(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114880098 A
(43)申请公布日 2022.08.09

(21)申请号 202210545419.X

(22)申请日 2022.05.19

(71)申请人 中国银行股份有限公司
地址 100818 北京市西城区复兴门内大街1号

(72)发明人 叶小谋 郑友杰 徐玉龙 郭凌星 叶雨凡

(74)专利代理机构 北京三友知识产权代理有限公司 11127
专利代理师 杨丹 沈珍珠

(51)Int.Cl.
G06F 9/48(2006.01)
G06F 16/33(2019.01)

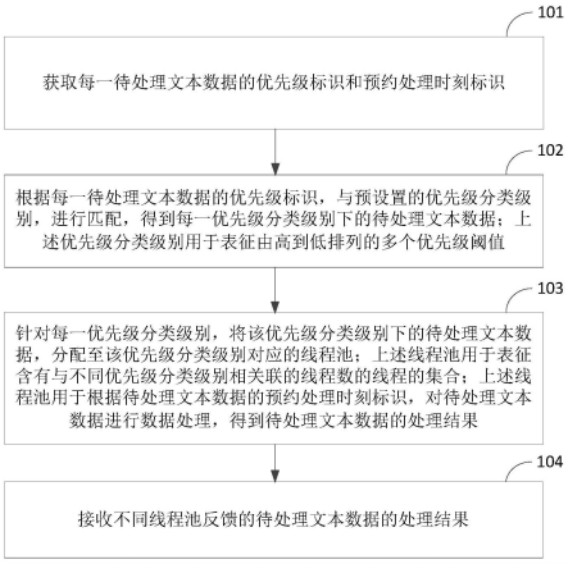
权利要求书2页 说明书9页 附图3页

(54)发明名称

文本数据的批量处理方法及装置

(57)摘要

本发明公开了一种文本数据的批量处理方法及装置，涉及大数据技术领域，该方法包括：根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据；所述优先级分类级别用于表征由高到低排列的多个优先级阈值；针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池；所述线程池用于根据待处理文本数据的预约处理时刻标识，对待处理文本数据进行数据处理，得到待处理文本数据的处理结果；接收不同线程池反馈的待处理文本数据的处理结果。本发明可保证重要任务数据处理的时效性。



CN 114880098 A

1.一种文本数据的批量处理方法,其特征在于,包括:

获取每一待处理文本数据的优先级标识和预约处理时刻标识;

根据每一待处理文本数据的优先级标识,与预设置的优先级分类级别,进行匹配,得到每一优先级分类级别下的待处理文本数据;所述优先级分类级别用于表征由高到低排列的多个优先级阈值;

针对每一优先级分类级别,将该优先级分类级别下的待处理文本数据,分配至该优先级分类级别对应的线程池;所述线程池用于表征含有与不同优先级分类级别相关联的线程数的线程的集合;所述线程池用于根据待处理文本数据的预约处理时刻标识,对待处理文本数据进行数据处理,得到待处理文本数据的处理结果;

接收不同线程池反馈的待处理文本数据的处理结果。

2.如权利要求1所述的方法,其特征在于,获取每一待处理文本数据的优先级标识和预约处理时刻标识,包括:

从预定义的文本数据存储目录中,获取每一文本数据存储目录中的待处理文本数据;

获取每一待处理文本数据的优先级标识和预约处理时刻标识。

3.如权利要求1所述的方法,其特征在于,获取每一待处理文本数据的优先级标识,包括:

获取每一待处理文本数据的数据特征向量;

将所述数据特征向量,输入至文本数据优先级识别模型中,得到文本数据的优先级;所述文本数据优先级识别模型以数据特征向量为输入数据,以优先级为输出数据;所述文本数据优先级识别模型是根据文本数据的数据特征向量的历史数据,对神经网络网络模型进行训练和验证得到的。

4.如权利要求1所述的方法,其特征在于,所述线程池具体用于:

在待处理文本数据的预约处理时刻标识表示:该待处理文本数据已指定预约处理时刻时,在所述预约处理时刻,对该待处理文本数据进行数据处理,得到待处理文本数据的处理结果;

在待处理文本数据的预约处理时刻标识表示:该待处理文本数据未指定预约处理时刻时,根据待处理文本数据的接收顺序,对该待处理文本数据进行数据处理,得到待处理文本数据的处理结果。

5.如权利要求1所述的方法,其特征在于,还包括:

将接收的不同线程池反馈的待处理文本数据的处理结果,存储至不同线程池对应的存储目录,并记录对应不同处理结果的处理时刻、处理耗时、处理线程池和处理结果。

6.一种文本数据的批量处理装置,其特征在于,包括:

待处理文本数据信息获取模块,用于获取每一待处理文本数据的优先级标识和预约处理时刻标识;

优先级标识匹配模块,用于根据每一待处理文本数据的优先级标识,与预设置的优先级分类级别,进行匹配,得到每一优先级分类级别下的待处理文本数据;所述优先级分类级别用于表征由高到低排列的多个优先级阈值;

线程池分配模块,用于针对每一优先级分类级别,将该优先级分类级别下的待处理文本数据,分配至该优先级分类级别对应的线程池;所述线程池用于表征含有与不同优先级

分类级别相关联的线程数的线程的集合;所述线程池用于根据待处理文本数据的预约处理时刻标识,对待处理文本数据进行数据处理,得到待处理文本数据的处理结果;

处理结果接收模块,用于接收不同线程池反馈的待处理文本数据的处理结果。

7.如权利要求6所述的装置,其特征在于,待处理文本数据信息获取模块,具体用于:

从预定义的文本数据存储目录中,获取每一文本数据存储目录中的待处理文本数据;

获取每一待处理文本数据的优先级标识和预约处理时刻标识。

8.如权利要求6所述的装置,其特征在于,待处理文本数据信息获取模块,具体用于:

获取每一待处理文本数据的数据特征向量;

将所述数据特征向量,输入至文本数据优先级识别模型中,得到文本数据的优先级;所述文本数据优先级识别模型以数据特征向量为输入数据,以优先级为输出数据;所述文本数据优先级识别模型是根据文本数据的数据特征向量的历史数据,对神经网络网络模型进行训练和验证得到的。

9.如权利要求6所述的装置,其特征在于,所述线程池具体用于:

在待处理文本数据的预约处理时刻标识表示:该待处理文本数据已指定预约处理时刻时,在所述预约处理时刻,对该待处理文本数据进行数据处理,得到待处理文本数据的处理结果;

在待处理文本数据的预约处理时刻标识表示:该待处理文本数据未指定预约处理时刻时,根据待处理文本数据的接收顺序,对该待处理文本数据进行数据处理,得到待处理文本数据的处理结果。

10.如权利要求6所述的装置,其特征在于,还包括:

数据存储模块,用于:

将接收的不同线程池反馈的待处理文本数据的处理结果,存储至不同线程池对应的存储目录,并记录对应不同处理结果的处理时刻、处理耗时、处理线程池和处理结果。

11.一种计算机设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,其特征在于,所述处理器执行所述计算机程序时实现权利要求1至5任一所述方法。

12.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质存储有计算机程序,所述计算机程序被处理器执行时实现权利要求1至5任一所述方法。

13.一种计算机程序产品,其特征在于,所述计算机程序产品包括计算机程序,所述计算机程序被处理器执行时实现权利要求1至5任一所述方法。

文本数据的批量处理方法及装置

技术领域

[0001] 本发明涉及大数据技术领域，尤其涉及文本数据的批量处理方法及装置。

背景技术

[0002] 本部分旨在为权利要求书中陈述的本发明实施例提供背景或上下文。此处的描述不因为包括在本部分中就承认是现有技术。

[0003] 越来越多业务的文本数据利用计算机进行数据处理，而目前进行文本数据批量处理时多采用多线程同步处理，但是当数据量不断增加时，会导致数据处理的时效性不够，尤其对于比较重要的任务，可能会导致处理滞后。

[0004] 目前采用多线程虽然可以提升文件处理速率，但是这样会有两个问题：

[0005] 1、无法区分文件处理的先后顺序，导致文本数据占用的系统资源无法区分，一些重要业务信息无法得到优先处理；

[0006] 2、无法指定处理时间，导致来了文件线程池立刻进行处理，因此有些时段可能会打扰到客户、或者处理文本数据的前置数据还未进行处理，导致处理出现错误。

[0007] 如何提出一种方案，能够保证正常的数据处理顺序，又可以提高关键性任务处理的时效性成为本领域亟需解决的技术问题。

发明内容

[0008] 本发明实施例提供一种文本数据的批量处理方法，用以保证重要任务数据处理的时效性，该方法包括：

[0009] 获取每一待处理文本数据的优先级标识和预约处理时刻标识；

[0010] 根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据；所述优先级分类级别用于表征由高到低排列的多个优先级阈值；

[0011] 针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池；所述线程池用于表征含有与不同优先级分类级别相关联的线程数的线程的集合；所述线程池用于根据待处理文本数据的预约处理时刻标识，对待处理文本数据进行数据处理，得到待处理文本数据的处理结果；

[0012] 接收不同线程池反馈的待处理文本数据的处理结果。

[0013] 本发明实施例还提供一种文本数据的批量处理装置，用以保证重要任务数据处理的时效性，该装置包括：

[0014] 待处理文本数据信息获取模块，用于获取每一待处理文本数据的优先级标识和预约处理时刻标识；

[0015] 优先级标识匹配模块，用于根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据；所述优先级分类级别用于表征由高到低排列的多个优先级阈值；

[0016] 线程池分配模块，用于针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池；所述线程池用于表征含有与不同优先级分类级别相关联的线程数的线程的集合；所述线程池用于根据待处理文本数据的预约处理时刻标识，对待处理文本数据进行数据处理，得到待处理文本数据的处理结果；

[0017] 处理结果接收模块，用于接收不同线程池反馈的待处理文本数据的处理结果。

[0018] 本发明实施例还提供一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现上述文本数据的批量处理方法。

[0019] 本发明实施例还提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述文本数据的批量处理方法。

[0020] 本发明实施例还提供一种计算机程序产品，所述计算机程序产品包括计算机程序，所述计算机程序被处理器执行时实现上述文本数据的批量处理方法。

[0021] 本发明实施例中，获取每一待处理文本数据的优先级标识和预约处理时刻标识；根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据；所述优先级分类级别用于表征由高到低排列的多个优先级阈值；针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池；所述线程池用于表征含有与不同优先级分类级别相关联的线程数的线程的集合；所述线程池用于根据待处理文本数据的预约处理时刻标识，对待处理文本数据进行数据处理，得到待处理文本数据的处理结果；接收不同线程池反馈的待处理文本数据的处理结果，与现有技术中仅以多线程进行批量处理的技术方案相比，通过将不同优先级标识的待处理文本数据分配至不同的线程池，使得对优先级高的文本数据，可使用线程多的线程池进行处理，可在保证文本数据的正常处理顺序时，保证重要任务可以尽可能优先处理，避免了重要任务被延后处理的问题，同时保证优先级低的任务也能够得到处理，保证了重要任务数据处理的时效性。

附图说明

[0022] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。在附图中：

[0023] 图1为本发明实施例中一种文本数据的批量处理方法的流程示意图；

[0024] 图2为本发明实施例中一种文本数据的批量处理方法的具体示例图；

[0025] 图3为本发明实施例中一种文本数据的批量处理方法的具体示例图；

[0026] 图4为本发明实施例中一种文本数据的批量处理装置的结构示例图；

[0027] 图5为本发明实施例中提供的一种计算机设备的示意图。

具体实施方式

[0028] 为使本发明实施例的目的、技术方案和优点更加清楚明白，下面结合附图对本发明实施例做进一步详细说明。在此，本发明的示意性实施例及其说明用于解释本发明，但并

不作为对本发明的限定。

[0029] 本文中术语“和/或”，仅仅是描述一种关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。另外，本文中术语“至少一种”表示多种中的任意一种或多种中的至少两种的任意组合，例如，包括A、B、C中的至少一种，可以表示包括从A、B和C构成的集合中选择的任意一个或多个元素。

[0030] 在本说明书的描述中，所使用的“包含”、“包括”、“具有”、“含有”等，均为开放性的用语，即意指包含但不限于。参考术语“一个实施例”、“一个具体实施例”、“一些实施例”、“例如”等的描述意指结合该实施例或示例描述的具体特征、结构或者特点包含于本申请的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。各实施例中涉及的步骤顺序用于示意性说明本申请的实施，其中的步骤顺序不作限定，可根据需要作适当调整。

[0031] 越来越多业务的文本数据利用计算机进行数据处理，而目前进行文本数据批量处理时多采用多线程同步处理，但是当数据量不断增加时，会导致数据处理的时效性不够，尤其对于比较重要的任务，可能会导致处理滞后。

[0032] 批量文件解析是一种文件读取和处理方法。随着业务的发展，批量文本解析的任务越来越多。如何在有限处理能力下满足上述解析任务，保证业务正常运行。常规的做法是采用多线程加速文件处理。批量文本数据可为一组待处理数据按照某种固定格式进行组装而形成。

[0033] 采用多线程是可以提升文件处理速率，但是各个批量文本之间并没有优先级之分。这样会有两个问题：1.无法区分优先级。导致优先级不同的批量文本占用的系统资源无法区分，一些重要业务信息无法优先处理。2.无法指定处理时间。导致来了文件就进行处理，有些时段可能会打扰到客户。

[0034] 为了解决上述问题，本发明实施例提供了一种文本数据的批量处理方法，参见图1，该方法可以包括：

[0035] 步骤101：获取每一待处理文本数据的优先级标识和预约处理时刻标识；

[0036] 步骤102：根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据；上述优先级分类级别用于表征由高到低排列的多个优先级阈值；

[0037] 步骤103：针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池；上述线程池用于表征含有与不同优先级分类级别相关联的线程数的线程的集合；上述线程池用于根据待处理文本数据的预约处理时刻标识，对待处理文本数据进行数据处理，得到待处理文本数据的处理结果；

[0038] 步骤104：接收不同线程池反馈的待处理文本数据的处理结果。

[0039] 本发明实施例中，获取每一待处理文本数据的优先级标识和预约处理时刻标识；根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据；上述优先级分类级别用于表征由高到低排列的多个优先级阈值；针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池；上述线程池用于表征含有与不同优先级分类级别相关联

的线程数的线程的集合;上述线程池用于根据待处理文本数据的预约处理时刻标识,对待处理文本数据进行数据处理,得到待处理文本数据的处理结果;接收不同线程池反馈的待处理文本数据的处理结果,与现有技术中仅以多线程进行批量处理的技术方案相比,通过将不同优先级标识的待处理文本数据分配至不同的线程池,使得对优先级高的文本数据,可使用线程多的线程池进行处理,可在保证文本数据的正常处理顺序时,保证重要任务可以尽可能优先处理,避免了重要任务被延后处理的问题,同时保证优先级低的任务也能够得到处理,保证了重要任务数据处理的时效性。

[0040] 具体实施时,首先获取每一待处理文本数据的优先级标识和预约处理时刻标识。

[0041] 在一个实施例中,获取每一待处理文本数据的优先级标识和预约处理时刻标识,如图2所示,可以包括:

[0042] 步骤201:从预定义的文本数据存储目录中,获取每一文本数据存储目录中的待处理文本数据;

[0043] 步骤202:获取每一待处理文本数据的优先级标识和预约处理时刻标识。

[0044] 举一实例,待处理文本数据的文件命名规则可如下所示:

[0045] 来源系统标识(2).接口类型(4).序列号(8).上送文件日期(YYYYMMDDHHMM).预约处理时间(YYYYMMDDHHMM).优先级(3).DAT

[0046] 因此,对应该文件命名规则的待处理文本数据的文件名可如下所示:

[0047] BF.BAMP.00000001.202204122200.202204130800.001.DAT。

[0048] 获取每一待处理文本数据的优先级标识和预约处理时刻标识的处理流程,可示例如下:

[0049] 1.从指定目录扫描到该待处理文本数据,即待处理文件;

[0050] 2.对待处理文件的文件名进行解析,获取待处理文件的预约处理时刻标识以及优先级字段;

[0051] 以上面的文件名为例,此文件预约处理时刻标识为2022-04-13 08:00,优先级为1。

[0052] 在一个实施例中,待处理文本数据的优先级标识和预约处理时刻标识,可设置于待处理文本数据的文件名称中。而通过对文件名设置优先级和预约处理时间。有如下两个优点:第一,让优先级高的批量文本分配更多的线程资源处理,实现优先的业务数据优先处理。第二,同根据预约处理时间灵活指定文本处理时间,避免打扰客户。

[0053] 实施例中,1.批量文本名指明了此文本处理的优先级,根据优先级分配处理线程数;2.批量文本名指明了此文本的预约处理时间,根据预约处理时间调度文本处理时间。

[0054] 在一个实施例中,获取每一待处理文本数据的优先级标识,可以包括:

[0055] 获取每一待处理文本数据的数据特征向量;

[0056] 将上述数据特征向量,输入至文本数据优先级识别模型中,得到文本数据的优先级;上述文本数据优先级识别模型以数据特征向量为输入数据,以优先级为输出数据;上述文本数据优先级识别模型是根据文本数据的数据特征向量的历史数据,对神经网络网络模型进行训练和验证得到的。

[0057] 实施例中,通过建立文本数据优先级识别模型,可准确根据待处理文本数据的数据特征向量,来确定文本数据的优先级。

[0058] 在一个实施例中，确定优先级可包括：一个是文件中名指定的优先级，一个批量文件的大小。

[0059] 举一例，假如文件名指定优先级是a，批量数据规模大小为b，则该文本数据的优先级可按如下公式计算：

[0060] $p=a\times\alpha+(b/c)\times\beta$

[0061] 其中，p为该文本数据的最终优先级，无量纲；a是该文本数据的优先级标识所表征的优先级具体数据，无量纲；α和β为比例参数，可根据实际使用场景进行灵活设置；b为该文本数据的数据规模大小；c为每分钟处理数据的数量，可根据实际使用场景进行灵活设置；如c可为40w，即系统一分钟处理文件的数据量为40万。

[0062] 借助上述公式，可确定文本数据的最终优先级，进而可根据此参数分配指定的线程数进行处理。

[0063] 具体实施时，在获取每一待处理文本数据的优先级标识和预约处理时刻标识后，根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据；上述优先级分类级别用于表征由高到低排列的多个优先级阈值。

[0064] 举一分配线程池的实例，如：

[0065] 系统中有高、中、低三个带有优先级的线程池，对应的最大核心线程数为70、20、10。p为文本数据的优先级，若$p>=70$，则分配到高优先级线程池；若$20<=p<70$，则分配到中优先级线程池；若$p<20$，则分配到低优先级线程池。

[0066] 具体实施时，在根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据后，针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池；上述线程池用于表征含有与不同优先级分类级别相关联的线程数的线程的集合；上述线程池用于根据待处理文本数据的预约处理时刻标识，对待处理文本数据进行数据处理，得到待处理文本数据的处理结果。

[0067] 实施例中，如图3所示，上述线程池具体用于：

[0068] 步骤301：在待处理文本数据的预约处理时刻标识表示：该待处理文本数据已指定预约处理时刻时，在上述预约处理时刻，对该待处理文本数据进行数据处理，得到待处理文本数据的处理结果；

[0069] 步骤302：在待处理文本数据的预约处理时刻标识表示：该待处理文本数据未指定预约处理时刻时，根据待处理文本数据的接收顺序，对该待处理文本数据进行数据处理，得到待处理文本数据的处理结果。

[0070] 举一实例，例如待处理文本数据的文件名为：BF.BAMP.00000001.202204122200.202204130800.001.DAT，通过获取该待处理文本数据的优先级标识和预约处理时刻标识，可得出：预约处理时刻标识为2022-04-13 08:00，则在这一具体时刻开始处理该待处理文本数据；

[0071] 而如待处理文本数据的文件名为：BF.BOMP.00000001.202204122200.000000000000.001.DAT，通过获取该待处理文本数据的优先级标识和预约处理时刻标识，可得出：预约处理时刻标识为12个0，则表明这个文件需要即刻处理。

[0072] 在上述实施例中，通过对批量文本引入优先级标识，当系统中检测到需要处理的批量文本时，可根据优先级标识分配不同的线程数，优先级高的批量文本多分配线程数进行处理。这样会保证优先级高的批量文本使用较多的系统资源进行处理。同时在文件名中指定文件预约处理时间，在解析时根据预约处理时间对批量文本进行调度。

[0073] 在上述实施例中，根据对批量文本预先设置优先级和预约处理时间，可对批量文本解析进行资源和时间维度的调度；同样，分配的线程池越多，文件处理的并发数更多，可以提升文件处理效率，减少文件处理时间；而待处理文件优先级越高，分配的线程池容量越大。线程池容量越大，任务并行数越高。预约处理时间，仅对任务执行时间进行规定，到了执行时间开始处理文本。

[0074] 具体实施时，在针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池后，接收不同线程池反馈的待处理文本数据的处理结果。

[0075] 具体实施时，本发明实施例提供的一种文本数据的批量处理方法，还可以包括：将接收的不同线程池反馈的待处理文本数据的处理结果，存储至不同线程池对应的存储目录，并记录对应不同处理结果的处理时刻、处理耗时、处理线程池和处理结果。

[0076] 实施例中，通过将接收的不同线程池反馈的待处理文本数据的处理结果，存储至不同线程池对应的存储目录，并记录对应不同处理结果的处理时刻、处理耗时、处理线程池和处理结果，有利于工作人员对文本数据的处理工程进行追溯，有利于保障数据安全。

[0077] 下面以本发明实施例中的一个具体实施例，来对本发明进行详细说明：

[0078] 1、批量文本扫描，根据文件目录获取文件，得到待处理批量文本文件名；

[0079] 2、批量文本预解析，根据文件名中优先级字段确定处理的分配的线程池大小；

[0080] 3、批量文本时间调度模块，根据预约处理时间和文件类型把解析任务提交给不同的线程；

[0081] 4、批量文本处理模块，根根解析规则，针对不同的文件类型采用不同的线程进行处理。

[0082] 当然，可以理解的是，上述详细流程还可以有其他变化例，相关变化例均应落入本发明的保护范围。

[0083] 本发明实施例中，获取每一待处理文本数据的优先级标识和预约处理时刻标识；根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据；上述优先级分类级别用于表征由高到低排列的多个优先级阈值；针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池；上述线程池用于表征含有与不同优先级分类级别相关联的线程数的线程的集合；上述线程池用于根据待处理文本数据的预约处理时刻标识，对待处理文本数据进行数据处理，得到待处理文本数据的处理结果；接收不同线程池反馈的待处理文本数据的处理结果，与现有技术中仅以多线程进行批量处理的技术方案相比，通过将不同优先级标识的待处理文本数据分配至不同的线程池，使得对优先级高的文本数据，可使用线程多的线程池进行处理，可在保证文本数据的正常处理顺序时，保证重要任务可以尽可能优先处理，避免了重要任务被延后处理的问题，同时保证优先级低的任务也能够得到处理，保证了重要任务数据处理的时效性。

[0084] 如上述，本发明实施例通过引入优先级和预约处理时间，可对批量文本进行优先级的标注。对于有优先级业务含义的批量文本，可以分配更多的处理资源，保证优先的业务优先处理。对于指定处理时间的文本，可以实施预约处理，灵活指定处理时间，避免打扰到客户。

[0085] 本发明实施例中还提供了一种文本数据的批量处理装置，如下面的实施例上述。由于该装置解决问题的原理与文本数据的批量处理方法相似，因此该装置的实施可以参见文本数据的批量处理方法的实施，重复之处不再赘述。

[0086] 本发明实施例还提供一种文本数据的批量处理装置，用以保证重要任务数据处理的时效性，如图4所示，该装置包括：

[0087] 待处理文本数据信息获取模块401，用于获取每一待处理文本数据的优先级标识和预约处理时刻标识；

[0088] 优先级标识匹配模块402，用于根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据；上述优先级分类级别用于表征由高到低排列的多个优先级阈值；

[0089] 线程池分配模块403，用于针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池；上述线程池用于表征含有与不同优先级分类级别相关联的线程数的线程的集合；上述线程池用于根据待处理文本数据的预约处理时刻标识，对待处理文本数据进行数据处理，得到待处理文本数据的处理结果；

[0090] 处理结果接收模块404，用于接收不同线程池反馈的待处理文本数据的处理结果。

[0091] 在一个实施例中，待处理文本数据信息获取模块，具体用于：

[0092] 从预定义的文本数据存储目录中，获取每一文本数据存储目录中的待处理文本数据；

[0093] 获取每一待处理文本数据的优先级标识和预约处理时刻标识。

[0094] 在一个实施例中，待处理文本数据信息获取模块，具体用于：

[0095] 获取每一待处理文本数据的数据特征向量；

[0096] 将上述数据特征向量，输入至文本数据优先级识别模型中，得到文本数据的优先级；上述文本数据优先级识别模型以数据特征向量为输入数据，以优先级为输出数据；上述文本数据优先级识别模型是根据文本数据的数据特征向量的历史数据，对神经网络网络模型进行训练和验证得到的。

[0097] 在一个实施例中，上述线程池具体用于：

[0098] 在待处理文本数据的预约处理时刻标识表示：该待处理文本数据已指定预约处理时刻时，在上述预约处理时刻，对该待处理文本数据进行数据处理，得到待处理文本数据的处理结果；

[0099] 在待处理文本数据的预约处理时刻标识表示：该待处理文本数据未指定预约处理时刻时，根据待处理文本数据的接收顺序，对该待处理文本数据进行数据处理，得到待处理文本数据的处理结果。

[0100] 在一个实施例中，还可以包括：

[0101] 数据存储模块，用于：

[0102] 将接收的不同线程池反馈的待处理文本数据的处理结果，存储至不同线程池对应

的存储目录，并记录对应不同处理结果的处理时刻、处理耗时、处理线程池和处理结果。

[0103] 下面给出一个具体实施例，来说明本发明的装置的具体应用，该实施例中，可以包括如下模块：

[0104] 批量文件扫描模块：用于根据事先定义的文件目录，获取待处理的批量文本。

[0105] 批量文本预解析模块：用于根据优先级对批量文本进行分类。

[0106] 批量文本时间调度模块：用于根据预约处理时间对批量文本进行调度。

[0107] 批量文本处处理模块：用于根据文本解析规则对文件进行处理。

[0108] 当然，可以理解的是，上述详细模块还可以有其他变化例，相关变化例均应落入本发明的保护范围。

[0109] 基于上述发明构思，如图5所示，本发明还提出了一种计算机设备500，包括存储器510、处理器520及存储在存储器510上并可在处理器520上运行的计算机程序530，所述处理器520执行所述计算机程序530时实现上述文本数据的批量处理方法。

[0110] 本发明实施例还提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述文本数据的批量处理方法。

[0111] 本发明实施例还提供一种计算机程序产品，所述计算机程序产品包括计算机程序，所述计算机程序被处理器执行时实现上述文本数据的批量处理方法。

[0112] 本发明实施例中，获取每一待处理文本数据的优先级标识和预约处理时刻标识；根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据；所述优先级分类级别用于表征由高到低排列的多个优先级阈值；针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池；所述线程池用于表征含有与不同优先级分类级别相关联的线程数的线程的集合；所述线程池用于根据待处理文本数据的预约处理时刻标识，对待处理文本数据进行数据处理，得到待处理文本数据的处理结果；接收不同线程池反馈的待处理文本数据的处理结果，与现有技术中仅以多线程进行批量处理的技术方案相比，通过将不同优先级标识的待处理文本数据分配至不同的线程池，使得对优先级高的文本数据，可使用线程多的线程池进行处理，可在保证文本数据的正常处理顺序时，保证重要任务可以尽可能优先处理，避免了重要任务被延后处理的问题，同时保证优先级低的任务也能够得到处理，保证了重要任务数据处理的时效性。

[0113] 本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0114] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0115] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0116] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0117] 以上所述的具体实施例，对本发明的目的、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，以上所述仅为本发明的具体实施例而已，并不用于限定本发明的保护范围，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

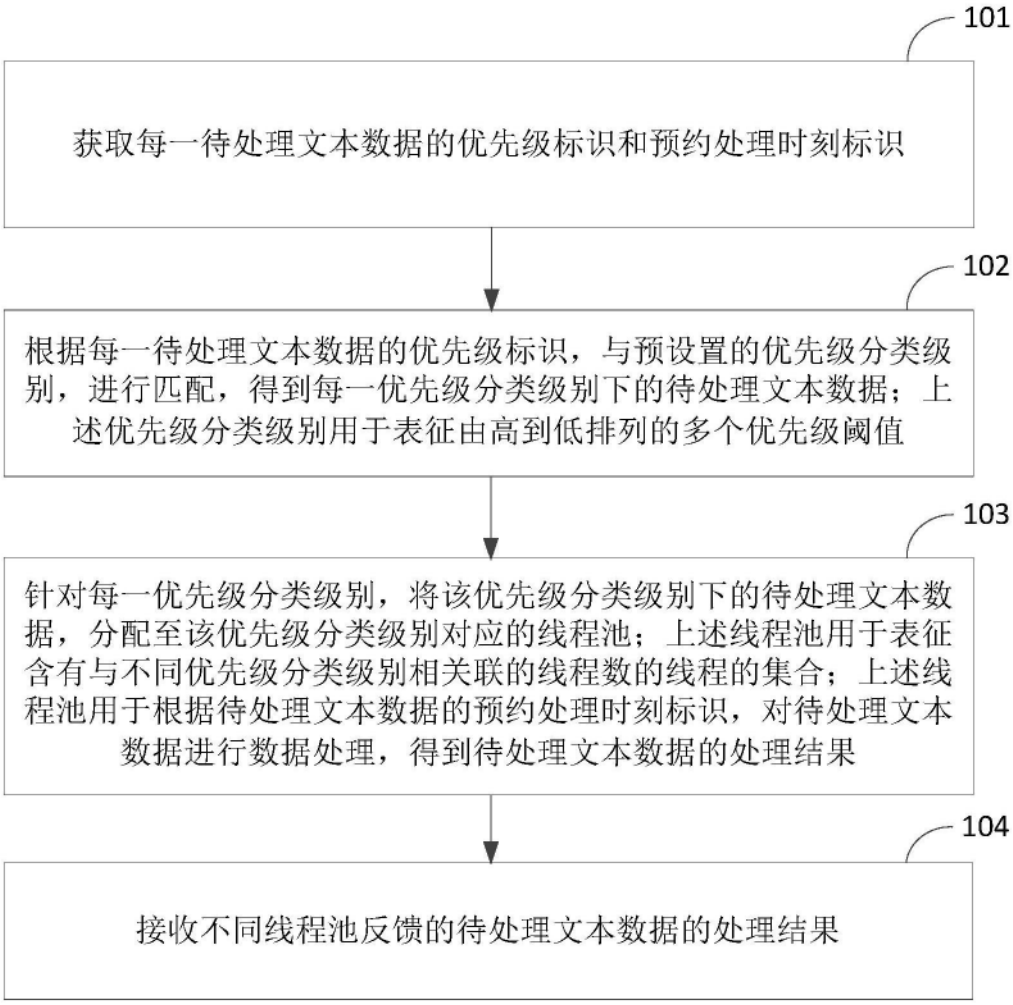

101
获取每一待处理文本数据的优先级标识和预约处理时刻标识
102
根据每一待处理文本数据的优先级标识，与预设置的优先级分类级别，进行匹配，得到每一优先级分类级别下的待处理文本数据；上述优先级分类级别用于表征由高到低排列的多个优先级阈值
103
针对每一优先级分类级别，将该优先级分类级别下的待处理文本数据，分配至该优先级分类级别对应的线程池；上述线程池用于表征含有与不同优先级分类级别相关联的线程数的线程的集合；上述线程池用于根据待处理文本数据的预约处理时刻标识，对待处理文本数据进行数据处理，得到待处理文本数据的处理结果
104
接收不同线程池反馈的待处理文本数据的处理结果

图1

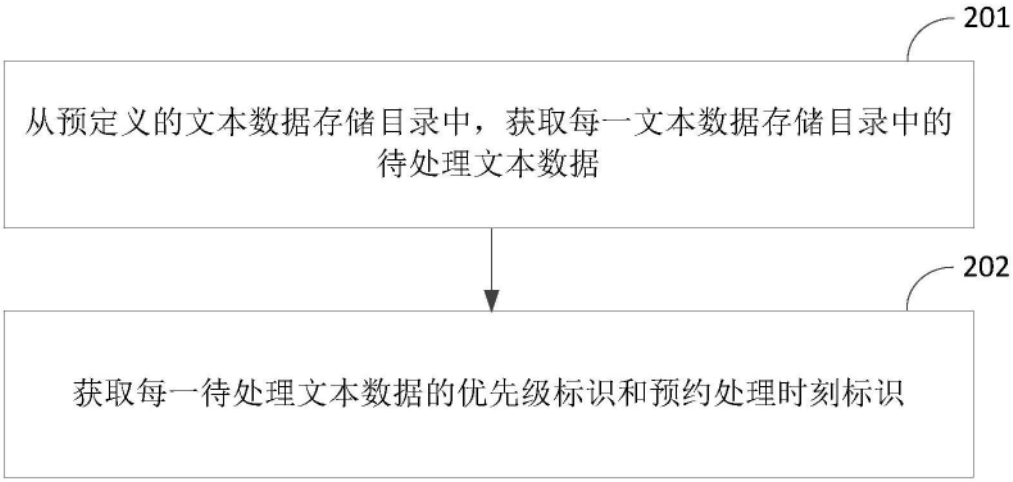

201
从预定义的文本数据存储目录中，获取每一文本数据存储目录中的待处理文本数据
202
获取每一待处理文本数据的优先级标识和预约处理时刻标识

图2

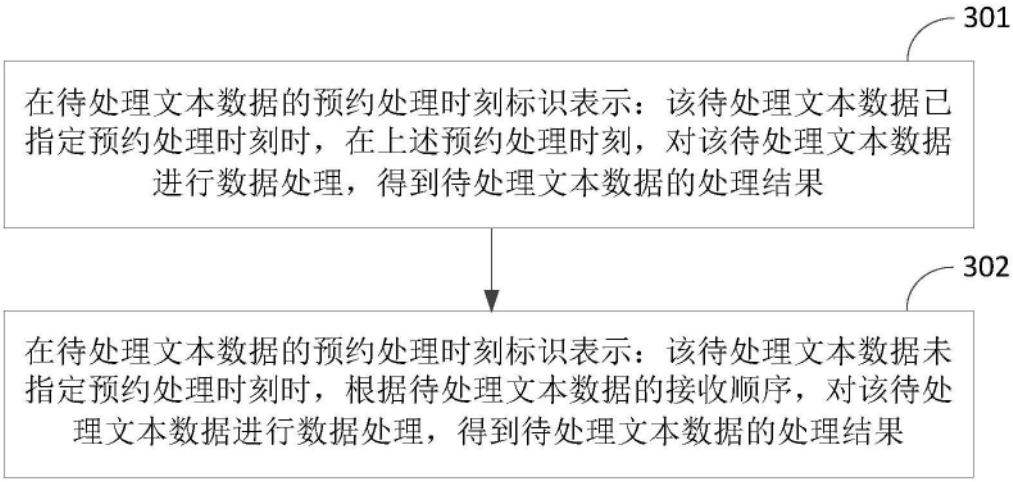

301
在待处理文本数据的预约处理时刻标识表示：该待处理文本数据已指定预约处理时刻时，在上述预约处理时刻，对该待处理文本数据进行数据处理，得到待处理文本数据的处理结果
302
在待处理文本数据的预约处理时刻标识表示：该待处理文本数据未指定预约处理时刻时，根据待处理文本数据的接收顺序，对该待处理文本数据进行数据处理，得到待处理文本数据的处理结果

图3

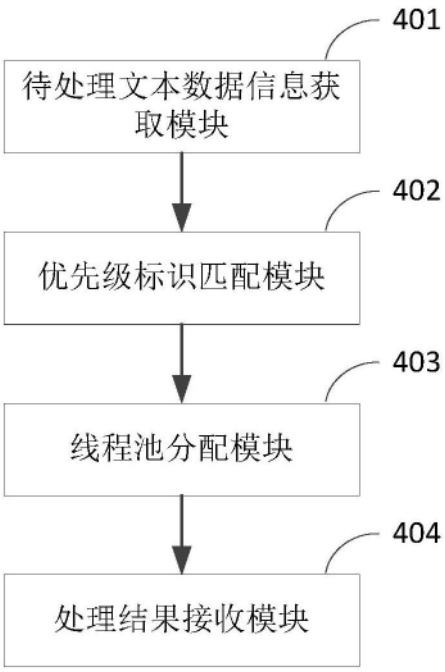

401
待处理文本数据信息获取模块
402
优先级标识匹配模块
403
线程池分配模块
404
处理结果接收模块

图4

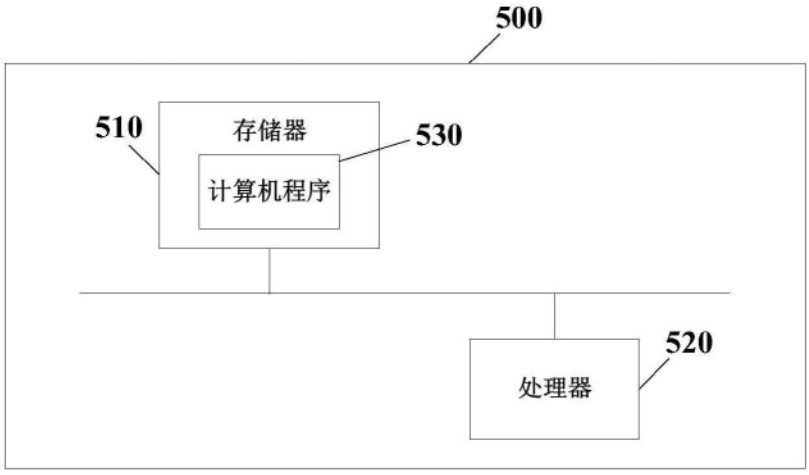


图5