连接结果;以及

查询单元,所述查询单元被配置为将所述连接结果在命名实体备选列表中进行查询,确定所述命名实体备选列表中与所述连接结果匹配的一个或多个命名实体,将所述一个或

多个命名实体作为识别的命名实体；

相似度计算单元，被配置为在所述命名实体备选列表中没有与所述连接结果匹配的命名实体的情况下，利用字符串相似度计算方法，计算所述连接结果与命名实体备选列表中的命名实体的相似度，并将相似度高于预定阈值的一个或多个命名实体作为相似的命名实体，并确定所述相似的命名实体中的最长公共子序列中的最短连续子串的长度大于第二阈值的命名实体作为识别的命名实体。

9.如权利要求8所述的装置，其中所述字符串相似度计算方法计算所述连接结果与命名实体备选列表中的命名实体的最长公共子序列长度，并且将所述最长公共子序列长度和连接结果的长度的比值作为所述相似度。

10.如权利要求9所述的装置，其中所述字符串相似度计算方法还包括利用动态规划矩阵，计算所述最长公共子序列长度。

11.如权利要求8所述的装置，其中所述相似度计算单元被配置为根据动态规划矩阵，确定所述相似的命名实体中的最长公共子序列中的最短连续子串的长度大于第二阈值的命名实体作为识别的命名实体。

12.如权利要求8所述的装置，其中所述的预定的逻辑关系包括：具有相同优先级的槽位中的字段之间逻辑关系设为“或”，并且将不同优先级的槽位中的字段之间的逻辑关系设为“与”。

13.如权利要求8所述的装置，其中所述查询单元还被配置为，在所述连接结果没有匹配到命名实体备选列表中的命名实体时，仅保留优先级最高的槽位中的第一字段，并在命名实体备选列表中进行查询，确定到所述命名实体备选列表中包含第一字段的一个或多个命名实体，将所述一个或多个命名实体作为识别的命名实体。

14.如权利要求8所述的装置，还包括优先级设定单元，所述优先级设定单元被配置为根据对命名实体备选列表中的所有命名实体的统计，将列表中的命名实体的字段归类为多个槽位，并对每个槽位赋予优先级而获得所述命名实体中槽位的预定的优先级。

15.一种识别命名实体的系统，其特征在于包括：

一个或更多个处理器；以及

一个或更多个存储器，被配置为存储一系列计算机可执行指令，

其中所述一系列计算机可执行指令在由所述一个或更多个处理器运行时使得所述一个或更多个处理器执行根据权利要求1-7中的任意一项所述的方法。

16.一种非暂态计算机可读存储介质，其上存储有程序，其特征在于，所述程序被处理器执行时实现权利要求1-7中的任意一项所述的方法的步骤。

一种识别命名实体的方法和装置

技术领域

[0001] 本公开涉及自然语言处理的方法，更具体地，本公开涉及一种识别自然语言中的命名实体的方法和装置。

背景技术

[0002] 在涉及自然语言处理(例如信息抽取、信息检索、机器翻译、问答系统等)的研究中，通常需要对自然语言的文本中的实体名称进行识别，即从自然叙述的非结构化信息中提取命名实体(命名实体(Named Entity)识别)。命名实体是指包括物品名称、人名、地名、机构名、时间、数量特征、专有名词等特定种类词的集合，更广泛地而言，命名实体可以是任何符合特定需要的特殊文本段落。命名实体识别是自然语言处理的一个极为重要的基础任务。在基于自然语言的信息提取和检索方面，命名实体识别有着广泛的应用。

[0003] 槽位(slot)与自然语言的处理中所需要获取的信息元素相对应。例如，要想查找对应于特定的电器的命名实体，需要知道品牌、型号等元素，这些元素可以被认为是槽位。

[0004] 通常采用机器学习(条件随机场等)、关键词匹配等方式进行命名实体识别。当输入的自然语言文本中的命名实体为常见名词时，命名实体识别较为简单。

[0005] 然而，在特定领域，命名实体可能会具有较为复杂的结构。例如，在电子产品名称领域，可能存在品牌名-系列-型号这样的多层结构的命名实体名称。如果使用传统的关键词匹配的方式进行命名实体的识别，会导致命名实体各种变形(例如，别称)、不完整的命名实体(例如，缺少系列名或型号)以及命名实体中字段是乱序的等情况无法被成功识别。如果采用基于统计的机器学习方法进行识别，则需要大量专业的人工标记语料，从而使得机器学习的方式的代价高的同时收获小。

[0006] 因此，需要一种高效准确地识别复杂命名实体的方法和系统。

发明内容

[0007] 在下文中给出了关于本公开的简要概述，以便提供关于本公开的一些方面的基本理解。但是，应当理解，这个概述并不是关于本公开的穷举性概述。它并不是意图用来确定本公开的关键性部分或重要部分，也不是意图用来限定本公开的范围。其目的仅仅是以简化的形式给出关于本公开的某些概念，以此作为稍后给出的更详细描述的前序。

[0008] 根据本公开的一个方面，提供了一种识别命名实体的方法，包括：匹配步骤，基于命名实体中槽位的预定的优先级，按照优先级从高到低的顺序，将文本中的字段与各个槽位进行匹配；连接步骤，按照优先级和预定的逻辑关系将匹配到的槽位中的字段连接以得到连接结果；以及查询步骤，将所述连接结果在命名实体备选列表中进行查询，确定所述命名实体备选列表中与所述连接结果匹配的一个或多个命名实体，将所述一个或多个命名实体作为识别的命名实体。

[0009] 根据本公开的另一个方面，提供一种识别命名实体的装置，包括：匹配单元，所述匹配单元被配置为基于命名实体中槽位的预定的优先级，按照优先级从高到低的顺序，将

文本中的字段与各个槽位进行匹配;连接单元,所述连接单元被配置为按照优先级和预定的逻辑关系将匹配到的槽位中的字段连接以得到连接结果;以及查询单元,所述查询单元被配置为将所述连接结果在命名实体备选列表中进行查询,确定所述命名实体备选列表中与所述连接结果匹配的一个或多个命名实体,将所述一个或多个命名实体作为识别的命名实体。

[0010] 根据本发明的另一个方面,提供了一种识别命名实体的系统,该系统包括:一个或更多个处理器;以及一个或更多个存储器,被配置为存储一系列计算机可执行指令,其中所述一系列计算机可执行指令在由所述一个或更多个处理器运行时使得所述一个或更多个处理器执行如上所述的方法。

[0011] 根据本发明的另一个方面,提供了一种非暂态的计算机可读介质,其上存储有计算机可执行指令,所述计算机可执行指令在由一个或更多个处理器运行时使得所述一个或更多个处理器执行如上所述的方法。

[0012] 通过以下参照附图对本公开的示例性实施例的详细描述,本公开的其它特征及其优点将会变得更为清楚。

附图说明

[0013] 构成说明书的一部分的附图描述了本公开的实施例,并且连同说明书一起用于解释本公开的原理。

[0014] 参照附图,根据下面的详细描述,可以更加清楚地理解本公开,其中:

[0015] 图1是示出了根据本发明一个示例性实施例的识别命名实体的方法的示例性流程图。

[0016] 图2是示出了根据本发明另一个示例性实施例的识别命名实体的方法的示例性流程图。

[0017] 图3是示出了根据本发明一个示例性实施例的识别命名实体的方法的具体流程图。

[0018] 图4是示出了根据本发明一个示例性实施例的利用动态规划矩阵的相似度计算步骤的具体流程图。

[0019] 图5是示出了根据本发明一个示例性实施例的系统的构成的示意图。

[0020] 图6是示出可以实现根据本发明的实施例的计算设备的示例性配置图。

[0021] 注意,在以下说明的实施方式中,有时在不同的附图之间共同使用同一附图标记来表示相同部分或具有相同功能的部分,而省略其重复说明。在一些情况中,使用相似的标号和字母表示类似项,因此,一旦某一项在一个附图中被定义,则在随后的附图中不需要对其进行进一步讨论。

[0022] 为了便于理解,在附图等中所示的各结构的位置、尺寸及范围等有时不表示实际的位置、尺寸及范围等。因此,本公开并不限于附图等所公开的位置、尺寸及范围等。

具体实施方式

[0023] 下面将参照附图来详细描述本公开的各种示例性实施例。应注意到:除非另外具体说明,否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本

公开的范围。

[0024] 以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本公开及其应用或使用的任何限制。也就是说，本文中的结构及方法是以示例性的方式示出，来说明本公开中的结构和方法的不同实施例。然而，本领域技术人员将会理解，它们仅仅说明可以用来实施的本公开的示例性方式，而不是穷尽的方式。此外，附图不必按比例绘制，一些特征可能被放大以示出具体组件的细节。

[0025] 本公开提出了一种识别命名实体名称的方法，包括匹配步骤，其中基于命名实体中槽位的预定的优先级，按照优先级从高到低的顺序，将文本中的字段与各个槽位进行匹配。如果优先级最高的槽位没有匹配到任何结果，则命名实体识别无结果。如果优先级最高的槽位匹配到了字段，则进行下一级别优先级的槽位匹配。将匹配到的槽位中的各个字段按照优先级的顺序和预定的逻辑关系进行连接，并输出所有连接结果。然后，将该连接结果在命名实体备选列表中进行查询，确定所述命名实体备选列表中与所述连接结果匹配的一个或多个命名实体，然后将所述一个或多个命名实体作为识别的命名实体。

[0026] 与传统的直接进行关键字匹配的方法相比，本公开的技术方案可以准确地覆盖各种命名实体的变形(例如，智能手机的别名、同一品牌下电器的不同型号、名称未按照品牌-型号的顺序连接等)的情况，而与机器学习的方法相比，本公开的技术方案可以减少大量人工标记语料以进行模型训练的工作量，效率更高。

[0027] 图1是示出了根据本发明一个示例性实施例的识别命名实体的方法100的示例性流程图。如图1所示，该识别命名实体的方法100可以包括：匹配步骤110、连接步骤120和查询步骤130。

[0028] 首先，在匹配步骤110中，基于预定的命名实体中槽位的优先级，按照优先级从高到低的顺序，将输入的文本(例如，自然语句)中的字段进行槽位匹配。如果优先级最高的槽位没有匹配到任何结果，则命名实体识别无结果。如果优先级最高的槽位匹配到了字段，则进行下一级别优先级的槽位匹配。然后，处理进行到连接步骤120。

[0029] 在一些实施方式中，该命名实体中槽位的预定的优先级是是根据对命名实体备选列表中的所有命名实体的统计，将备选列表中的命名实体的字段归类为多个槽位，并对每个槽位赋予优先级而获得。例如，在待识别的自然语言中的命名实体可能属于智能终端领域时，首先统计所有终端的名称信息(例如，可以基于各个终端品牌网站上收集/抓取的信息)。在统计到相应的信息后，设计针对智能终端信息的槽位的类别。

[0030] 在一些实施方式中，可能存在多个优先级相同的待匹配槽位。在这种情况下，可以并列匹配优先级相同的多个槽位，然后再匹配下一优先级的槽位。继续以智能终端为例，设计的槽位可以包括智能终端的品牌、系列、型号、别名等。随后，对这些槽位赋予优先级，例如，优先级最高的是品牌和系列，优先级较低的是型号和别名。

[0031] 在连接步骤120中，将匹配到的多个槽位中的字段按照优先级顺序和预定的逻辑关系进行连接，并输出所有连接结果。例如，预定的逻辑关系可以包括“与”、“或”等。基于连接逻辑关系，连接结果通常包括匹配到的多个槽位的字段的不同组合形式。然后，处理进行到查询步骤130。

[0032] 在一些实施方式中，预定的逻辑关系被设置为具有相同优先级的槽位中的字段之间逻辑关系设为“或”，并且将不同优先级的槽位中的字段之间的逻辑关系设为“与”。

[0033] 在查询步骤130中，将所有连接结果在命名实体备选列表中进行查询，如果可以匹配到所述命名实体备选列表中的一个或多个命名实体，将所述一个或多个命名实体均输出为命名实体识别结果。

[0034] 在一些实施方式中，存在在查询步骤130中输入的所有的连接结果均无法匹配到命名实体备选列表中的命名实体的情况，在这种情况下，可以在识别自然语言中的命名实体的方法中加入相似度计算步骤来寻找相似的命名实体。下面将参考图2示出了包括相似度计算步骤的识别自然语言中的命名实体的方法200的示例性流程图。如图2所示，该识别自然语言中的命名实体的方法200可以包括：匹配步骤210、连接步骤220、查询步骤230和相似度计算步骤240。为了简洁起见，仅详细描述与方法100不同的相似度计算步骤240。

[0035] 当在查询步骤230中，所有连接结果均没有匹配到命名实体备选列表中的任何命名实体时，处理进行到相似度计算步骤240。在相似度计算步骤240中，利用字符串相似度计算方法，计算各个连接结果与命名实体备选列表中的命名实体的相似度，并将相似度高于预定阈值的一个或多个命名实体输出为命名实体识别结果。

[0036] 在一些实施方式中，上述字符串相似度计算方法可以包括计算所述连接结果与命名实体备选列表中的命名实体的最长公共子序列长度。另外，还可以利用动态规划矩阵，计算与所述连接结果相似度最大的命名实体备选列表中的命名实体并将其输出为命名实体识别结果。在下文中，将参考图4详细说明利用动态规划矩阵的字符串相似度计算方法的具体示例。

[0037] 在一些实施方式中，如果基于字符串相似度计算方法仍没有查询到命名实体备选列表中的命名实体时，可以仅保留优先级最高的槽位中的字段，并在命名实体备选列表中进行查询，如果可以匹配到所述命名实体备选列表中的一个或多个命名实体，则将一个或多个命名实体输出为命名实体识别结果。

[0038] 为了更清楚明了地体现出本发明的方法流程，下面将参考图3描述根据本发明的一个具体实施例。图3是示出了根据本发明一个示例性实施例的识别智能终端名称的命名实体的方法的详细步骤图。

[0039] 首先，在步骤S301处，接收待识别命名实体的自然语言输入的语句。该语句可以包括任何多层次的待识别的命名实体，例如，电子产品名称、商品套餐名称、地址等。在一些实施方式中，例如，该自然语言输入可以是查询某个特定的智能终端的名称的问题。

[0040] 如上文参考图1所描述的，在本示例中，已经统计了从各个终端品牌网站上收集/抓取的信息、整理针对智能终端的命名实体备选列表，并设计针对智能终端信息的槽位的类别，从而获取了相应的槽位优先级信息。如下表1所示，针对每个终端名称设计对应四个槽位，分别是品牌、系列、型号、别名，并赋予优先级信息，优先级最高的槽位对应于品牌和系列，优先级较低的是型号和别名。

[0041]

槽位	华为AscendP7	三星GalaxyA3
品牌	华为	三星
系列	Ascend	Galaxy
型号	P7	A3
别名	/	A3009

[0042] 表1

[0043] 随后，处理进行到步骤S302。在步骤S302处，使用槽位优先级信息在自然语言输入的语句中进行匹配。在一些实施方式中，可以先并列匹配优先级最高的槽位(即，品牌名称与系列名称)。

[0044] 在步骤S303处判断是否成功匹配，如果优先级最高的槽位没有匹配成功，则命名实体识别结束，并且不返回任何结果。如果匹配成功，处理进行到步骤S304处。

[0045] 在步骤S304处，再并列匹配优先级较低的槽位(例如，型号名称与别名)。如果在步骤S305处匹配成功，则处理进行到步骤S306。

[0046] 在步骤S306处，将匹配出的各个槽位中的字段信息按照预先设定的逻辑关系进行连接，例如，可以按照[(“品牌”或“系列”)与(“型号”或“别名”)]的逻辑关系进行连接。例如，在品牌槽位对应的字段为三星，系列为Galaxy，型号为A3，别名为3009的情况下，输出的连接结果可以包括三星A3、GalaxyA3、三星3009以及Galaxy3009。随后，处理进行到步骤S308处。在一些其他的实施方式中，可以按照任意逻辑关系连接槽位中的字段，例如，也可以按照品牌-系列-型号-别名顺序连接各个槽位中的字段。

[0047] 在步骤S308处，在预先整理的针对智能终端的命名实体备选列表中查询所有的连接结果，若可匹配到备选列表中的命名实体，则输出为最终的命名实体识别结果。如果无法匹配到任何命名实体(例如，以智能终端名称为例，如果在输入的自然语言的语句中匹配出(品牌＝红米)和(型号＝9)的情况下，因为红米9这款手机不存在，所以将无法匹配到任何命名实体)，则处理进行到步骤S310。

[0048] 在步骤S310处，通过基于改进的动态规划矩阵的字符串相似度计算方法，计算连接结果与命名实体备选列表中的命名实体的相似度。如果相似度大于或等于预先设定的阈值，则在步骤S311处将该连接结果输出为相似的命名实体识别的结果并且处理进行到步骤S312处。在步骤S312处，进一步对相似的命名实体进行检查，并输出识别的命名实体，下面将参考图4详细描述该检查步骤。如果小于预先设定的阈值，则处理进行到步骤S313处。

[0049] 在步骤S313处，则仅保留优先级最高的槽位中的字段(即，对应于品牌和系列槽位信息)，再按照[(“品牌”或“系列”)]的逻辑关系进行连接，随后处理返回到S308处，并在预先整理的针对智能终端的命名实体备选列表中查询该修改后的连接结果的字段，若可匹配到备选列表中包含该字段的命名实体，则输出为最终的命名实体识别结果。

[0050] 图3中示出了以智能终端为例的识别的命名实体的方法300，本领域技术人员应当理解，命名实体不限于此，并且其他领域的命名实体的识别方法与方法300类似，区别仅在于针对各个领域的统计信息，槽位的优先级将被设计为不同。例如，在待识别的命名实体是手机通话/流量套餐的情况下，可以收集并整理了运营商提供的各个套餐的相关信息。例如，可以为每个套餐设计对应五个槽位，分别是名称、价格、流量、语音、别名，并赋予相应的优先级。在这个示例中，优先级较高的是名称和别名，优先级较低的是流量、价格、语音。以“天翼畅享69元套餐”为例，在统计过程中，该命名实体可以如下表2进行槽位分类：

[0051]

槽位	天翼畅享69元套餐
名称	天翼
价格	69元
流量	/
语音	500分钟

别名	畅享

[0052] 表2

[0053] 如上文中方法200的步骤240以及方法300的步骤S310所述的，在查询步骤无法匹配到命名实体时，可以进行相似度计算步骤。在相似度计算步骤中，除了传统的相似度计算，还可以通过基于改进的动态规划矩阵的字符串相似度计算方法，计算连接结果与命名实体备选列表中的命名实体的相似度，并基于动态规划矩阵来检查相似的命名实体，得到最优化的命名实体识别结果。下面将参考图4，详细描述通过基于改进的动态规划矩阵的字符串相似度计算方法的一个具体实例。

[0054] 首先，在步骤410处，将字符串A设置为连接步骤中输出的连接结果(即，将匹配到的多个槽位中的字段按照优先级顺序和预定的逻辑关系进行连接的所有连接结果)，字符串设置为备选的命名实体列表中给的各个命名实体。随后，处理进行到步骤420处。

[0055] 在步骤420，获取字符串A和字符串B的最长公共子序列并获得其长度。本领域技术人员应当理解，最长公共子序列是指两个或多个字符串之间最长子序列，其中子序列不需要在原序列中占用连续的位置。与之相对，子串则需要是连续的。

[0056] 在步骤430处，通过归一化的最长公共子序列(LCS，Longest Common Subsequence)长度计算字符串A与字符串B的相似度，相似度如公式1所示：

[0057] $$\text{similarity}\quad(A,B)=\frac{LCS}{len\ (A)}$$

[0058] 公式1

[0059] 在一些实施方式中，可以利用动态规划矩阵来循环检查字符串A与字符串B的相似度。例如，在动态规划矩阵中，定义字符串A的长度为m，字符串B的长度为n，dp[i][j]为字符串A的第一个字符到第i个字符串和字符串B的第一个字符到第j个字符的最长公共子序列，整个矩阵的大小为(n+1)x(m+1)。在初始状态时：dp[i][0]＝0，dp[0][j]＝0。

[0060] 随后的状态转移方程为：

[0061] 当A[i-1]!＝B[j-1]时，dp[i][j]＝max{dp[i-1][j]，dp[i][j-1]}

[0062] 当A[i-1]＝＝B[j-1]时，dp[i][j]＝dp[i-1][j-1]+1

[0063] 随后，处理进行到步骤440。在该步骤中，获得大于等于预定第一阈值的所有相似结果。例如，以上文中无法匹配到任何命名实体的智能终端名称“红米9”为例，在预定阈值被设置为0.6的情况下，可以查询到红米8、小米9、小米CC9、红辣椒9X与“红米9”的相似度均为0.667(2/3)。随后，处理进行到步骤450。

[0064] 在步骤450处，检查字符串A和字符串B的最长连续公共子序列中最短的子串(即，连续的字符)的长度是否符合预定的第二阈值，并将不符合条件的相似结果筛除。在一些实施方式中，这样的检查可以通过直接检查步骤440中生成的动态规划矩阵的方式来完成。例如，该检查方法可以包括如下步骤：

[0065] 1)找出初始重合的位置，检查标记值设为1，再以此位置沿着矩阵斜对角线开始循环检查；

[0066] 2)若斜对角线上数值呈现递增的状态，则检查标记值加1；

[0067] 3)若斜对角线上的数值不增不减，

[0068] a)当前检查标记值不为1时，将检查标记值归0；

[0069] b)当检查标记值为1时，则直接跳过4)，结束检查，将相似度置为0，也就是不认为两字符串匹配；

[0070] 4)继续检查斜对角线上的下一数值，重复2)和3)操作，直至对角线上的数值为空。

[0071] 最后相似度没有被置为0的字符串B作为检索出的实体名称。

[0072] 以上文中未能匹配到命名实体的“红米9”(字符串A)与备选列表中的“红米8”(字符串B)以及“红辣椒9X”(字符串B')为例，其动态规划矩阵可以被表示为如下表3A和表3B所示：

[0073]

	0	红	米	8
0	0	0	0	0
红	0	1	1	1
米	0	1	2	2
9	0	1	2	2

[0074] 表3A

[0075]

	0	红	辣	椒	9	X
0	0	0	0	0	0	0
红	0	1	1	1	1	1
米	0	1	1	1	1	1
9	0	1	1	1	2	2

[0076] 表3B

[0077] 基于上述检查方法进行检查后，由于“红辣椒9X”的公共子序列中的最短子串长度仅为1，因此它与“红米9”相似度被置为0，该相似结果被排除。类似的，“小米CC9”也将被排除。仅有红米8、小米9作为被识别为相似的命名实体。由此可见，基于动态规划矩阵的字符相似度比较方法及基于动态规划矩阵的检查方法保证了命名实体匹配结果的相似度和合理性。

[0078] 图5是示出了根据本发明一个示例性实施例的用于识别命名实体的装置500的基本配置的框图。

[0079] 如图5所示，该识别命名实体的装置500包括：匹配单元510、连接单元520、查询单元530和相似度计算单元540。其中，该匹配单元510基于命名实体中槽位的预定的优先级，按照优先级从高到低的顺序，将输入的自然语句中的字段进行槽位匹配；该连接单元520，匹配到的槽位中的字段按照优先级和预定的逻辑关系进行连接，并输出所有连接结果；该查询单元530将连接结果在命名实体备选列表中进行查询，如果可以匹配到所述命名实体备选列表中的一个或多个命名实体，将一个或多个命名实体输出为命名实体识别结果；当连接结果没有匹配到命名实体备选列表中的命名实体时，相似度计算单元540利用字符串相似度计算方法，计算所述连接结果与命名实体备选列表中的命名实体的相似度，并将相似度高于预定阈值的一个或多个命名实体输出为命名实体识别结果。本领域技术人员应当理解，识别命名实体的装置500所包含的部件可以不限于上述部件510-540，而是可以包括用于实现根据本发明实施例的前述方法的其他步骤的部件。装置500的各个部件可以由硬件、软件、固件或其任意组合来实现。另外，本领域技术人员也应当理解，装置500的各个部件可以根据需要被组合或分割成子部件。装置500的上述各个部件不限于上述的各个功能，

而是可以实现如前所述的根据本发明实施例的各种方法的相应步骤的功能。

[0080] 图6示出了可以实现根据本发明的实施例的计算设备2000的示例性配置。计算设备2000是可以应用本发明的上述方面的硬件设备的实例。计算设备2000可以是被配置为执行处理和/或计算的任何机器。计算设备2000可以是但不限制于工作站、服务器、台式计算机、膝上型计算机、平板计算机、个人数据助手(PDA)、智能电话、车载计算机或以上组合。前述装置500可以全部或至少部分地由上述计算设备2000或与其相似的设备或系统实现。

[0081] 如图6所示,计算设备2000可以包括可能经由一个或多个接口与总线2002连接或通信的一个或多个元件。例如,计算设备2000可以包括总线2002、一个或多个处理器2004、一个或多个输入设备2006以及一个或多个输出设备2008。总线2002可以包括但不限于,工业标准架构(Industry Standard Architecture,ISA)总线、微通道架构(Micro Channel Architecture,MCA)总线、增强ISA(EISA)总线、视频电子标准协会(VESA)局部总线、以及外设组件互连(PCI)总线等。一个或多个处理设备2004可以是任何种类的处理器,并且可以包括但不限于一个或多个通用处理器或专用处理器(诸如专用处理芯片)。输入设备2006可以是能够向计算设备输入信息的任何类型的输入设备,并且可以包括但不限于鼠标、键盘、触摸屏、麦克风和/或远程控制器。输出设备2008可以是能够呈现信息的任何类型的设备,并且可以包括但不限于显示器、扬声器、视频/音频输出终端、振动器和/或打印机。计算设备2000还可以包括或被连接至非暂态存储设备2010,该非暂态存储设备2010可以是任何非暂态的并且可以实现数据存储的存储设备,并且可以包括但不限于盘驱动器、光存储设备、固态存储器、软盘、柔性盘、硬盘、磁带或任何其他磁性介质、压缩盘或任何其他光学介质、ROM(只读存储器)、RAM(随机存取存储器)、缓存存储器和/或任何其他存储芯片或单元、和/或计算机可以从其中读取数据、指令和/或代码的其他任何介质。非暂态存储设备2010可以与任何接口可拆卸地连接。非暂态存储设备2010可以具有存储于其上的、用于实现前述用于在区块链网络中进行共识的方法和/或步骤的数据/指令/代码。计算设备2000还可以包括通信设备2012,该通信设备2012可以是能够启用与外部装置和/或网络通信的任何种类的设备或系统,并且可以包括但不限于调制解调器、网络卡、红外线通信设备、无线通信设备和/或芯片集(诸如蓝牙™设备、1302.11设备、WiFi设备、WiMax设备、蜂窝通信设施等)。

[0082] 计算设备2000还可以包括工作存储器2014。该工作存储器2014可以是能够存储对于处理器2004有用的指令和/或数据的任何类型的工作存储器,并且可以包括但不限于随机存取存储器(RAM)和只读存储器(ROM)。

[0083] 位于上述工作存储器上的软件元件可以包括但不限于操作系统2016、一个或多个应用程序2018、驱动器和/或其他数据和代码。上述一个或多个应用程序2018可以包括用于执行如上所述的识别命名实体的各方法及各步骤的指令。可以通过读取和执行一个或多个应用程序2018的处理器实现前述识别命名实体的系统300的各部件/单元/元件,例如匹配单元310、连接单元320、查询单元330及相似度比较单元340等等。软件元件的指令的可执行代码或源代码可以存储在非暂态计算机可读存储介质(诸如如上所述的存储设备2010)中,并且可以通过编译和/或安装读入工作存储器2014中。还可以从远程位置下载软件元件的指令的可执行代码或源代码。

[0084] 应当理解,可以根据特定要求进行变型。例如,可以使用定制的硬件和/或特定元件可以以硬件、软件、固件、中间件、微代码、硬件描述语言或其任何组合的方式实现。此外,

可以采用与其他计算设备(诸如网络输入/输出设备)的连接。例如,本发明的方法和设备中的一些或全部可以根据本发明通过使用汇编语言编程硬件(例如,包括现场可编程门阵列(FPGA)和/或可编程逻辑阵列(PLA)的可编程逻辑电路)或逻辑和算法的硬件编程语言(例如VERILOG,VHDL,C++)来实现。

[0085] 应当进一步理解,计算设备2000的元件可以分布在整个网络上。例如,可以在使用一个处理器执行一些处理的同时,使用其他远程处理器执行其他处理。计算机系统2000的其他元件也可以类似地分布。因此,计算设备2000可以被理解为在多个地点执行处理的分布式计算系统。

[0086] 可以通过许多方式来实施本发明的方法和设备。例如,可以通过软件、硬件、固件、或其任何组合来实施本发明的方法和设备。上述的方法步骤的次序仅是说明性的,本发明的方法步骤不限于以上具体描述的次序,除非以其它方式明确说明。此外,在一些实施例中,本发明还可以被实施为记录在记录介质中的程序,其包括用于实现根据本发明的方法的机器可读指令。因而,本发明还覆盖存储用于实现根据本发明的方法的程序的记录介质。

[0087] 虽然已通过示例详细展示了本发明的一些具体实施例,但是本领域技术人员应当理解,上述示例仅意图是说明性的而不限制本发明的范围。本领域技术人员应该理解,上述实施例可以在不脱离本发明的范围和实质的情况下被修改。本发明的范围是通过所附的权利要求限定的。

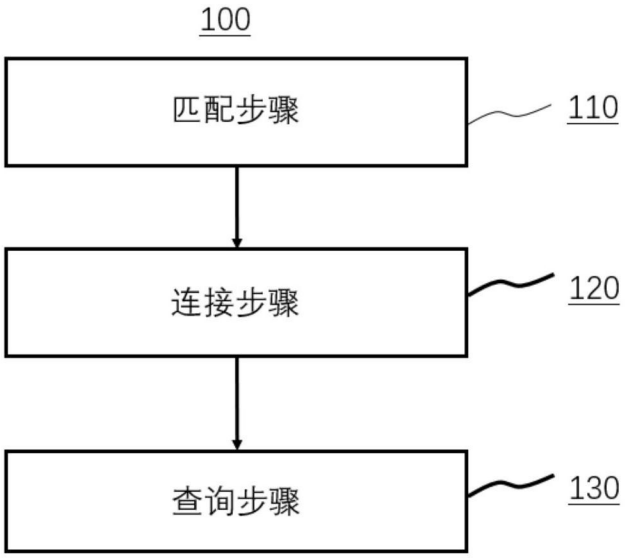

100
匹配步骤
110
连接步骤
120
查询步骤
130

图1

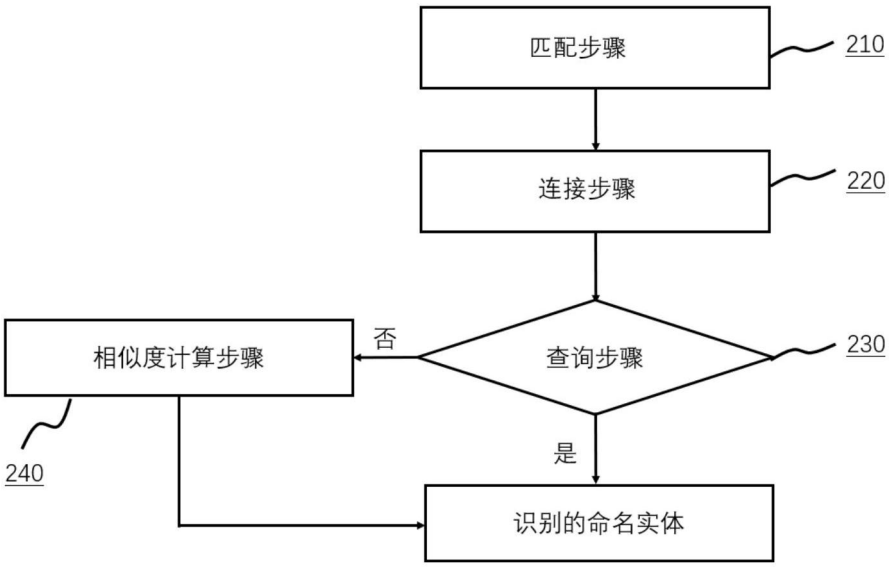

匹配步骤
210
连接步骤
220
查询步骤
230
否
相似度计算步骤
240
是
识别的命名实体

图2

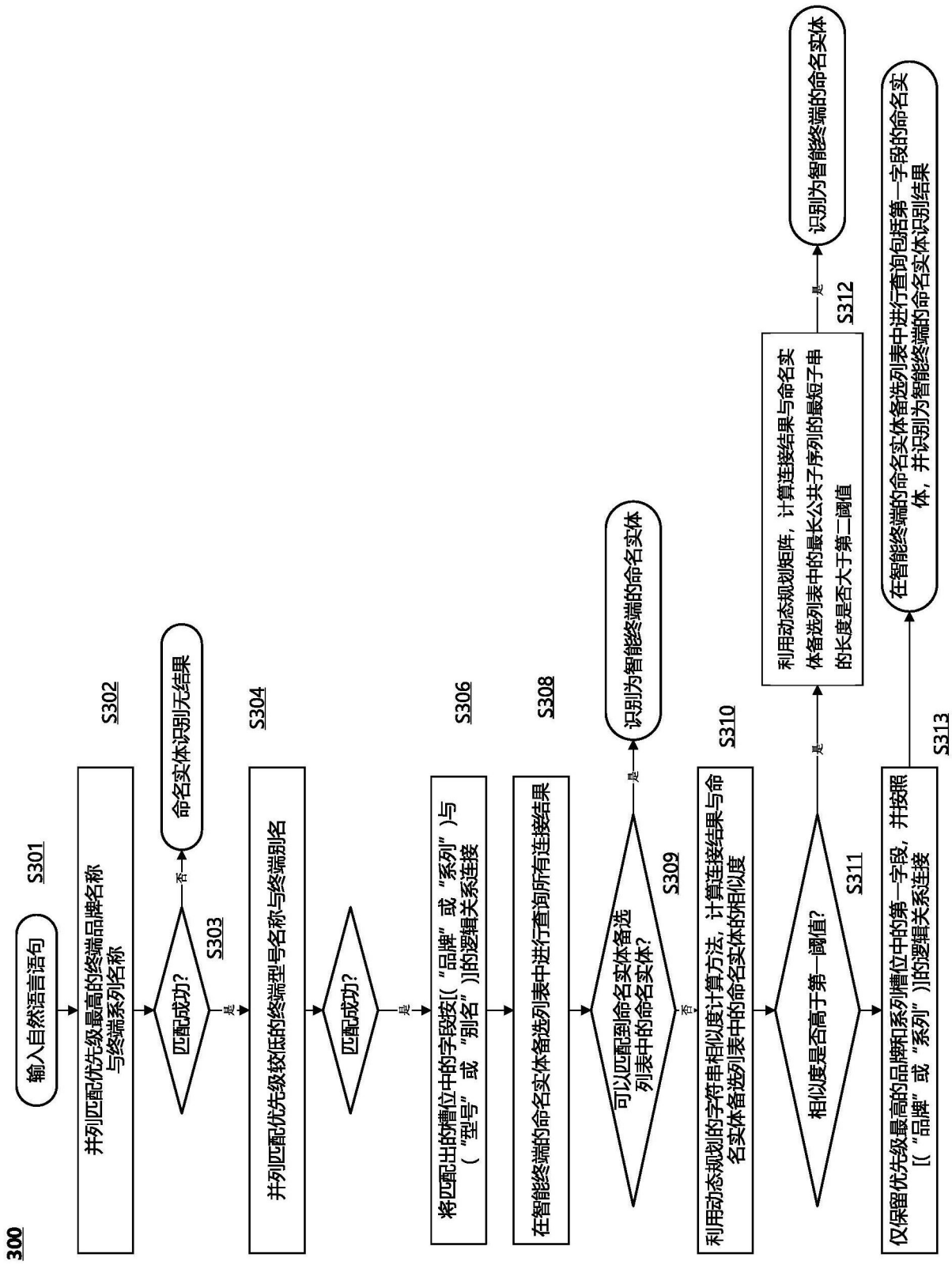

300
输入自然语言语句
S301
并列匹配优先级最高的终端品牌名称与终端系列名称
S302
匹配成功?
否
命名实体识别无结果
S303
是
并列匹配优先级较低的终端型号名称与终端别名
S304
匹配成功?
是
将匹配出的槽位中的字段按(“品牌”或“系列”)与(“型号”或“别名”)的逻辑关系连接
S306
在智能终端的命名实体备选列表中进行查询所有连接结果
S308
可以匹配到命名实体备选列表中的命名实体?
是
识别为智能终端的命名实体
否
S309
利用动态规划的字符串相似度计算方法，计算连接结果与命名实体备选列表中的命名实体的相似度
S310
相似度是否高于第一阈值?
是
利用动态规划矩阵，计算连接结果与命名实体备选列表中的最长公共子序列的最短子串的长度是否大于第二阈值
S312
是
识别为智能终端的命名实体
S311
仅保留优先级最高的品牌和系列槽位中的第一字段，并按照[(“品牌”或“系列”)]的逻辑关系连接
S313
在智能终端的命名实体备选列表中进行查询包括第一字段的命名实体，并识别为智能终端的命名实体识别结果

图3

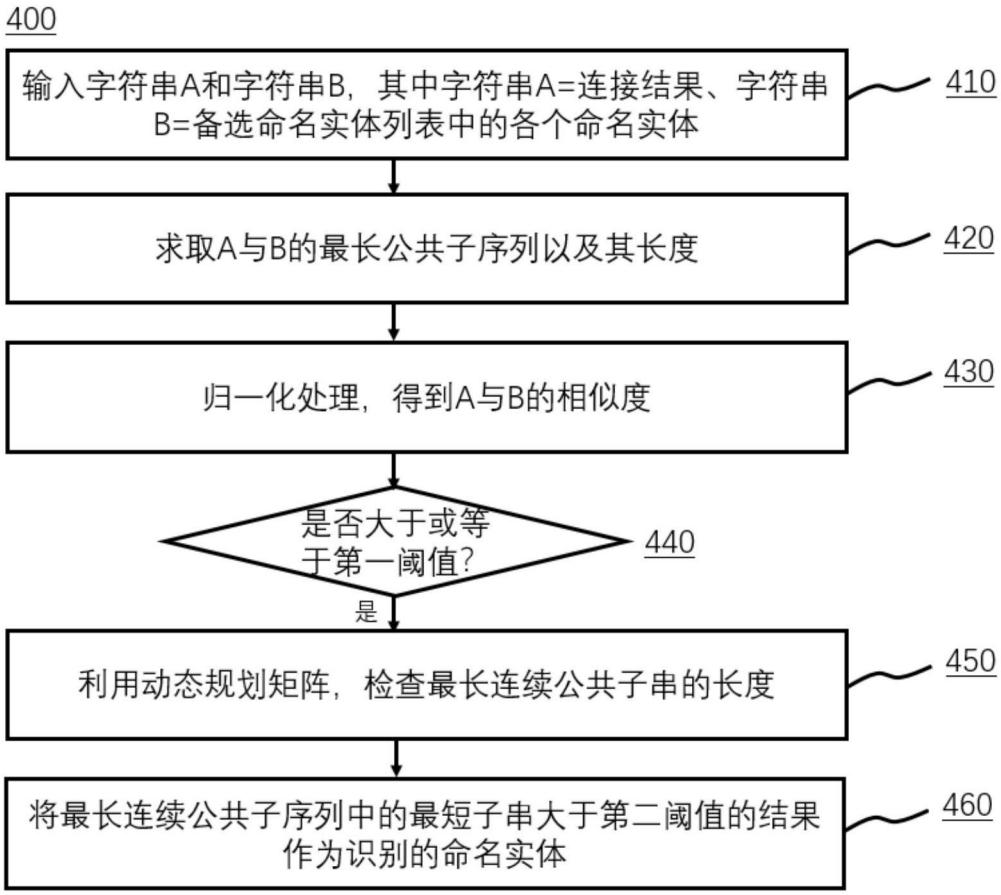

400
输入字符串A和字符串B，其中字符串A=连接结果、字符串B=备选命名实体列表中的各个命名实体
410
求取A与B的最长公共子序列以及其长度
420
归一化处理，得到A与B的相似度
430
是否大于或等于第一阈值?
440
是
利用动态规划矩阵，检查最长连续公共子串的长度
450
将最长连续公共子序列中的最短子串大于第二阈值的结果作为识别的命名实体
460

图4

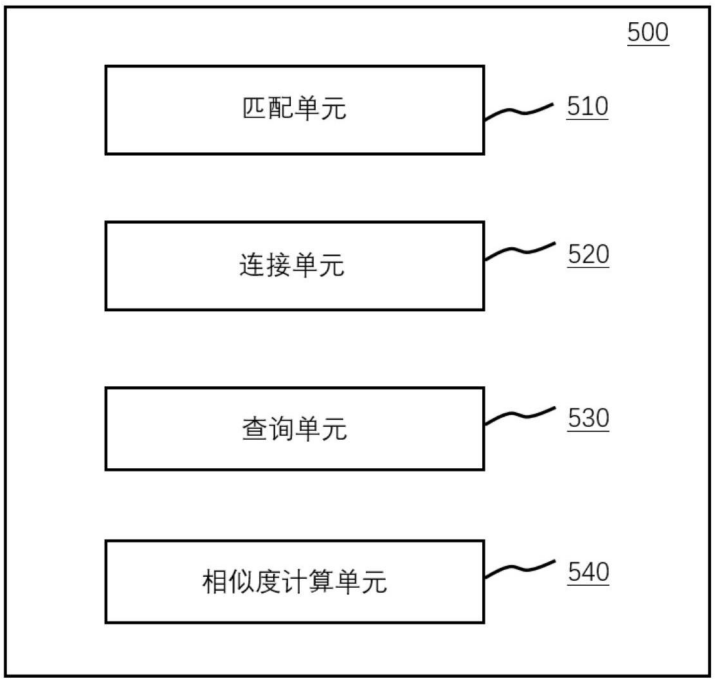

500
匹配单元
510
连接单元
520
查询单元
530
相似度计算单元
540

图5

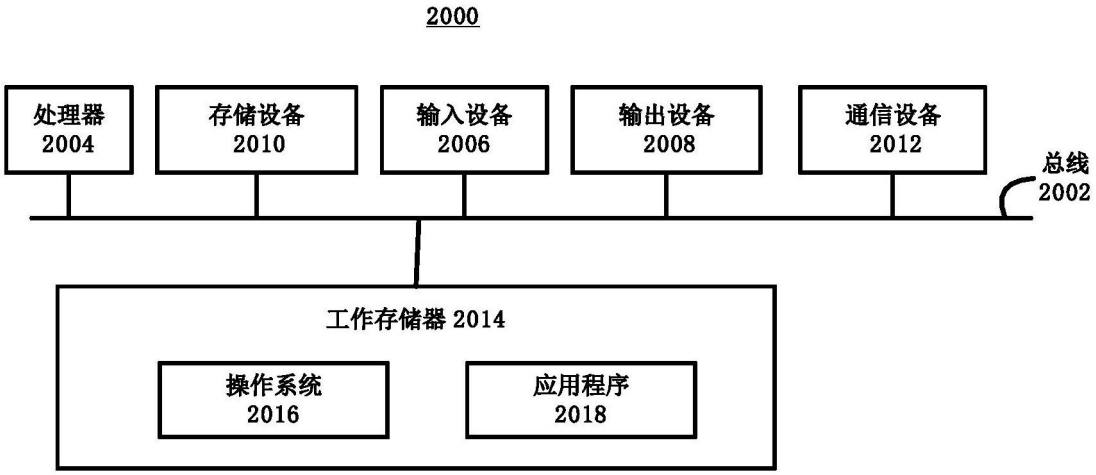

2000
处理器
2004
存储设备
2010
输入设备
2006
输出设备
2008
通信设备
2012
总线
2002
工作存储器 2014
操作系统
2016
应用程序
2018

图6