(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 105989022 A
(43)申请公布日 2016.10.05

(21)申请号 201510050583.3

(22)申请日 2015.01.30

(71)申请人 北京陌陌信息技术有限公司
地址 100102 北京市朝阳区阜通东大街1号望京SOHO塔2B座20层

(72)发明人 张艳魁 高永芝

(74)专利代理机构 北京东方亿思知识产权代理有限责任公司 11258
代理人 付乐

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

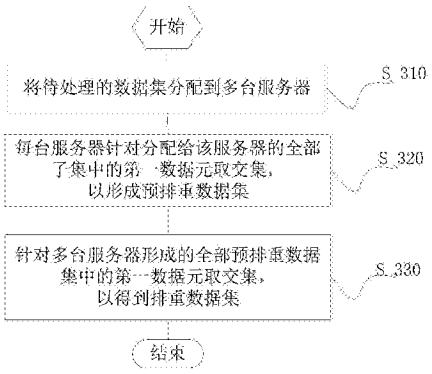
权利要求书2页 说明书6页 附图4页

(54)发明名称

数据排重方法及系统

(57)摘要

本申请提供了数据排重方法及系统。该方法包括：将待处理的数据集分配到多台服务器，其中，每台服务器被分配数据集中的多个子集，每个子集中的每个数据均包括具有相同属性的第一数据元；每台服务器至少针对分配给该服务器的全部子集中的第一数据元取交集，以形成预排重数据集；以及至少针对多台服务器形成的全部预排重数据集中的第一数据元取交集，以得到排重数据集。在本申请中还给出了数据排重系统的示例架构，并给出了这样的数据排重方法和系统的具体用例。本申请的技术方案大大减轻了每台服务器的工作负载、节省了数据排重时间，有效地提高了数据排重的效率和精度、提高了系统响应速度。



300

CN 105989022 A

1. 一种数据排重方法，包括：

将待处理的数据集分配到多台服务器，其中，每台服务器被分配所述数据集中的多个子集，每个子集中的每个数据均包括具有相同属性的第一数据元；

每台服务器至少针对分配给该服务器的全部子集中的第一数据元取交集，以形成预排重数据集；以及

至少针对所述多台服务器形成的全部预排重数据集中的第一数据元取交集，以得到排重数据集。

2. 根据权利要求 1 所述的方法，其中，每个子集对应于多个网格式的地理区域中的不同地理区域内的用户信息。

3. 根据权利要求 1 所述的方法，其中，每台服务器监视多个网格式的地理区域中的不同地理区域内的用户信息，并将每个地理区域内的用户信息存储为一个子集。

4. 根据权利要求 2 或 3 所述的方法，其中，所述第一数据元是用户 ID。

5. 根据权利要求 2 所述的方法，其中，将待处理的数据集分配到多台服务器包括：

将与所述多个网格式的地理区域中的相邻地理区域对应的子集分配给所述多台服务器中的不同服务器。

6. 根据权利要求 1 所述的方法，其中，每个子集中的各个数据具有不同的第一数据元。

7. 根据权利要求 1 所述的方法，其中，所述数据集与指定的时间单元相关联。

8. 根据权利要求 4 所述的方法，其中，所述数据集中的每个数据还包括与用户 ID 对应的经纬度信息。

9. 一种数据排重系统，包括：

多台第一服务器，其中，每台第一服务器被分配有待处理的数据集中的多个子集，每个子集中的每个数据均包括具有相同属性的第一数据元；

每台第一服务器被配置为至少针对分配给该服务器的全部子集中的第一数据元取交集，以形成预排重数据集；以及

第二服务器，所述第二服务器被配置为至少针对所述多台第一服务器形成的全部预排重数据集中的第一数据元取交集，以得到排重数据集。

10. 根据权利要求 9 所述的系统，其中，每个子集对应于多个网格式的地理区域中的不同地理区域内的用户信息。

11. 根据权利要求 9 所述的系统，其中，每台服务器监视多个网格式的地理区域中的不同地理区域内的用户信息，并将每个地理区域内的用户信息存储为一个子集。

12. 根据权利要求 10 或 11 所述的系统，其中，所述第一数据元是用户 ID。

13. 根据权利要求 10 所述的系统，其中，与所述多个网格式的地理区域中的相邻地理区域对应的子集被分配到所述多台第一服务器中的不同第一服务器上。

14. 根据权利要求 9 所述的系统，其中，每个子集中的各个数据具有不同的第一数据元。

15. 根据权利要求 9 所述的系统，其中，所述数据集与指定的时间单元相关联。

16. 根据权利要求 12 所述的系统，其中，所述数据集中的每个数据还包括与用户 ID 对应的经纬度信息。

17. 根据权利要求 12 所述的系统，其中，所述第二服务器可由所述第一服务器之一来

担任。

18.一种存储有指令的有形、非易失性计算机可读介质，当所述指令被一个或多个处理器运行时，使得所述一个或多个处理器执行如权利要求1-8中的任一项所述的数据排重方法。

数据排重方法及系统

技术领域

[0001] 本申请一般地涉及数据处理领域，更具体地，本申请涉及数据排重方法及系统。

背景技术

[0002] 随着信息技术的不断发展，各类数据大量涌现，其中不乏重复性数据，这些重复性数据不但给数据的存储造成了负担，有时数据的重复还会给后续操作带来困难，比如，在电信行业中如果存在重复电话单就会造成重复收费、在互联网广告领域重复点击也会造成重复计费、在文献管理领域对同一文献的重复收录和索引会给后续检索带来麻烦、在根据用户量来预付费的情况下对用户的重复计数也会导致重复计费等等。因此，对大量数据进行排重的需求越来越多。

[0003] 随着计算机的普及使用，人们很自然地想到使用计算机或服务器来进行海量数据的处理。但是，已有的数据排重方法通常使用单一服务器对大量数据进行处理，给服务器带来巨大的工作负荷，并且数据排重过程往往是繁琐且耗时的。

[0004] 因此，希望提供简单易行的数据排重解决方案，以快速、准确地对大量数据进行排重。

发明内容

[0005] 为了提供快速准确的数据排重解决方案，本申请提供了数据排重方法及系统。

[0006] 一方面，本申请提供了一种数据排重方法。该方法包括：

[0007] 将待处理的数据集分配到多台服务器，其中，每台服务器被分配所述数据集中的多个子集，每个子集中的每个数据均包括具有相同属性的第一数据元；

[0008] 每台服务器至少针对分配给该服务器的全部子集中的第一数据元取交集，以形成预排重数据集；以及

[0009] 至少针对所述多台服务器形成的全部预排重数据集中的第一数据元取交集，以得到排重数据集。

[0010] 可选地，在上述数据排重方法中，每个子集对应于多个网格式的地理区域中的不同地理区域内的用户信息。

[0011] 可选地，在上述数据排重方法中，每台服务器监视多个网格式的地理区域中的不同地理区域内的用户信息，并将每个地理区域内的用户信息存储为一个子集。

[0012] 可选地，在上述数据排重方法中，所述第一数据元是用户 ID。

[0013] 可选地，在上述数据排重方法中，所述位置信息包括用户所处的经度和纬度信息。

[0014] 可选地，在上述数据排重方法中，与相邻的地理区域对应的子集被分配到所述多台服务器中的不同服务器上。

[0015] 可选地，在上述数据排重方法中，每个子集中的各个数据具有不同的第一数据元。

[0016] 可选地，在上述数据排重方法中，所述数据集与指定的时间单元相关联。

[0017] 可选地，在上述数据排重方法中，所述数据集中的每个数据还包括与用户 ID 对应

的经纬度信息。

[0018] 另一方面，本申请还提供了一种数据排重系统。该系统包括：

[0019] 多台第一服务器，其中，每台第一服务器被分配有待处理的数据集中的多个子集，每个子集中的每个数据均包括具有相同属性的第一数据元；

[0020] 每台第一服务器被配置为至少针对分配给该服务器的全部子集中的第一数据元取交集，以形成预排重数据集；以及

[0021] 第二服务器，所述第二服务器被配置为至少针对所述多台第一服务器形成的全部预排重数据集中的第一数据元取交集，以得到排重数据集。

[0022] 再一方面，本申请还提供了一种有形、非易失性计算机可读介质，其上存储有指令，当这些指令被一个或多个处理器执行时，使得这些处理器，执行上面描述的数据排重方法。

[0023] 本申请的实施例的技术方案通过将待处理的数据集分配到多台服务器，由每台服务器分别对各自的数据集执行预排重，然后对多台服务器形成的全部预排重数据集执行排重得到排重数据集，将存储和计算负载分布到多台服务器上，通过并行处理若干被切分的小规模问题（例如，每台服务器各自的预排重运算），然后再对运算结果进行进一步求精来实现一项计算任务（例如，对多台服务器形成的全部预排重数据集执行排重），大大减轻了每台服务器的工作负载、节省了数据处理时间，有效地提高了数据处理的效率，提高了系统响应速度。

附图说明

[0024] 为了更清楚地说明本申请的实施例的技术方案，下面将结合附图对实施例进行描述，这些附图形成了本文的一部分并通过图解的方式示出了具体示例实施例，其中：

[0025] 图 1 为根据一个实施例的数据排重系统的架构示意图。

[0026] 图 2 为根据另一实施例的据排重系统的另一架构示意图。

[0027] 图 3 为示出了根据实施例的数据排重方法的过程的流程图。

[0028] 图 4 为示出了根据一具体用例的数据排重方法的过程的流程图。

具体实施方式

[0029] 在下面对本申请的详细描述中阐述了很多具体细节，以便于充分理解本申请。但是，在没有这些具体细节的情况下也可以实施本申请，对于本领域的技术人员来说是很明显的。在另外一些示例里，没有对公知的方法、过程和部件进行详细的描述，以避免喧宾夺主、淡化了本申请的主要内容。

[0030] 在整个说明书和权利要求书中，术语或短语可能具有明确表述的意思之外的在上下文中暗示或暗指的有细微差别的含义。例如，术语“数据集”可以表示任何数据的集合，可以是多元数据也可以是一元数据的集合，可以是数字、字母、字符串、符号、文字等等多种形式。术语“数据元”表示数据集中的数据的组成元素，数据可以由一个或多个数据元组成。并且数据集可以根据预定的规则（例如，可以根据其中的某一数据元）被拆分成预定数目的子集。术语“服务器”也可以指代处理器、运算器等能够实现本申请描述的技术方案的任何计算器件与存储器件的组合。短语“网格式的地理区域”在本申请中意指在地图上进行

划分得到一系列网格，其中每个格子表示一个网格式的地理区域。短语“在一个实施例中”不一定指相同的实施例，并且短语“在另一实施例中”不一定指不同的实施例。例如，要求保护的主题旨在包括示例实施例全部或部分的组合。

[0031] 为了提供快速准确的数据排重解决方案，本申请的实施例提供了数据排重的方法和系统。

[0032] 图 1 给出了这样的数据排重系统 100 的一个示例。如图所示，系统 100 包括多台第一服务器 101 和第二服务器 102。在本申请中，服务器可以是传统意义上的服务器，也可以是们能够实现其功能的其他器件，比如，处理器或运算器与存储器的组合等等。在一个实施例中，第一服务器 101 和第二服务器 102 可以是相同的服务器。在另一实施例中，第二服务器 102 可由多台第一服务器 101 之一来担任。这些服务器通过网络互连，以发送或接收数据和信令，并进行同步。网络可以是任何类型的网络，例如，公共交换电话网 (PSTN)、互联网 (Internet)、局域网 (LAN)、广域网 (WAM)、有线网络、无线网络等的任意组合。

[0033] 首先，假定待处理的数据集中的每个数据存在至少一个可供比较的数据元，称为相同属性的数据元。优选地，数据集中的每个数据的每个相应数据元属性都相同，即数据集中的数据全部按照相同的格式进行存储。待处理的数据集中可能存在一个或多个重复数据。所谓重复数据并不是指数据完全相同，而是在数据处理中将要考虑的一个或多个数据元相同。待处理的数据集被划分为多个子集。优选地，数据集的每个子集中的数据具有不同的第一数据元。这些子集被分配给多台第一服务器 101，使得每台第一服务器 101 被分配有待处理的数据集中的一个或多个子集。接下来，每台第一服务器 101 针对分配给该服务器的全部子集中的至少第一数据元取交集。全部多台第一服务器 101 的处理结果组成预排重数据集。预排重数据集被传输到第二服务器 102。第二服务器 102 同样针对预排重数据集中的至少第一数据元取交集。从而得到针对至少第一数据元的排重数据集。在其他实施例中，排重可以针对两个或更多个数据元进行。

[0034] 在需要排重的数据量非常庞大，而且可以用于排重的服务器足够多的情况下，可以使用图 2 所提供的数据排重系统 200。在图 2 中，数据排重系统 200 包括多台第一服务器 201、多台第二服务器 202 和第三服务器 203，其中第二服务器 202 的数目小于第一服务器 201 的数目。在一个实施例中，第一服务器 201、第二服务器 202 和第三服务器 203 可以是相同的服务器。在另一实施例中，多台第二服务器 202 和第三服务器 203 可由多台第一服务器 201 中的一部分来担任。这些服务器通过网络互连，以发送或接收数据和信令，并进行同步。

[0035] 首先，假定待处理的数据集中的每个数据存在至少一个可供比较的数据元，称为相同属性的数据元。优选地，数据集中的每个数据的每个相应数据元属性都相同，即数据集中的数据全部按照相同的格式进行存储。待处理的数据集中可能存在一个或多个重复数据。所谓重复数据并不是指数据完全相同，而是在数据处理中将要考虑的一个或多个数据元相同。待处理的数据集被划分为多个子集。优选地，数据集的每个子集中的数据具有不同的第一数据元。这些子集被分配给多台第一服务器 201，使得每台第一服务器 201 被分配有待处理的数据集中的一个或多个子集。接下来，每台第一服务器 201 针对分配给该服务器的全部子集中的至少第一数据元取交集。此形成第一预排重数据集。第一预排重数据集被进一步分配给多台第二服务器 202。第二服务器 202 同样针对预排重数据集中的至少第

一数据元取交集。从而得到针对至少第一数据元的第二预排重数据集。第二预排重数据集被传输至第三服务器 203，由第三服务器 203 进一步针对至少第一数据元取交集，得到排重后的数据集。

[0036] 可选地，在其他实施例中，多台第一服务器 201 中的每台第一服务器 201 针对第一数据元对分配给该服务器的全部子集求交集，得到第一预排重数据集。接着，多台第一服务器 202 中的每台第一服务器 202 针对第二数据元对分配给该服务器的全部第一预排重数据集的子集求交集，得到第二预排重数据集。而第三服务器 203 可以同时针对第一数据元和第二数据元对第二预排重数据集的各个子集求交集，从而得出待处理数据集关于第一数据元和第二数据元的排重数据集。当然，排重也可以针对更多的数据元进行。

[0037] 虽然，在上面的实施例中仅给出了 2 级和 3 级服务器架构，但是本领域普通技术人员将理解的是，在适当的情况下，本申请的数据排重系统可以具有更多级的服务器架构，可以处理更多的数据，可以针对更多的数据元对数据集进行排重运算。上面的实施例仅是示例性的而不是限制性的。

[0038] 图 3 示出了根据实施例的数据排重方法 300 的流程图。如前面所描述的，假定待处理的数据集中的每个数据存在至少一个可供比较的数据元，称为相同属性的数据元。优选地，数据集中的每个数据的每个相应数据元属性都相同，即数据集中的数据全部按照相同的格式进行存储。待处理的数据集中可能存在一个或多个重复数据。所谓重复数据并不是指数据完全相同，而是在数据处理中将要考虑的一个或多个数据元相同。待处理的数据集被划分为多个子集。优选地，数据集的每个子集中的数据具有不同的第一数据元。过程 300 开始于步骤 S310，在步骤 S310 中，将数据集的多个子集分配给多台服务器。每台服务器分配有待处理数据的多个子集，每个子集中的每个数据均至少包括具有相同属性的第一数据元。在步骤 S320 中，多台服务器中的每台服务器针对分配给该服务器的全部子集中的第一数据元取交集，以形成预排重数据集。接下来，在步骤 S330 中，针对多台服务器形成的全部预排重数据集中的第一数据元取交集，以得到排重数据集。过程 300 结束。

[0039] 应该指出的是，上面仅描述了数据排重的一般过程，即包括预排重和排重，在一些实施例中，这样的预排重步骤可能不止执行一次，而且可以针对不止一个数据元进行，比如，第一预排重可以由多台第一服务器中的每台第一服务器分别针对第一数据元对待处理数据的多个子集进行预排重，得到第一预排重数据集；第二预排重可以由多台第二服务器中的每台第二服务器分别针对第二数据元（也可以继续针对第一数据元）对第一预排重数据集的多个子集进行预排重，得到第二预排重数据集；接着，可以由第三服务器针对第一数据元和第二数据元（也可以只针对第一数据元）对第二预排重数据集的多个子集进行排重，最后得到针对第一数据元和第二数据元的排重数据集。根据上述描述，本领域普通技术人员可以想到本申请的技术方案可以通过更多级预排重针对数据所包含的更多数据元对待处理数据进行排重。

[0040] 下面将结合具体实例介绍上面描述的数据排重方法的应用，但是本领域普通技术人员应当理解本申请的数据排重方法还可以应用于其他实例。

[0041] 随着智能移动终端的普及使用，越来越多得用户开始借助于移动终端上安装的各种应用来获取信息，这使得商家看到了商机，越来越多的商家借助移动终端上安装的各种应用来向移动终端的用户推送信息，从而推广其产品或服务。往往，商家所推送的产品或服

务信息对其周围一定距离范围内的移动终端的用户而言是有用信息，而对于处于较远地域的用户而言就不那么有用了。因此，商家希望仅向其周围一定距离范围内的移动终端推送产品或服务信息，从而控制成本、提高回报率。这就涉及到预估一定距离范围内、在某一时间单元（比如，一天）内的净用户数，从而合理规划信息投放时间段和地域范围，获得期望的回报率。

[0042] 在指定时间单元内，用户可能被重复登记在关于一定距离范围的统计数据中（比如，用户移动到一定距离范围之外的地域之后又返回一定距离范围之内，或者用户退出或重新登录相关应用等）。因此，简单地记录在在指定时间单元内出现在一定距离范围内的用户，会存在重复计数，需要对用户信息进行排重以得出一定距离范围内、在指定时间单元内的净用户数。

[0043] 在本实例中，应用上面描述的数据排重方法来对一定距离范围内、在指定时间单元内的用户信息进行排重。

[0044] 在现有技术中，通过墨卡托投影将地球球面投影在二维平面上，对投影所得的二维平面进行四等分，并根据预定的距离单位（比如，1 千米、2 千米等等）逐步细分（比如，N 次），得到多个（比如，4^N个）网格式的地理区域。从而，一定的地理范围由一个或多个网格式的地理区域构成。

[0045] 在此情景下，数据集是一定距离范围内、在指定时间单元内的用户信息。数据集中的每个子集与多个网格式的地理区域中的不同的地理区域内的用户信息相对应。在实施例中，待处理的数据集被分配到多台服务器，可以是将已有的用户信息的多个子集分配到多台服务器，或者，将多个网格式的地理区域分配给多台服务器中的每台服务器，由每台服务器分别监测被分配给该服务器的所有网格式的地理区域内的用户信息，并针对指定的时间单元（比如，一天）存储为用户信息的一个子集。

[0046] 为了统计一定距离范围内、在指定时间单元内的用户信息，为每个用户分配一个用户 ID，并将用户 ID 作为用户信息的第一数据元来存储用户信息。可选地，用户信息还可以包括对应的位置信息（比如，经纬度、地理标志等），以及用户出现在该地理位置的时间信息。

[0047] 如上面所述的，将待处理的数据集分配到多台服务器例如可以是将与相邻的地理区域对应的子集分配到多台服务器中的不同服务器。优选地，每个子集中的各条用户信息具有不同的用户 ID，即在每个子集中，针对同一用户 ID 仅保存一条用户信息。由于同一用户在同一时间单元内出现在不相邻地理区域内的几率相对出现在相邻地理区域内的几率小得多，将相邻地理区域对应的子集分配给不同服务器，这大大减轻了每台服务器的运算负载，有效地提高了运算速率。

[0048] 在可替代的实施例中，将待处理的数据集分配到多台服务器例如可以通过将多个网格式的地理区域分配给多台服务器中的每台服务器，由每台服务器分别监测被分配给该服务器的所有网格式的地理区域内的用户信息，并针对指定的时间单元（比如，一天）将每个地理区域内的用户信息存储为一个子集。优选地，每个子集中的各条用户信息具有不同的用户 ID，即每台服务器在指定时间单元内针对同一用户 ID 在属于一个地理区域的子集内仅存储一条用户信息。由于同一用户在同一时间单元内出现在不相邻地理区域内的几率相对出现在相邻地理区域内的几率小得多，将相邻地理区域分配给不同服务器进行监测并

存储，这大大减轻了每台服务器的运算负载，有效地提高了运算速率。下面参考图 4 描述根据该具体实施例的数据排重方法。

[0049] 图 4 示出了在该实施例中的数据排重方法 400 的流程图。过程 400 开始于步骤 S410，首先，在步骤 S410，将二维地图划分为多个网格式的地理区域。在步骤 S420，将多个网格式的地理区域分配到多台服务器。优选地，相邻地理区域被分配给多台服务器中的不同服务器。在步骤 S430 中，每台服务器监测被分配给该服务器的每个地理区域内的用户，并针对指定时间单元将用户 ID 作为第一数据元来存储用户信息，将每个地理区域内的用户信息存储为用户信息的一个子集。优选地，每个子集中的各个用户信息具有不同的用户 ID。另外，用户 ID 可以与用户所处的地理位置信息，以及用户在该地理位置的时间信息相关联地存储。在步骤 S440，每台服务器针对该服务器监测得到的全部子集中的用户 ID 取交集，获得预排重用户信息。接下来，在步骤 S450，针对多台服务器形成的全部预排重用户信息中的用户 ID 取交集，以得到排重用户信息。过程 400 结束。

[0050] 上述实施例的技术方案，通过将网格状的地理区域均匀地分配到多台服务器，由每台服务器分别监测分配给该服务器的每个地理区域内的用户，并针对指定时间单元将用户 ID 作为第一数据元来存储用户信息，将每个地理区域内的用户信息存储为一个子集，将监测、存储和计算负载分布到多台服务器上进行，通过并行处理若干被切分的小规模问题（例如，每台服务器各自的交集运算），然后再对处理结果进行汇总和进一步求精来实现一项计算任务（例如，针对预排重用户信息再进行交集运算得到排重用户信息），大大减轻了每台服务器的工作负载、节省了数据处理时间，有效地提高了数据处理的效率和精度。

[0051] 上面给出了本申请所提供的数据排重方法和系统的一个用例，但是本申请普通技术人员应当理解，本申请所提供的数据排重方法和系统还可以用于其他各种数据的排重，比如，在电信领域用于话单的排重、在文献管理领域用于文献资料的排重、等等，在此不逐一例举。

[0052] 另外，本申请的数据排重方法可以作为逻辑指令被编码在一个或多个有形计算机可读介质中以供一个或多个处理器执行。例如，计算机可读介质例如可以是电子介质（例如，RAM（随机存取存储器）、ROM（只读存储器）、EPROM（可擦除可编程只读存储器））、磁介质、光介质（例如，CD、DVD）、电磁介质、半导体技术介质或任意其他合适的介质。

[0053] 以上在实施例中描述了本申请的优选实施例。尽管在特定实施例中描述了本申请，但是应当理解在不脱离本发明的范围的情况下可以进行许多变化和修改。因此，希望以上详细描述被认为是示意性的而不是限制性的，并且要理解意欲限定本发明的精神和范围的是所附的权利要求，包括所有等同物。

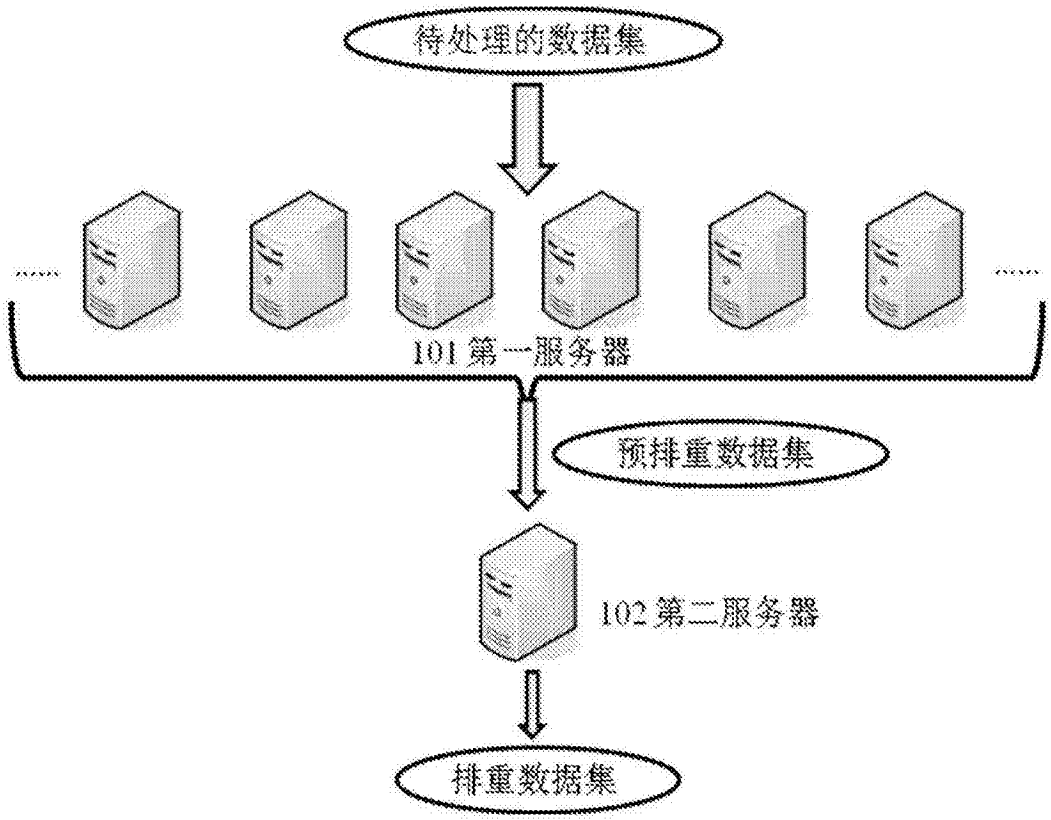
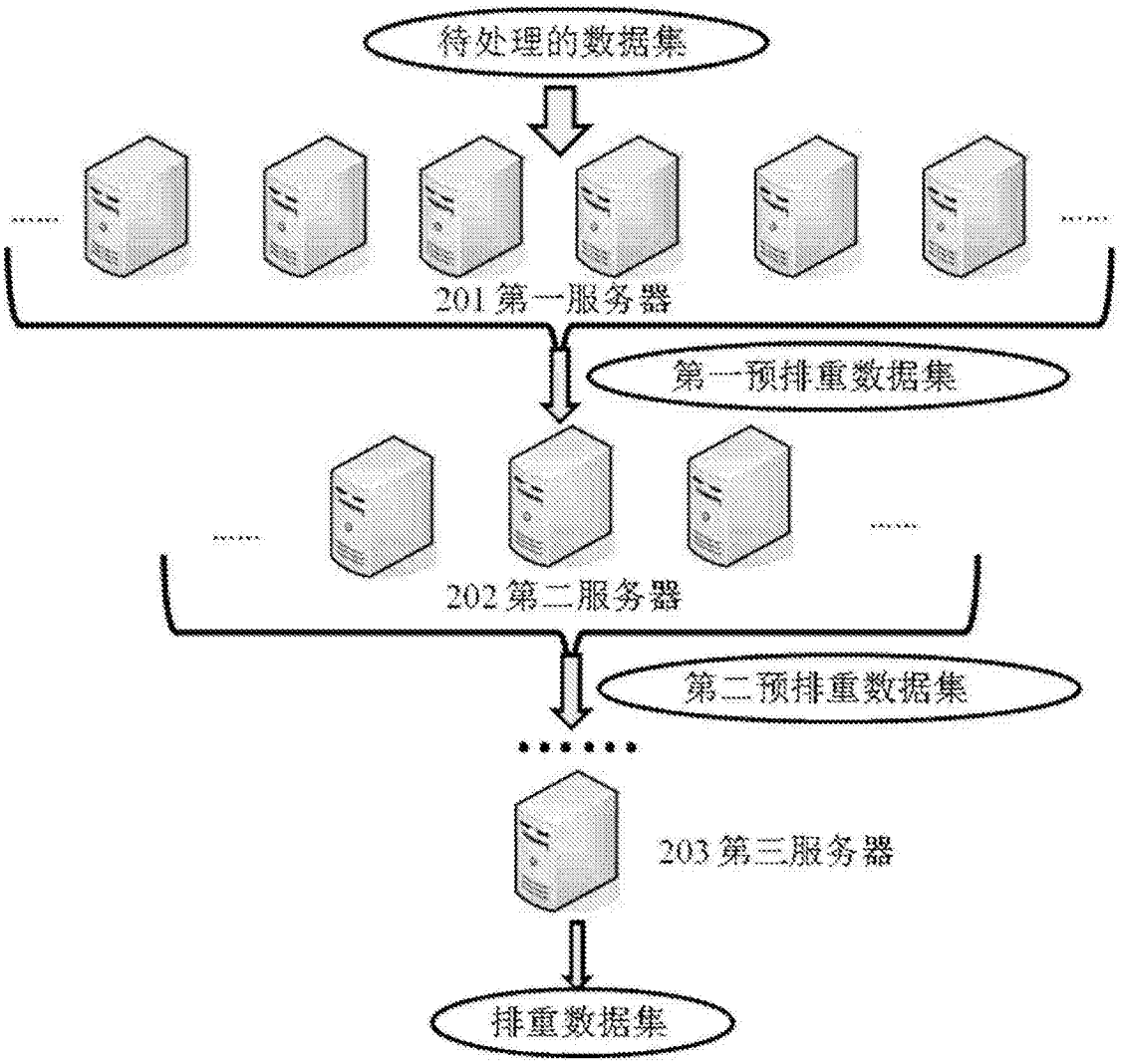


图 1


待处理的数据集
201 第一服务器
第一预排重数据集
202 第二服务器
第二预排重数据集
203 第三服务器
排重数据集

200

图 2

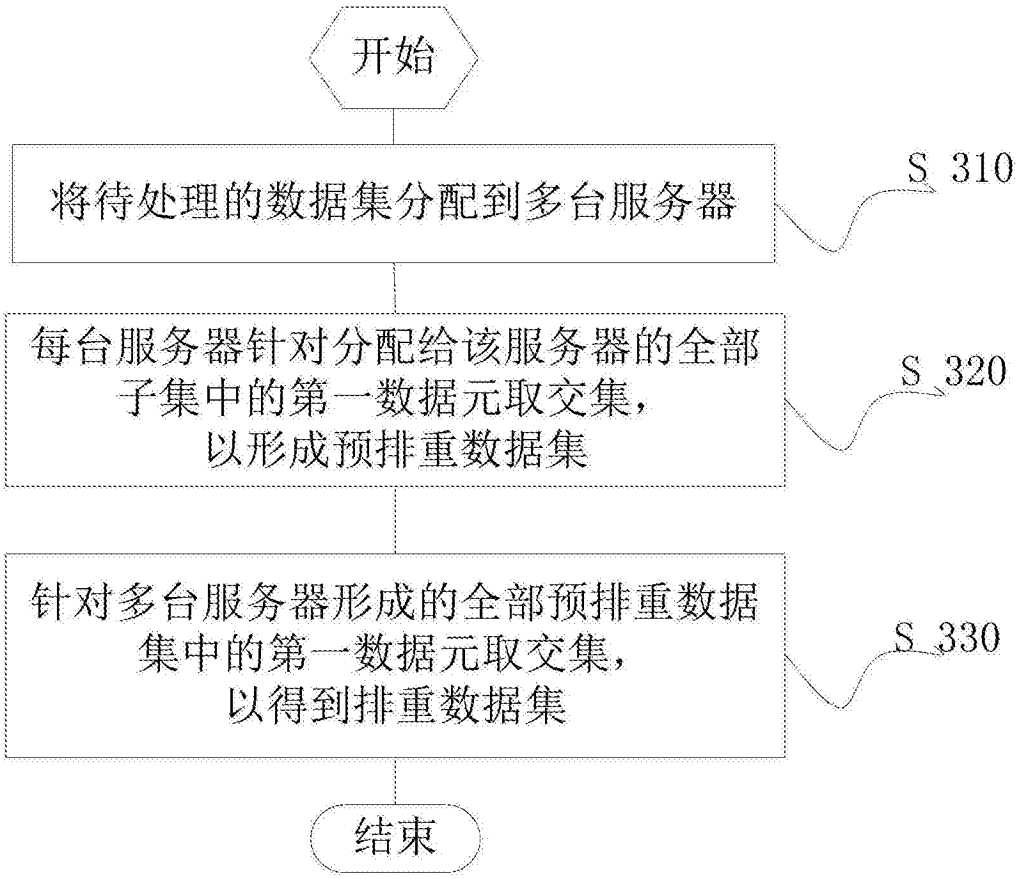

开始
将待处理的数据集分配到多台服务器
S 310
每台服务器针对分配给该服务器的全部子集中的第一数据元取交集，以形成预排重数据集
S 320
针对多台服务器形成的全部预排重数据集中的第一数据元取交集，以得到排重数据集
S 330
结束
300

图 3

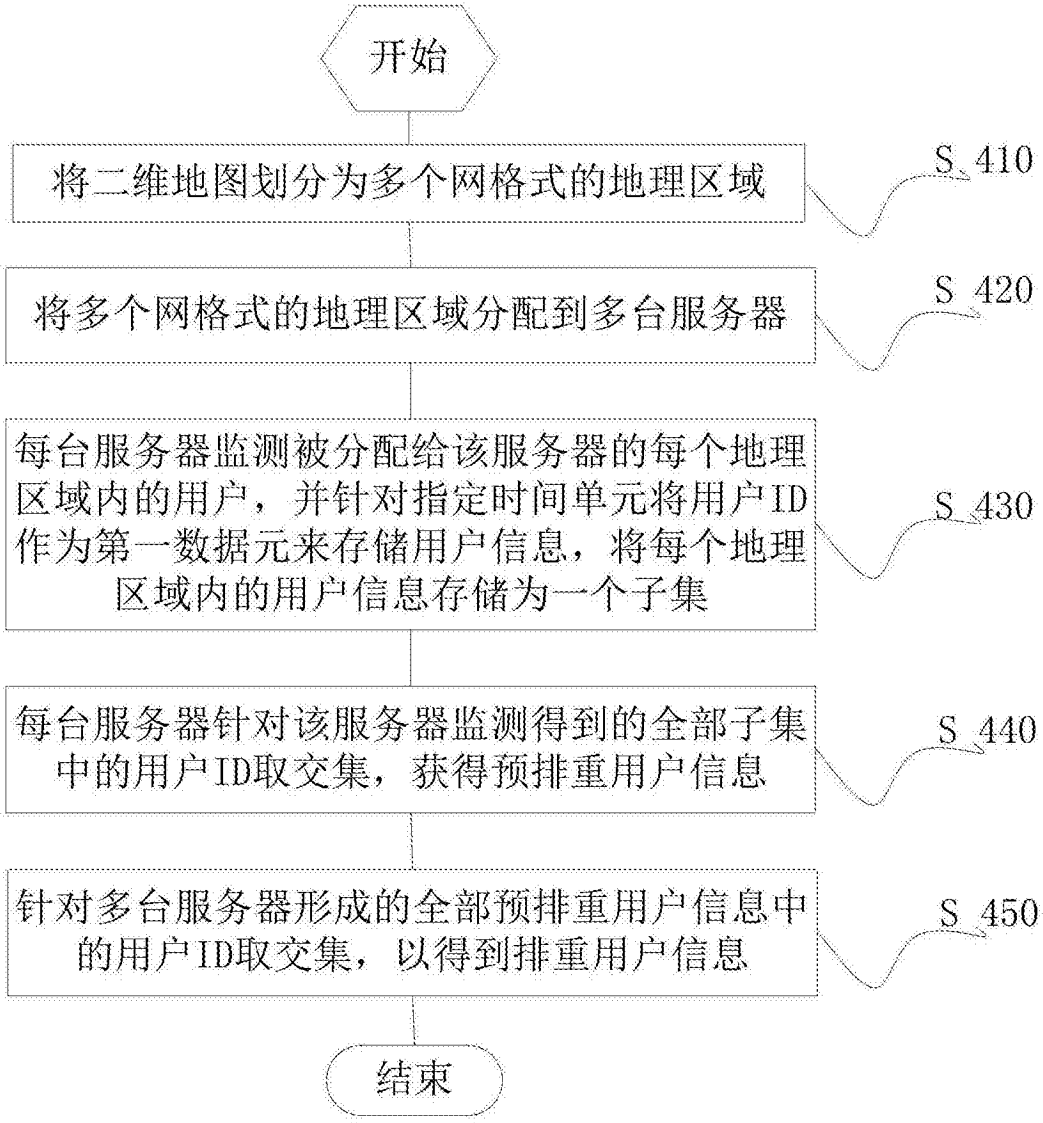

开始
将二维地图划分为多个网格式的地理区域
S 410
将多个网格式的地理区域分配到多台服务器
S 420
每台服务器监测被分配给该服务器的每个地理区域内的用户，并针对指定时间单元将用户ID作为第一数据元来存储用户信息，将每个地理区域内的用户信息存储为一个子集
S 430
每台服务器针对该服务器监测得到的全部子集中的用户ID取交集，获得预排重用户信息
S 440
针对多台服务器形成的全部预排重用户信息中的用户ID取交集，以得到排重用户信息
S 450
结束
400

图 4