(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113408660 B
(45)授权公告日 2024.05.24

(21)申请号 202110803686.8

(22)申请日 2021.07.15

(65)同一申请的已公布的文献号
申请公布号 CN 113408660 A

(43)申请公布日 2021.09.17

(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 柳正青 蓝琰佳 赵旭

(74)专利代理机构 中科专利商标代理有限责任公司 11021
专利代理师 李世阳

(51)Int.Cl.
G06F 18/23(2023.01)
G06F 18/22(2023.01)

(56)对比文件
CN 101350032 A,2009.01.21
CN 103034656 A,2013.04.10
CN 103559259 A,2014.02.05
CN 104615768 A,2015.05.13
CN 106445967 A,2017.02.22
CN 107908650 A,2018.04.13
CN 108304379 A,2018.07.20
CN 110489558 A,2019.11.22
CN 110888981 A,2020.03.17
CN 111353296 A,2020.06.30
CN 112084776 A,2020.12.15
CN 112329548 A,2021.02.05
CN 112560444 A,2021.03.26
CN 112926308 A,2021.06.08
US 2013268554 A1,2013.10.10
WO 2017149711 A1,2017.09.08
黄永;陆伟;程齐凯.学术文本的结构功能识别——基于章节内容的识别.情报学报.2016,(03),全文.

审查员 赵鹏翔

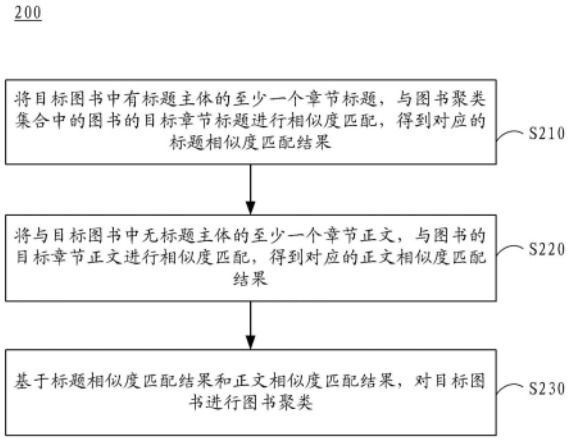
权利要求书2页 说明书11页 附图3页

(54)发明名称

图书聚类方法、装置、设备和存储介质

(57)摘要

本公开公开了一种图书聚类方法,涉及互联网技术领域,尤其涉及大数据和智能搜索等技术领域,可以应用于从不同小说来源找出相同小说的相关场景。具体实现方案为:将目标图书中有标题主体的至少一个章节标题,与图书聚类集合中的图书的目标章节标题进行相似度匹配,得到对应的标题相似度匹配结果,目标章节标题包括:图书中有标题主体的章节标题;将与目标图书中无标题主体的至少一个章节正文,与图书的目标章节正文进行相似度匹配,得到对应的正文相似度匹配结果;以及基于标题相似度匹配结果和正文相似度匹配结果,对目标图书进行图书聚类。



CN 113408660 B

1.一种图书聚类方法,包括:

将目标图书中有标题主体的至少一个章节标题,与图书聚类集合中的图书的目标章节标题进行相似度匹配,得到对应的标题相似度匹配结果,所述目标章节标题包括:所述图书中有标题主体的章节标题;

将与所述目标图书中无标题主体的至少一个章节正文,与所述图书的目标章节正文进行相似度匹配,得到对应的正文相似度匹配结果;以及

基于所述标题相似度匹配结果和所述正文相似度匹配结果,对所述目标图书进行图书聚类;

其中,将与所述目标图书中无标题主体的至少一个章节正文,与所述图书的目标章节正文进行相似度匹配,包括:

针对所述目标图书,获取所述至少一个章节正文中每个章节正文中的至少一个长句对应的至少一个转换值;

针对所述图书,获取所述目标章节正文中每个章节正文中的至少一个长句对应的至少一个转换值;以及

将针对所述目标图书中每个无标题主体的章节正文获得的至少一个转换值,与基于所述图书中每个目标章节正文获得的至少一个转换值进行相似度匹配;

其中,基于所述目标图书和所述图书分别获得的长句均不包括预先设定的无效长句。

2.根据权利要求1所述的方法,其中,所述将目标图书中有标题主体的至少一个章节标题,与图书聚类集合中的图书的目标章节标题进行相似度匹配,包括:

针对所述目标图书,获取所述至少一个章节标题中每个章节标题的标题主体;

针对所述图书,获取所述目标章节标题中每个章节标题的标题主体;以及

将基于所述目标图书获得的每个标题主体与基于所述图书获得的每个标题主体进行相似度匹配。

3.根据权利要求1所述的方法,还包括:在针对所述目标图书进行标题相似度匹配之前,

基于图书标签,将所述目标图书与所述图书聚类集合中的图书进行相似度匹配,得到对应的标签相似度匹配结果,其中,所述图书标签包括书名信息和/或作者信息;

其中,在所述标签相似度匹配结果表征所述目标图书与所述图书聚类集合中的图书相似的情况下,执行针对所述目标图书进行标题相似度匹配的相关操作。

4.一种图书聚类装置,包括:

章节标题相似度匹配模块,用于将目标图书中有标题主体的至少一个章节标题,与图书聚类集合中的图书的目标章节标题进行相似度匹配,得到对应的标题相似度匹配结果,所述目标章节标题包括:所述图书中有标题主体的章节标题;

章节正文相似度匹配模块,用于将与所述目标图书中无标题主体的至少一个章节正文,与所述图书的目标章节正文进行相似度匹配,得到对应的正文相似度匹配结果;以及

图书聚类模块,用于基于所述标题相似度匹配结果和所述正文相似度匹配结果,对所述目标图书进行图书聚类;

其中,所述章节正文相似度匹配模块包括:

第五获取单元,用于针对所述目标图书,获取所述至少一个章节正文中每个章节正文

中的至少一个长句对应的至少一个转换值;

第六获取单元,用于针对所述图书,获取所述目标章节正文中每个章节正文中的至少一个长句对应的至少一个转换值;以及

正文长句转换值相似度匹配单元,用于将针对所述目标图书中每个无标题主体的章节正文获得的至少一个转换值,与基于所述图书中每个目标章节正文获得的至少一个转换值进行相似度匹配;

其中,基于所述目标图书和所述图书分别获得的长句均不包括预先设定的无效长句。

5.根据权利要求4所述的装置,其中,所述章节标题相似度匹配模块包括:

第一获取单元,用于针对所述目标图书,获取所述至少一个章节标题中每个章节标题的标题主体;

第二获取单元,用于针对所述图书,获取所述目标章节标题中每个章节标题的标题主体;以及

标题主体相似度匹配单元,用于将基于所述目标图书获得的每个标题主体与基于所述图书获得的每个标题主体进行相似度匹配。

6.根据权利要求4所述的装置,还包括:

图书标签相似度匹配模块,用于在所述章节标题相似度匹配模块针对所述目标图书进行标题相似度匹配之前,基于图书标签,将所述目标图书与所述图书聚类集合中的图书进行相似度匹配,得到对应的标签相似度匹配结果,其中,所述图书标签包括书名信息和/或作者信息;

其中,在所述标签相似度匹配结果表征所述目标图书与所述图书聚类集合中的图书相似的情况下,通过所述章节标题相似度匹配模块执行针对所述目标图书进行标题相似度匹配的相关操作。

7. 一种电子设备,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-3中任一项所述的方法。

8.一种存储有计算机指令的非瞬时计算机可读存储介质,其中,所述计算机指令用于使所述计算机执行根据权利要求1-3中任一项所述的方法。

9.一种计算机程序产品,包括计算机程序,所述计算机程序在被处理器执行时实现根据权利要求1-3中任一项所述的方法。

图书聚类方法、装置、设备和存储介质

技术领域

[0001] 本公开涉及互联网技术领域，尤其涉及大数据和智能搜索等技术领域，可以应用于从不同小说来源找出相同小说的相关场景。具体涉及一种图书聚类方法、装置、设备和存储介质。

背景技术

[0002] 目前，数字图书(简称图书)网站众多，离线处理图书信息时，通常需要按图书维度进行处理。比如，数字小说(又称网络小说，简称小说)网站众多，离线处理小说信息时，通常需要按小说维度进行处理，如将不同网站上挂载的相同小说聚类在一起进行处理。

发明内容

[0003] 本公开提供了一种图书聚类方法、装置、设备、存储介质以及计算机程序产品。

[0004] 根据本公开的一方面，提供了一种图书聚类方法，包括：将目标图书中有标题主体的至少一个章节标题，与图书聚类集合中的图书的目标章节标题进行相似度匹配，得到对应的标题相似度匹配结果，所述目标章节标题包括：所述图书中有标题主体的章节标题；将与所述目标图书中无标题主体的至少一个章节正文，与所述图书的目标章节正文进行相似度匹配，得到对应的正文相似度匹配结果；以及基于所述标题相似度匹配结果和所述正文相似度匹配结果，对所述目标图书进行图书聚类。

[0005] 根据本公开的另一方面，提供了一种图书聚类装置，包括：章节标题相似度匹配模块，用于将目标图书中有标题主体的至少一个章节标题，与图书聚类集合中的图书的目标章节标题进行相似度匹配，得到对应的标题相似度匹配结果，所述目标章节标题包括：所述图书中有标题主体的章节标题；章节正文相似度匹配模块，用于将与所述目标图书中无标题主体的至少一个章节正文，与所述图书的目标章节正文进行相似度匹配，得到对应的正文相似度匹配结果；以及图书聚类模块，用于基于所述标题相似度匹配结果和所述正文相似度匹配结果，对所述目标图书进行图书聚类。

[0006] 根据本公开的另一方面，提供了一种电子设备，包括：至少一个处理器；以及与所述至少一个处理器通信连接的存储器；其中，所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行本公开实施例所述的方法。

[0007] 根据本公开的另一方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，其中，所述计算机指令用于使所述计算机执行根据本公开实施例所述的方法。

[0008] 根据本公开的另一方面，提供了一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现根据本公开实施例所述的方法。

[0009] 应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0010] 附图用于更好地理解本方案,不构成对本公开的限定。其中:

[0011] 图1示例性示出了适于本公开实施例的系统架构;

[0012] 图2示例性示出了根据本公开实施例的图书聚类方法的流程图;

[0013] 图3示例性示出了根据本公开实施例的图书聚类的示意图;

[0014] 图4示例性示出了根据本公开实施例的基于二分图进行相似度判断的示意图;

[0015] 图5示例性示出了根据本公开实施例的图书聚类装置的框图;以及

[0016] 图6示例性示出了用来实现本公开实施例的图书聚类方法的电子设备的框图。

具体实施方式

[0017] 以下结合附图对本公开的示范性实施例做出说明,其中包括本公开实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本公开的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0018] 相关技术中,通常简单地根据书名和作者对网络小说进行聚类分析。但是,小说网站中,书名和作者命名并不规范。比如,某些小说网站为了提高网络小说被检索到的概率,可能会将书名改为主角名。因此,简单地通过书名和作者进行小说聚类,会导致很多相同的网络小说无法聚类在一起。

[0019] 对此,本公开实施例提供了一种改进后的图书聚类方法,该方法通过对各图书的章节标题和章节正文进行相似度联合匹配,可以尽最大可能地将相同的图书聚类在一起,比如可以将书名和/或作者不同,但实质内容相同的图书聚类在一起,因而可以提高图书聚类效果。

[0020] 在本公开实施例中,所谓联合匹配,可以理解为,在对图书进行聚类时,对于有章节标题的章节,可以基于章节标题进行相似度判断;对于无章节标题的章节,可以基于章节正文进行相似度判断。最后,可以联合章节标题相似度判断结果和章节正文相似度判断结果这两部分来判断两本图书是否是同一本图书,从而将相同的图书聚合到一起。

[0021] 以下将结合附图和具体实施例详细阐述本公开。

[0022] 适于本公开实施例的图书聚类方法和装置的系统架构介绍如下。

[0023] 图1示例性示出了适于本公开实施例的系统架构。需要注意的是,图1所示仅为可以应用本公开实施例的系统架构的示例,以帮助本领域技术人员理解本公开的技术内容,但并不意味着本公开实施例不可以用于其他环境或场景。

[0024] 如图1所示,系统架构100可以包括:服务器101,阅读终端102、103、104,网站A、网站B、网站C。

[0025] 应该理解,市面上小说网站众多,比如网站A、网站B、网站C都可以是小说网站。这些网站提供的小说文本质量可能参差不齐,比如网站A提供的某小说文本只有前3章,网站B提供的该小说文本章节顺序混乱,网站C提供的该小说文本章节存在重复现象,等等,这些都会影响用户的阅读体验。

[0026] 在本公开实施例中,服务器101可以对多个网站(如网站A、网站B、网站C等)上挂载的小说进行聚类,从而将不同网站上的相同小说聚类在一起,即将不同网站上的多本相同

小说聚合在同一个小说聚类集合中。同时,服务器101还可以基于同一小说聚类集合中的多本相同小说,提供高质量的小说转码服务,使得用户可以读到质量更好的小说版本,提高用户的阅读体验。比如,在进行小说转码时,可以获取集合中多本小说的不同章节的内容,然后组合成一本内容相对较为完整且不存在章节重复和乱序等问题的小说文本,最后响应于用户的访问请求,如阅读终端102、103、104中的任意一个或多个发起的访问请求,对组合得到的小说文本进行转码并反馈给用户阅读,以提高转码小说的质量,同时提升用户的阅读体验。

[0027] 应该理解,图1中的网站、服务器和阅读终端的数目仅仅是示意性的。根据实现需要,可以具有任意数目的网站、服务器和阅读终端。

[0028] 适于本公开实施例的图书聚类方法和装置的应用场景介绍如下。

[0029] 应该理解,本公开实施例提供的图书聚类方法和装置可以用于对任意类型的图书进行聚类,本公开在此不做限定。

[0030] 以小说为例,本公开实施例提供的图书聚类方法和装置可以应用于小说转码、智能搜索、图书查重等场景,本公开在此也不做限定。

[0031] 根据本公开的实施例,本公开提供了一种图书聚类方法。

[0032] 图2示例性示出了根据本公开实施例的图书聚类方法的流程图。

[0033] 如图2所示,图书聚类方法200可以包括:操作S210～S230。

[0034] 在操作S210,将目标图书中有标题主体的至少一个章节标题,与图书聚类集合中的图书的目标章节标题进行相似度匹配,得到对应的标题相似度匹配结果,目标章节标题包括:图书中有标题主体的章节标题。

[0035] 在操作S220,将与目标图书中无标题主体的至少一个章节正文,与图书的目标章节正文进行相似度匹配,得到对应的正文相似度匹配结果。

[0036] 在操作S230,基于标题相似度匹配结果和正文相似度匹配结果,对目标图书进行图书聚类。

[0037] 在本公开的一些实施例中,对于任意一本图书而言,可以基于其所有章节执行上述操作,以实现图书聚类。或者,在本公开的另一些实施例中,对于任意一本图书而言,可以基于其部门章节(如前N章,N为整数)执行上述操作,以实现图书聚类。

[0038] 示例性的,在本公开实施例中,可以获取目标图书的前N章的章节标题,并将这些章节标题分为两类,即有标题主体的章节标题和无标题主体的章节标题两类。然后,对目标图书中有标题主体的章节标题,执行操作S210,得到关于目标图书的标题相似度匹配结果。对目标图书中无标题主体的章节正文,执行操作S220,得到关于目标图书的正文相似度匹配结果。最后,执行操作S230,同时利用目标图书的标题相似度匹配结果及其正文相似度匹配结果,对目标图书进行图书聚类。

[0039] 应该理解,网络上可以存在一些特殊图书,比如每个章节的章节标题都有标题主体的图书(称为第一类图书)或者每个章节的章节标题都没有标题主体的图书(称为第二类图书)。

[0040] 在本公开实施例中,对于第一类图书,实际上执行上述的操作S210可以正常得到对应的标题相似度匹配结果,而执行上述的操作S220实际上无法正常得到对应的正文相似度匹配结果,因而这种情况下,在执行上述的操作S230进行图书聚类时,实际上只是基于标

题相似度匹配结果进行的。

[0041] 类似地,在本公开实施例中,对于第二类图书,实际上执行上述的操作S210无法正常得到对应的标题相似度匹配结果,而执行上述的操作S220实际上可以正常得到对应的正文相似度匹配结果,因而这种情况下,在执行上述的操作S230进行图书聚类时,实际上只是基于正文相似度匹配结果进行的。

[0042] 应该理解,简单地通过书名和作者进行图书聚类,会导致很多相同的网络小说无法聚类在一起,因而聚类结果准确性不高。

[0043] 还应该理解,很多图书,尤其是网络小说的字数通常很多,动辄上万字,因此,在通过图书聚类寻找相同图书的过程中,如果直接对图书全文进行内容相似度匹配,则计算量很大,且计算耗时太长。

[0044] 因而,本公开实施例中,采用基于图书的章节标题和章节正文进行相似度联合匹配的方式进行图书聚类,可以保证聚类结果的准确性,尽最大可能地将相同的图书聚类在一起,比如可以将书名和/或作者不同,但实质内容相同的图书聚类在一起,因而可以提高图书聚类效果,同时还可以兼顾图书聚类的处理速度。

[0045] 作为一种可选的实施例,将目标图书中有标题主体的至少一个章节标题,与图书聚类集合中的图书的目标章节标题进行相似度匹配,可以包括如下操作。

[0046] 针对目标图书,获取其中有标题主体的至少一个章节标题中每个章节标题的标题主体。

[0047] 针对图书聚类集合中的图书,获取其中有标题主体的目标章节标题中每个章节标题的标题主体。

[0048] 将基于目标图书获得的每个标题主体与基于图书聚类集合中的图书获得的每个标题主体进行相似度匹配。

[0049] 在本公开实施例中,对于目标图书中所有有标题主体的章节标题,可以获取其中每个章节标题的标题主体。类似地,对于图书聚类集合中的图书,也可以针对该图书中所有有标题主体的章节标题,获取每个章节标题的标题主体。最后将目标图书的每个标题主体与图书聚类集合中该图书的每个标题主体一一进行相似度匹配,得到对应的标题相似度匹配结果。

[0050] 在本公开实施例中,可以通过过滤冗余信息来清洗章节标题,进而提取每个章节标题中的标题主体。

[0051] 进一步,上述的冗余信息可以包括但不限于章节标题中的以下信息中的一种或多种:标点符合,杂质信息(如书名、作者、空格、无效字符等),标题前缀和标题后缀。

[0052] 此外,在本公开的其他实施例中,提取标题主体时,如果章节标题中有标点符合,除了可以进行通过过滤冗余信息去除之外,可以进行全角转半角或者半角转全角变换,以保证目标图书的章节标题中的标点符合与图书聚类集合中的图书的章节标题中的标点符合格式一致。

[0053] 应该理解,在本公开实施例中,标题前缀可以包括章节前的序号信息。

[0054] 示例性的,在本公开实施例中,对于标题前缀,可以通过正则表达式或公共前缀过滤来去除。

[0055] 应该理解,公共前缀过滤包括如下操作:将章节标题中的数字先统一改写为0;然

后建立Tire树;然后对于出现次数超过预定次数(如10次)的公共前缀,进行过滤去除;过滤完成后,将标题中剩余的0再还原为原始数字。

[0056] 还应该理解,上述的标题后缀包括标题无效后缀。标题无效后缀过滤包括如下操作:如果章节标题后有括号,先去掉括号及其中的内容,再对比前、后章节的标题;如果此种情况下前、后章节的标题相同,则认为此种标题后缀是有效后缀,不进行过滤;否则,如果此种情况下前、后章节的标题不同,则认为括号及其中的内容为标题无效后缀,需要过滤掉。

[0057] 过滤掉标题前缀、标题无效后缀以及标点符合和杂质信息后,余下的部分则为章节标题的标题主体。

[0058] 示例性的,如果章节标题为“第一章金莲火树(求月票~)”,则按照本公开实施例提供的上述操作,提取出的标题主体应为“金莲火树”。

[0059] 需要说明的是,在本公开实施例中,对于部分图书而言,如果章节标题类似于“第1章”,则可以认为其没有标题主体,对于这种没有标题主体的章节,可以依靠对应的章节正文进行相似度匹配。

[0060] 通过本公开实施例,使用标题主体代替章节标题本身进行标题相似度匹配,可以避免因章节标题中的标点符合、杂质信息、标题前缀和标题后缀等信息干扰而导致误判,进而影响图书聚类效果。

[0061] 作为一种可选的实施例,将与目标图书中无标题主体的至少一个章节正文,与图书的目标章节正文进行相似度匹配,可以包括如下操作。

[0062] 针对目标图书,获取其中无章节标题的至少一个章节正文中每个章节正文中的至少一个长句。

[0063] 针对图书聚类集合中的图书,获取其中的目标章节正文(目标章节正文可以有标题主体或无标题主体,本公开实施例在此不做限定)中每个章节正文中的至少一个长句。

[0064] 将基于目标图书获得的每个章节正文中的至少一个长句,与基于该图书获得的每个章节正文中的至少一个长句进行相似度匹配。

[0065] 即,在本公开的一些实施例中,在对目标图书的相关章节正文与图书聚类集合中的图书的相关章节正文进行相似度匹配时,可以使用正文中的一个或者多个长句(如使用正文中的top K长句)代替正文的全文内容进行相似度匹配。采用该方法可以进一步提高图书聚类的处理速度,同时还可以兼顾聚类结果的准确性。

[0066] 应该理解,在本公开实施例中,一个章节中的top K长句,可以理解为,该章节正文中,长度排名在前K位的K个句子。

[0067] 在本公开实施例中,可以通过预设标点符号(如句号、问号等)切分章节正文,并从章节正文中选取长度排名前K位的K个句子作为本章节正文top K长句。

[0068] 通过本公开实施例,使用章节正文中的top K长句代替章节正文本身进行正文相似度匹配,可以避免减少计算量,进一步提高图书聚类的处理速度。

[0069] 此外,通过本公开实施例,使用章节正文中的top K长句而不是短句来代替章节正文本身进行正文相似度匹配,这是因为实际实验发现,取top K长句可以取得较好的识别效果,取短句更容易产生误报。

[0070] 此外,与对正文内容切词,生成内容的指纹或特征向量,然后基于指纹或特征向量进行正文相似度匹配相比,本公开实施例采用的基于top K长句进行正文相似度匹配,可以

尽量避免产生误报。这是因为,基于内容切词进行正文相似度匹配方案,更适合用于对比图书语意的相似度。而当前小说内容普遍同质化,因此,通过内容切词进行正文相似度匹配,容易产生误报。比如,两本小说如果仅仅桥段类似,则通过该相似度匹配方法,可能被误认为是同一本小说。

[0071] 或者,作为一种可选的实施例,将与目标图书中无标题主体的至少一个章节正文,与图书的目标章节正文进行相似度匹配,可以包括如下操作。

[0072] 针对目标图书,获取其中无章节标题的至少一个章节正文中每个章节正文中的至少一个长句对应的至少一个转换值。

[0073] 针对图书聚类集合中的图书,获取其中的目标章节正文(目标章节正文可以有标题主体或无标题主体,本公开实施例在此不做限定)中每个章节正文中的至少一个长句对应的至少一个转换值。

[0074] 将针对目标图书中每个无标题主体的章节正文获得的至少一个转换值,与基于图书中每个目标章节正文获得的至少一个转换值进行相似度匹配。

[0075] 即,在本公开的另一些实施例中,在对目标图书的相关章节正文与图书聚类集合中的图书的相关章节正文进行相似度匹配时,可以使用正文中的一个或者多个长句(如使用正文中的top K长句)的对应转换值(如哈希值等)代替正文的全文内容或者该一个或者多个长句进行正文相似度匹配。采用该方法可以进一步提高图书聚类的处理速度,同时还可以兼顾聚类结果的准确性。

[0076] 并且,采用该方法,可以不必维护每一章的top K长句集合,而是维护于top K长句对应的转换值集合即可,因而更便于存储和处理。

[0077] 在本公开实施例中,可以采用哈希变换等变换手段将相关长句变换为对应的转换值。

[0078] 作为一种可选的实施例,基于目标图书和图书分别获得的长句均不包括预先设定的无效长句。

[0079] 在本公开实施例中,可以维护一个常见的无效句子的集合,用于过滤明显的无效句子。示例性的,如果某一个句子,在多个章节中重复出现,则可以认为该句子为无效子句。比如,句子“本书最新章节内容未完,更多精彩内容手机请扫描下方二维码下载app”,可以为一个典型的无效子句。

[0080] 在本公开的一些实施例中,可以获取目标图书中的每个无章节标题的章节正文,然后通过标点符号(如句号、问号等),将这些章节正文切成多个句子,并过滤掉其中的无效句子,最后再按长度对每个章节正文中剩下的句子进行排序,取长度排位位于前K位的top K句子,即为该章节正文的top K句子。

[0081] 进一步,将新发现的无效句子加入无效句子的集合中后,还可以重新计算对应章节的top K子句,以便为后续相似度判断提供更准确的数据。

[0082] 通过本公开实施例,对相关章节正文进行无效句子过滤,可以避免因无效句子的干扰而导致误判,进而影响图书聚类效果。

[0083] 作为一种可选的实施例,该方法还可以包括:在针对目标图书进行标题相似度匹配之前,执行以下操作。

[0084] 基于图书标签,将目标图书与图书聚类集合中的图书进行相似度匹配,得到对应

的标签相似度匹配结果,其中,图书标签包括书名信息和/或作者信息。

[0085] 其中,在标签相似度匹配结果表征目标图书与图书聚类集合中的图书相似的情况下,执行针对目标图书进行标题相似度匹配的相关操作。

[0086] 应该理解,直接基于图书内容(包括章节标题和章节内容)进行图书聚合,则计算量相对较大。

[0087] 因而,在本公开实施例中,在基于图书内容进行图书聚合之前,如在针对目标图书进行标题相似度匹配之前,可以先基于书名信息和/或作者信息等图书标签对图书进行相似度匹配,找出书名或作者相同的图书,然后在基于图书内容进行图书聚合阶段,仅仅对书名或作者相同的图书进行图书聚合即可。由此可以加快图书相似度的匹配速度,提高图书聚合效果。

[0088] 进一步,在本公开实施例中,在判断出目标图书的相关标题主体与图书聚类集合中的图书的相关标题主体是否相似,以及判断出目标图书的相关章节正文与图书聚类集合中的图书的相关章节正文是否相似之后,可以确定目标图书中相似章节的占比,进而根据该相似章节的占比确定目标图书是否与图书聚类集合中的图书相似。由此,最终可以将相似的图书(认为实际上是相同图书)聚合到同一个图书聚类集合,得到对应的聚类结果。

[0089] 示例性的,本公开的一个实施例中,图书聚类流程可以包括如下操作。

[0090] 遍历多个网站上挂载的所有图书,找出其中书名或作者名相同的图书作为候选图书。

[0091] 对候选图书进行两两判断,找出其中文本内容相似的图书作为相同图书,并添加两两图书相似的相关记录,最后根据相似性记录将相同的图书聚合在同一个图书聚类集合中。

[0092] 如图3所示,图书聚类集合301中的图书表示来自不同网站的相同图书,图书聚类集合302中的图书表示来自不同网站的另一相同图书,因此各图书聚类集合之间相互没有交集。此外,如图3所示,两个集合外边的图书与两个集合中的任一图书均不相同。此外,如果通过图书聚类发现,图书A与图书聚类集合301中的图书相似,同时图书A与图书聚类集合302中的图书也相似,则可以合并图书聚类集合301和图书聚类集合302为同一图书聚类集合。

[0093] 示例性的,上述的文本内容相似性判断方法具体可以如下。

[0094] 如图4所示,可以取图书401的前4章和图书402的前5章,并基于图书401的前4章和图书402的前5章组成如图所示的二分图。

[0095] 对于有标题主体的章节,比较这两本书的章节的标题主体。如果这两本书的相关章节的标题主体相同,则添加二分图的一个边。

[0096] 同理,对于无标题主体的章节,比较这两本书的章节的top K句子或者top K句子的转换值,如果相关章节中的top K句子或者top K句子的转换值有一半以上重合,则添加二分图的一个边。

[0097] 两本书的相似度=二分图的最大匹配数/两图书中的最小章节数。如果两本书的相似度超过某个阈值,则认为这两本书相似,即认为这两本书实际上是相同的书。

[0098] 如图4所示,图书401和图书402中章节数较小的是图书401,共4章。图4中二分图的最大匹配度为3。因此,图书401和图书402之间的相似度为(3/4)。假设相似度阈值为80%,

由于(3/4)小于80％,因此最终得出图书401和图书402不相似。即,图书401和图书402不属于相同的图书,应该聚合到两个不同的图书聚类集合中。

[0099] 需要说明的是,在本公开实施例中,对标题主体进行相似性比较时,可以比较两个标题主体是否完全相同,也可以通过莱文斯坦比等方法比较两个标题主体是否相似。如果两个标题主体的莱文斯坦比小于某个阈值,则认为这两者相似。

[0100] 通过本公开实施例,最终可以将众多网站上挂载的全部书籍聚合在多个图书聚类集合中,且这些集合之间互不相交。即,每一个集合中的图书,被认为其为相同的图书。由此,可以得到涉及众多网站的图书聚类结果。

[0101] 根据本公开的实施例,本公开还提供了一种图书聚类装置。

[0102] 图5示例性示出了根据本公开实施例的图书聚类装置的框图。

[0103] 如图5所示,图书聚类装置500可以包括:章节标题相似度匹配模块510、章节正文相似度匹配模块520和图书聚类模块530。

[0104] 章节标题相似度匹配模块510,用于将目标图书中有标题主体的至少一个章节标题,与图书聚类集合中的图书的目标章节标题进行相似度匹配,得到对应的标题相似度匹配结果,该目标章节标题包括:该图书中有标题主体的章节标题。

[0105] 章节正文相似度匹配模块520,用于将与该目标图书中无标题主体的至少一个章节正文,与该图书的目标章节正文进行相似度匹配,得到对应的正文相似度匹配结果。

[0106] 图书聚类模块530,用于基于该标题相似度匹配结果和该正文相似度匹配结果,对该目标图书进行图书聚类。

[0107] 作为一种可选的实施例,该章节标题相似度匹配模块包括:第一获取单元,用于针对该目标图书,获取该至少一个章节标题中每个章节标题的标题主体;第二获取单元,用于针对该图书,获取该目标章节标题中每个章节标题的标题主体;以及标题主体相似度匹配单元,用于将基于该目标图书获得的每个标题主体与基于该图书获得的每个标题主体进行相似度匹配。

[0108] 作为一种可选的实施例,该章节正文相似度匹配模块包括:第三获取单元,用于针对该目标图书,获取该至少一个章节正文中每个章节正文中的至少一个长句;第四获取单元,用于针对该图书,获取该目标章节正文中每个章节正文中的至少一个长句;以及正文长句相似度匹配单元,用于将基于该目标图书获得的每个章节正文中的至少一个长句,与基于该图书获得的每个章节正文中的至少一个长句进行相似度匹配。

[0109] 作为一种可选的实施例,该章节正文相似度匹配模块包括:第五获取单元,用于针对该目标图书,获取该至少一个章节正文中每个章节正文中的至少一个长句对应的至少一个转换值;第六获取单元,用于针对该图书,获取该目标章节正文中每个章节正文中的至少一个长句对应的至少一个转换值;以及正文长句转换值相似度匹配单元,用于将针对该目标图书中每个无标题主体的章节正文获得的至少一个转换值,与基于该图书中每个目标章节正文获得的至少一个转换值进行相似度匹配。

[0110] 作为一种可选的实施例,基于该目标图书和该图书分别获得的长句均不包括预先设定的无效长句。

[0111] 作为一种可选的实施例,该装置还包括:该图书标签相似度匹配模块,用于在该章节标题相似度匹配模块针对该目标图书进行标题相似度匹配之前,基于图书标签,将该目

标图书与该图书聚类集合中的图书进行相似度匹配,得到对应的标签相似度匹配结果,其中,该图书标签包括书名信息和/或作者信息;其中,在该标签相似度匹配结果表征该目标图书与该图书聚类集合中的图书相似的情况下,通过该章节标题相似度匹配模块执行针对该目标图书进行标题相似度匹配的相关操作。

[0112] 应该理解,本公开装置部分的实施例与本公开方法部分的实施例对应相同或类似,所解决的技术问题和所达到的技术效果也对应相同或类似,本公开在此不再赘述。

[0113] 根据本公开的实施例,本公开还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

[0114] 图6示出了可以用来实施本公开的实施例的示例电子设备600的示意性框图。电子设备旨在表示各种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机和其它适合的计算机。电子设备还可以表示各种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的和/或者要求的本公开的实现。

[0115] 如图6所示,电子设备600包括计算单元601,其可以根据存储在只读存储器(ROM)602中的计算机程序或者从存储单元608加载到随机访问存储器(RAM)603中的计算机程序,来执行各种适当的动作和处理。在RAM 603中,还可存储电子设备600操作所需的各种程序和数据。计算单元601、ROM 602以及RAM 603通过总线604彼此相连。输入/输出(I/O)接口605也连接至总线604。

[0116] 电子设备600中的多个部件连接至I/O接口605,包括:输入单元606,例如键盘、鼠标等;输出单元607,例如各种类型的显示器、扬声器等;存储单元608,例如磁盘、光盘等;以及通信单元609,例如网卡、调制解调器、无线通信收发机等。通信单元609允许设备600通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0117] 计算单元601可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元601的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。计算单元601执行上文所描述的各个方法和处理,例如图书聚类方法。例如,在一些实施例中,图书聚类方法可被实现为计算机软件程序,其被有形地包含于机器可读介质,例如存储单元608。在一些实施例中,计算机程序的部分或者全部可以经由ROM 602和/或通信单元609而被载入和/或安装到设备600上。当计算机程序加载到RAM 603并由计算单元601执行时,可以执行上文描述的图书聚类方法的一个或多个步骤。备选地,在其他实施例中,计算单元601可以通过其他任何适当的方式(例如,借助于固件)而被配置为执行图书聚类方法。

[0118] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出

装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0119] 用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器或控制器,使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行,作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0120] 在本公开的上下文中,机器可读介质可以是有形的介质,其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备,或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0121] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0122] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0123] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器,又称为云计算服务器或云主机,是云计算服务体系中的一项主机产品,以解决了传统物理主机与VPS服务("Virtual Private SerVer",或简称"VPS")中,存在的管理难度大,业务扩展性弱的缺陷。服务器也可以为分布式系统的服务器,或者是结合了区块链的服务器。

[0124] 本公开的技术方案中,所涉及的图书数据的记录,存储和应用等,均符合相关法律法规的规定,且不违背公序良俗。

[0125] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本发公开中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本公开公开的技术方案所期望的结果,本文在此不进行限制。

[0126] 上述具体实施方式,并不构成对本公开保护范围的限制。本领域技术人员应该明白的是,根据设计要求和其他因素,可以进行各种修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进等,均应包含在本公开保护范围之内。

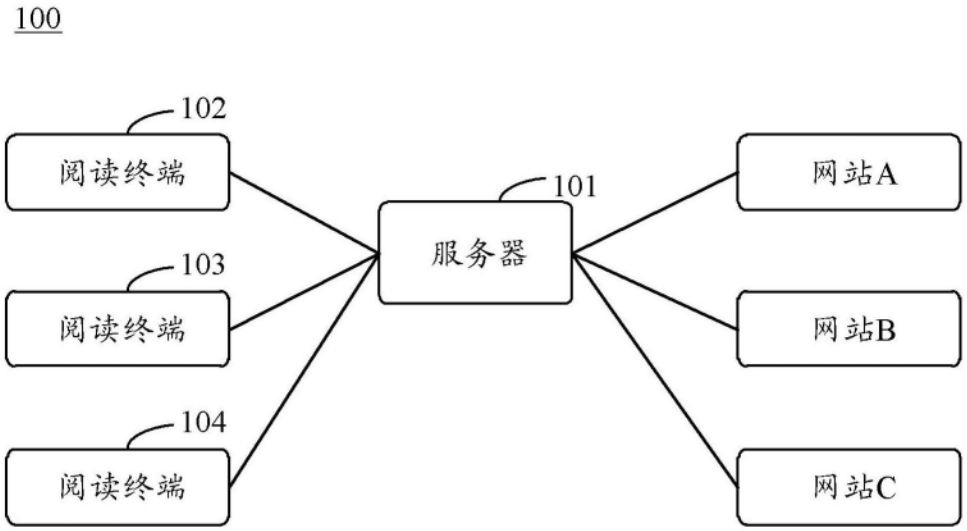

100
102
阅读终端
103
阅读终端
104
阅读终端
101
服务器
网站A
网站B
网站C

图1

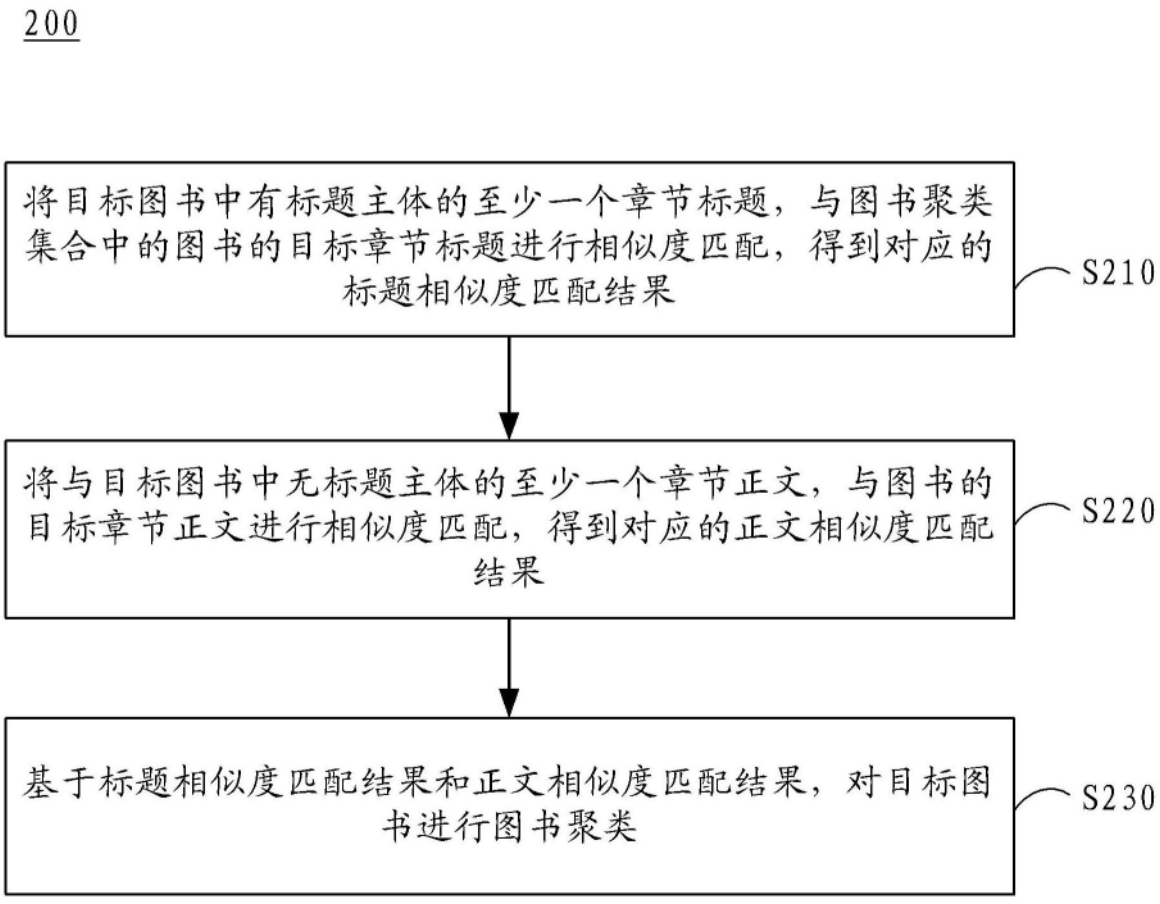

200
将目标图书中有标题主体的至少一个章节标题，与图书聚类集合中的图书的目标章节标题进行相似度匹配，得到对应的标题相似度匹配结果
S210
将与目标图书中无标题主体的至少一个章节正文，与图书的目标章节正文进行相似度匹配，得到对应的正文相似度匹配结果
S220
基于标题相似度匹配结果和正文相似度匹配结果，对目标图书进行图书聚类
S230

图2

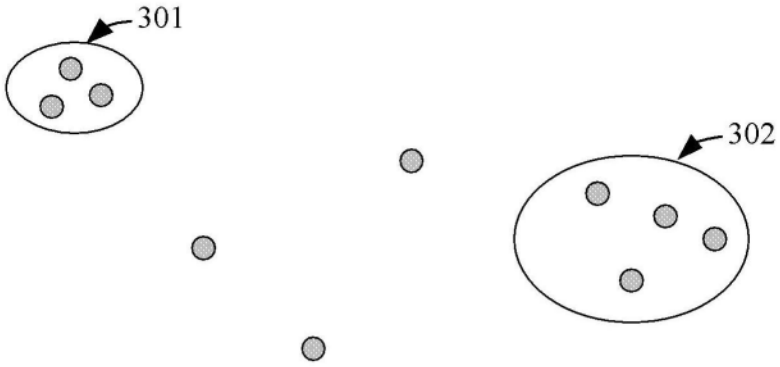

301
302

图3

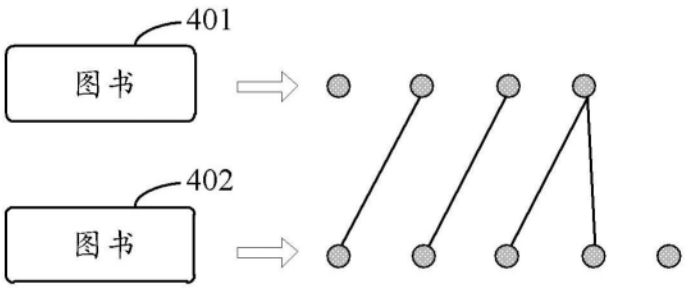

401
图书
402
图书

图4

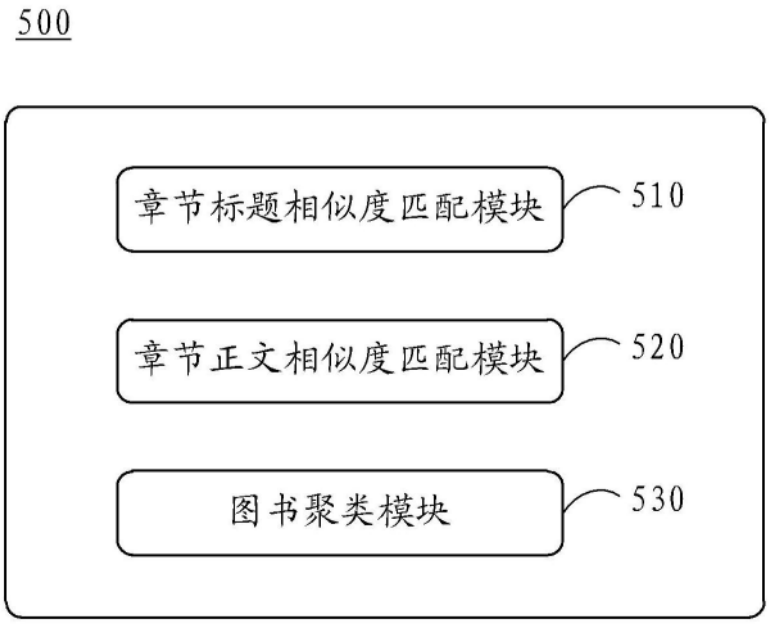

500
章节标题相似度匹配模块
510
章节正文相似度匹配模块
520
图书聚类模块
530

图5

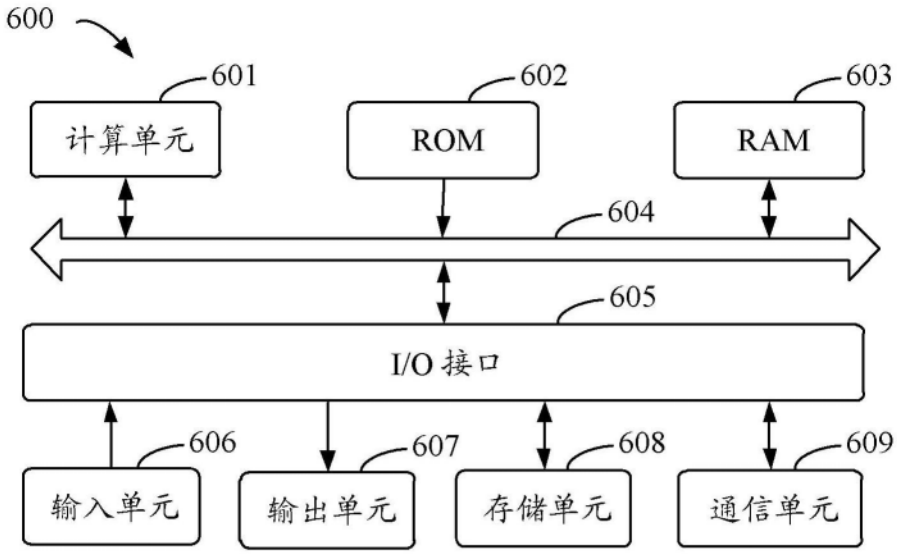

600
601
计算单元
602
ROM
603
RAM
604
605
I/O 接口
606
输入单元
607
输出单元
608
存储单元
609
通信单元

图6